CCIT Report #482 May 2004

On Joint Coding for Watermarking and Encryption

Neri Merhav

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

Abstract

In continuation to earlier works where the problem of joint information embedding and lossless compression (of the composite signal) was studied in the absence [8] and in the presence [9] of attacks, here we consider the additional ingredient of protecting the secrecy of the watermark against an unauthorized party, which has no access to a secret key shared by the legitimate parties. In other words, we study the problem of joint coding for three objectives: information embedding, compression, and encryption. Our main result is a coding theorem that provides a single–letter characterization of the best achievable tradeoffs among the following parameters: the distortion between the composite signal and the covertext, the distortion in reconstructing the watermark by the legitimate receiver, the compressibility of the composite signal (with and without the key), and the equivocation of the watermark, as well as its reconstructed version, given the composite signal. In the attack–free case, if the key is independent of the covertext, this coding theorem gives rise to a threefold *separation principle* that tells that asymptotically, for long block codes, no optimality is lost by first applying a rate–distortion code to the watermark source, then encrypting the compressed codeword, and finally, embedding it into the covertext using the embedding scheme of [8]. In the more general case, however, this separation principle is no longer valid, as the key plays an additional role of side information used by the embedding unit.

Index Terms: Information hiding, watermarking, encryption, data compression, separation principle, side information, equivocation, rate–distortion.

1 Introduction

It is common to say that encryption and watermarking (or information hiding) are related but they are substantially different in the sense that in the former, the goal is to protect the secrecy of the *contents* of information, whereas in the latter, it is the very *existence* of this information that is to be kept secret.

In the last few years, however, we are witnessing increasing efforts around the *combination* of encryption and watermarking, which is motivated by the desire to further enhance the security of sensitive information that is being hidden in the host signal. This is to guarantee that even if the watermark is somehow detected by a hostile party, its contents still remain secure due to the encryption. This combination of watermarking and encryption can be seen both in recently reported research work (see, e.g., [1],[2],[6],[7],[13] and references therein) and in actual technologies used in commercial products with a copyright protection framework, such as the CD and the DVD. Also, some commercial companies that provide Internet documents, have in their websites links to copyright warning messages, saying that their data are protected by digitally encrypted watermarks (see, e.g., `http://genealogy.lv/1864Lancaster/copyright.htm`).

This paper is devoted to the information–theoretic aspects of joint watermarking and encryption together with lossless compression of the composite signal that contains the encrypted watermark. Specifically, we extend the framework studied in [8] and [9] of joint watermarking and compression, so as to include encryption using a secret key. Before we describe the setting of this paper concretely, we pause then to give some more detailed background on the work reported in [8] and [9].

In [8], the following problem was studied: Given a covertext source vector $X^n = (X_1, \ldots, X_n)$, generated by a discrete memoryless source (DMS), and a message m, uniformly distributed in $\{1, 2, \ldots, 2^{nR_e}\}$, independently of X^n, with R_e designating the embedding rate, we wish to generate a composite (stegotext) vector $Y^n = (Y_1, \ldots, Y_n)$ that satisfies the following requirements: (i) Similarity to the covertext, in the sense that a distortion constraint, $Ed(X^n, Y^n) = \sum_{t=1}^{n} Ed(X_t, Y_t) \leq nD$, holds, (ii) compressibility, in the sense that the normalized entropy, $H(Y^n)/n$, does not exceed some threshold R_c, and (iii) reliability in decoding the message m from Y^n, in the sense that the decoding error probability is arbitrarily small for large n. A single–letter characterization of the best achievable

tradeoffs among R_c, R_e, and D was given in [8], and was shown to be achievable by an extension of the ordinary lossy source coding theorem, giving rise to the existence of 2^{nR_e} *disjoint* rate–distortion codebooks (one per each possible watermark message) as long as R_e does not exceed a certain fundamental limit. In [9], this setup was extended to include a given memoryless attack channel, $P(Z^n|Y^n)$, where item (iii) above was redefined such that the decoding was based on Z^n rather than on Y^n. This extension required a completely different approach, which was in the spirit of the Gel'fand–Pinsker coding theorem for a channel with non–causal side information (SI) at the transmitter [5]. The role of SI, in this case, was played by the covertext.

In this paper, we extend the settings of [8] and [9] to include encryption. For the sake of clarity of the exposition, we do that in several steps.

In the first step, we extend the attack–free setting of [8]: In addition to including encryption, we also extend the model of the watermark message source to be an arbitrary DMS, $U_1, U_2, \ldots$, independent of the covertext, and not necessarily a binary symmetric source (BSS) as in [8] and [9]. Specifically, we now assume that the encoder has three inputs (see Fig. 1): The covertext source vector, X^n, an independent (watermark) message source vector $U^N = (U_1, \ldots, U_N)$, where N may differ from n if the two sources operate in different rates, and a secret key (shared also with the legitimate decoder) $K^n = (K_1, \ldots, K_n)$, which, for mathematical convenience, is assumed to operate at the same rate as the covertext. It is assumed, at this stage, that K^n is independent of U^N and X^n. Now, in addition to requirements (i)-(iii), we impose a requirement on the equivocation of the message source relative to an eavesdropper that has access to Y^n, but not to K^n. Specifically, we would like the normalized conditional entropy, $H(U^N|Y^n)/N$, to exceed a prescribed threshold, h (e.g., $h = H(U)$ for perfect secrecy). Our first result is a coding theorem that gives a set of necessary and sufficient conditions, in terms of single–letter inequalities, such that a triple (D, R_c, h) is achievable while maintaining reliable reconstruction of U^N at the legitimate receiver.

In the second step, we relax the requirement of perfect (or almost perfect) watermark reconstruction, and assume that we willing to tolerate a certain distortion between the watermark message U^N and its reconstructed version $\hat{U}^N$, that is, $Ed'(U^N, \hat{U}^N) = \sum_{i=1}^{N} Ed'(U_i, \hat{U}_i) \leq ND'$. For example, if d' is the Hamming distortion measure then D', of course, designates the maximum allowable bit error probability (as opposed to the block

error probability requirement of [8] and [9]). Also, in this case, it makes sense, in some applications, to impose a requirement regarding the equivocation of the *reconstructed* message, $\hat{U}^N$, namely, $H(\hat{U}^N|Y^n)/N \geq h'$, for some prescribed constant h'. The rationale is that it is $\hat{U}^N$, not U^N, that is actually conveyed to the legitimate receiver. For the sake of generality, however, we will take into account both equivocation requirements, with the understanding that if one of them is superfluous, then the corresponding threshold (h or h' accordingly) can always be set to zero. Our second result then extends the above–mentioned coding theorem to a single–letter characterization of achievable quintuples (D, D', R_c, h, h'). As will be seen, this coding theorem gives rise to a threefold separation theorem, that separates, without asymptotic loss of optimality, between three stages: rate–distortion coding of U^N, encryption of the compressed bitstream, and finally, embedding the resulting encrypted version using the embedding scheme of [8]. The necessary and sufficient conditions related to the the encryption are completely decoupled from those of the embedding and the stegotext compression.

In the third and last step, we drop the assumption of an attack–free system and we assume a memoryless attack channel, in analogy to [9]. As it will turn out, in this case there is interaction between the encryption and the embedding, even if the key is still assumed independent of the covertext. In particular, it will be interesting to see that the key, in addition to its original role in encryption, serves as SI that is available to both encoder and decoder (see Fig. 2). Also, because of the dependence between the key and the composite signal, and the fact that the key is available to the legitimate decoder as well, it may make sense, at least in some applications, to let the compressibility constraint correspond to the the conditional entropy of Y^n given K^n. Again, for the sake of generality, we will consider both the conditional and the unconditional entropies of Y^n, i.e., $H(Y^n)/n \leq R_c$ and $H(Y^n|K^n)/n \leq R_c'$. Our final result then is a coding theorem that provides a single–letter characterization of the region of achievable six–tuples $(D, D', R_c, R_c', h, h')$. Interestingly, this characterization remains essentially unaltered even if there is dependence between the key and the covertext, which is a reasonable thing to have once the key and the stegotext interact anyhow.[1] In this context, the system designer confronts an interesting dilemma regarding the desirable degree of statistical dependence between the key and the covertext,

[1] In fact, the choice of the conditional distribution $P(K^n|X^n)$ is a degree of freedom that can be optimized subject to the given randomness resources.

which affects the dependence between the key and the stegotext. On the one hand, strong dependence can reduce the entropy of Y^n given K^n (and thereby reduce R'_c), and also help in the embedding process: For example, the extreme case of $K^n = X^n$ (which corresponds to *private* watermarking since the decoder actually has access to the covertext) is particularly interesting because in this case, for the encryptyion key, there is no need for any external resources of randomness, in addition to the randomness of the covertext that is already available. On the other hand, when there is strong dependence between K^n and Y^n, the secrecy of the watermark might be sacrificed since $H(K^n|Y^n)$ decreases as well. An interesting point, in this context, is that the Slepian–Wolf encoder [12] (see Fig. 2) is used to generate, from K^n, random bits that are essentially independent of Y^n (as Y^n is generated only after the encryption). These aspects will be seen in detail in Section 3.

The remaining parts of this paper are organized as follows: In Section 2, we set some notation conventions. Section 3 will be devoted to a formal problem description and to the presentation of the main result for the attack–free case with distortion–free watermark reconstruction (first step described above). In Section 4, the setup and the results will be extended along the lines of the second and the third steps, detailed above, i.e., a given distortion level in the watermark reconstruction and the incorporation of an attack channel. Finally, Sections 5 and 6 will be devoted to the proof of the last (and most general) version of the coding theorem, with Section 5 devoted to the converse part, and Section 6 – to the direct part.

2 Notation Conventions

We begin by establishing some notation conventions. Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors and their sample values, which will be denoted with same symbols superscripted by the dimension. Thus, for example, A^ℓ (ℓ – positive integer) will denote a random ℓ-vector $(A_1, ..., A_\ell)$, and $a^\ell = (a_1, ..., a_\ell)$ is a specific vector value in $\mathcal{A}^\ell$, the ℓ-th Cartesian power of $\mathcal{A}$. The notations a_i^j and A_i^j, where i and j are integers and $i \leq j$, will designate segments $(a_i, \ldots, a_j)$ and $(A_i, \ldots, A_j)$, respectively, where for $i = 1$, the subscript will be omitted (as above). For $i > j$, a_i^j (or A_i^j) will be understood as the null string. Sequences without specifying indices

are denoted by $\{\cdot\}$.

Sources and channels will be denoted generically by the letter P, or Q, subscripted by the name of the RV and its conditioning, if applicable, e.g., $P_U(u)$ is the probability function of U at the point $U = u$, $P_{K|X}(k|x)$ is the conditional probability of $K = k$ given $X = x$, and so on. Whenever clear from the context, these subscripts will be omitted. Information theoretic quantities like entropies and mutual informations will be denoted following the usual conventions of the Information Theory literature, e.g., $H(U^N)$, $I(X^n; Y^n)$, and so on. For single–letter infomation quantities (i.e., when $n = 1$ or $N = 1$), subscripts will be omitted, e.g., $H(U^1) = H(U_1)$ will be denoted by $H(U)$, similarly, $I(X^1; Y^1) = I(X_1; Y_1)$ will be denoted by $I(X; Y)$, and so on.

3 Problem Definition and Main Result for Step 1

We now turn to the formal description of the model and the problem setting for step 1, as described in the Introduction. A source P_X, henceforth referred to as the *covertext source* or the *host source*, generates a sequence of independent copies, $\{X_t\}_{t=-\infty}^{\infty}$, of a finite–alphabet RV, $X \in \mathcal{X}$. At the same time and independently, another source P_U, henceforth referred to as the *message source*, or the *watermark source*, generates a sequence of independent copies, $\{U_i\}_{i=-\infty}^{\infty}$, of a finite–alphabet RV, $U \in \mathcal{U}$. The relative rate between the message source and the covertext source is λ message symbols per covertext symbol. This means that while the covertext source generates a block of n symbols, say, $X^n = (X_1, \ldots, X_n)$, the message source generates a block of $N = \lambda n$ symbols, $U^N = (U_1, \ldots, U_N)$ (assuming, without essential loss of generality, that λn is a positive integer). In addition to the covertext source and the message source, yet another source, P_K, henceforth referred to as the *key source*, generates a sequence of independent copies, $\{K_t\}_{t=-\infty}^{\infty}$, of a finite–alphabet RV, $K \in \mathcal{K}$, independently[2] of both $\{X_t\}$ and $\{U_i\}$. They key source is assumed to operate at the same rate as the covertext source, that is, while the covertext source generates the block of length n, X^n, the key source generates a block of n symbols as well, $K^n = (K_1, \ldots, K_n)$.

Given n and λ, a block code for *joint watermarking, encryption, and compression* is a mapping $f_n : \mathcal{U}^N \times \mathcal{X}^n \times \mathcal{K}^n \to \mathcal{Y}^n$, $N = \lambda n$, whose output $y^n = (y_1, \ldots, y_n) = f_n(u^N, x^n, k^n) \in \mathcal{Y}^n$ is referred to as the *stegotext* or the *composite signal*, and accordingly,

[2]The assumption of independence between $\{K_t\}$ and $\{X_t\}$ is temporary and made now primarily for the sake of simplicity of the exposition. It will be dropped later on.

the finite alphabet $\mathcal{Y}$ is referred to as the *stegotext alphabet*. Let $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ denote a single–letter distortion measure between covertext symbols and stegotext symbols, and let the distortion between the vectors, $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$, be defined additively across the corresponding components, as usual.

An $(n, \lambda, D, R_c, h, \delta)$ code is a block code for joint watermarking, encryption, and compression with parameters n and λ that satisfies the following requirements:

1. The expected distortion between the covertext and the stegotext satisfies

$$\sum_{t=1}^{n} Ed(X_t, Y_t) \leq nD. \tag{1}$$

2. The entropy of the stegotext satisfies

$$H(Y^n) \leq nR_c. \tag{2}$$

3. The equivocation of the message source satisfies

$$H(U^N|Y^n) \geq Nh. \tag{3}$$

4. There exists a decoder $g_n : \mathcal{Y}^n \times \mathcal{K}^n \to \mathcal{U}^N$ such that

$$P_e \stackrel{\Delta}{=} \Pr\{g_n(Y^n, K^n) \neq U^N\} \leq \delta. \tag{4}$$

For a given λ, a triple (D, R_c, h) is said to be *achievable* if for every $\epsilon > 0$, there is a sufficiently large n for which $(n, \lambda, D + \epsilon, R_c + \epsilon, h - \epsilon, \epsilon)$ codes exist. The *achievable region* of triples (D, R_c, h) is the set of all achievable triples (D, R_c, h). It is assumed that $H(K) \leq \lambda H(U)$ as this upper limit on $H(K)$ suffices to achieve perfect secrecy.

Our first coding theorem is the following:

Theorem 1 *A triple (D, R_c, h) is achievable if and only if the following conditions are both satisfied:*

(a) $h \leq H(K)$.

(b) There exists a channel $\{P_{Y|X}(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ such that: (i) $H(Y|X) \geq \lambda H(U)$, (ii) $R_c \geq \lambda H(U) + I(X;Y)$, and (iii) $D \geq Ed(X,Y)$.

As can be seen, the encryption, on the one hand, and the embedding and the compression, on the other hand, do not interact at all in this theorem. There is a complete decoupling between them: While condition (a) refers solely to the key and the secrecy of the watermark, condition (b) is only about the embedding–compression part, and it is a replica of the conditions of the coding theorem in [8], where the role of the embedding rate, R_e (see Introduction above), is played by the product $\lambda H(U)$. This suggests a very simple separation principle, telling that in order to attain a given achievable triple (D, R_c, h), first compress the watermark U^N to its entropy, then encrypt Nh bits (out of the $NH(U)$) of the compressed bit–string (by bit–by–bit XORing with the same number of compressed key bits), and finally, embed this partially encrypted compressed bit–string into the covertext, using the coding theorem of [8] (again, see the Introduction above for a brief description of this).

4 Extensions to Steps 2 and 3

Moving on to Step 2, we now relax requirement no. 4 in the above definition of an $(n, \lambda, D, R_c, h, \delta)$ code, and allow a certain distortion between U^N and its reconstruction $\hat{U}^N$ at the legitimate decoder. More precisely, let $\hat{\mathcal{U}}$ denote a finite alphabet, henceforth referred to as the *message reconstruction alphabet*. Let $d' : \mathcal{U} \times \hat{\mathcal{U}} \to \mathbb{R}^+$ denote a single–letter distortion measure between message symbols and message reconstruction symbols, and let the distortion between vectors $u^N \in \mathcal{U}^N$ and $\hat{u}^N \in \hat{\mathcal{U}}^N$ be again, defined additively across the corresponding components. Finally, let $R_U(D')$ denote the rate–distortion function of the source P_U w.r.t. d', i.e.,

$$R_U(D') = \min\{I(U;\hat{U}) : \ Ed'(U,\hat{U}) \le D'\}. \tag{5}$$

It will now be assumed that $H(K) \le \lambda R_U(D')$, for the same reasoning as before.

Requirement no. 4 is now replaced by the following requirement: There exists a decoder $g_n : \mathcal{Y}^n \times \mathcal{K}^n \to \hat{\mathcal{U}}^N$ such that $\hat{U}^N = (\hat{U}_1, \ldots, \hat{U}_N) = g_n(Y^n, K^n)$ satisfies:

$$\sum_{i=1}^{N} Ed'(U_i, \hat{U}_i) \le ND'. \tag{6}$$

In addition to this modification of requirement no. 4, we add, to requirement no. 3, a specification regarding the minimum allowed equivocation w.r.t. the reconstructed message:

$$H(\hat{U}^N|Y^n) \ge Nh', \tag{7}$$

in order to guarantee that the secrecy of the reconstructed message is also secure enough. Accordingly, we modify the above definition of a block code as follows: An $(n, \lambda, D, D', R_c, h, h')$ code is a block code for joint watermarking, encryption, and compression with parameters n and λ that satisfies requirements 1–4, with the above modifications of requirements 3 and 4. For a given λ, a quintuple (D, D', R_c, h, h') is said to be *achievable* if for every $\epsilon > 0$, there is a sufficiently large n for which $(n, \lambda, D + \epsilon, D' + \epsilon, R_c + \epsilon, h - \epsilon, h' - \epsilon)$ codes exist.

Our second theorem extends Theorem 1 to this setting:

Theorem 2 *A quintuple (D, D', R_c, h, h') is achievable if and only if the following conditions are all satisfied:*

(a) $h \leq H(K)/\lambda + H(U) - R_U(D')$.

(b) $h' \leq H(K)/\lambda$.

(c) There exists a channel $\{P_{Y|X}(y|x),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ such that: (i) $\lambda R_U(D') \leq H(Y|X)$, (ii) $R_c \geq \lambda R_U(D') + I(X;Y)$, and (iii) $D \geq Ed(X,Y)$.

As can be seen, the passage from Theorem 1 to Theorem 2 includes the following modifications: In condition (c), $H(U)$ is simply replaced by $R_U(D')$ as expected. This means that the lossless compression code of U^N, in the achievability of Theorem 1, is now replaced by a rate–disortion code for distortion level D'. Conditions (a) and (b) now tell us that the key rate (in terms of entropy) should be sufficiently large to satisfy both equivocation requirements. Note that the condition regarding the equivocation w.r.t. the clean message source is softer than in Theorem 1 as $H(U) - R_U(D') \geq 0$. This is because the rate–distortion code for U^N already introduces an uncertainty of $H(U) - R_U(D')$ bits per symbol, and so, the encryption should only complete it to the desired level of h bits per symbol. This point is discussed in depth in [14]. Of course, by setting $D' = 0$ (and hence also $h' = h$), we are back to Theorem 1.

We also observe that the encryption and the embedding are still decoupled in Theorem 2, and that an achievable quintuple can still be attained by separation: First, apply a rate–distortion code to U^N, as mentioned earlier, then encrypt $N \cdot \max\{h + R_U(D') - H(U), h'\}$ bits of the compressed codeword (to satisfy both equivocation requirements), and finally, embed the (partially) encrypted codeword into X^n, again, by using the scheme of [8]. Note that without the encryption and without requirement no. 2 of the compressibility of Y^n,

this separation principle is a special case of the one in [10], where a separation theorem was established for the Wyner–Ziv source (with SI correlated to the source at the decoder) and the Gel'fand–Pinsker channel (with channel SI at the encoder). Here, there is no SI correlated to the source and the role of channel SI is fulfilled by the covertext. Thus, the new observation here is that the separation theorem continues to hold in the presence of encryption and requirement no. 2.

Finally, we turn to step 3, of including an attack channel (see Fig. 1). Let $\mathcal{Z}$ be a finite alphabet, henceforth referred to as the *forgery alphabet*, and let $\{P_{Z|Y}(z|y),\ y \in \mathcal{Y},\ z \in \mathcal{Z}\}$ denote a set of conditional PMF's from the stegotext alphabet to the forgery alphabet. We now assume that the stegotext vector is subjected to an attack modeled by the memoryless channel,

$$P_{Z^n|Y^n}(z^n|y^n) = \prod_{t=1}^{n} P_{Z|Y}(z_t|y_t). \tag{8}$$

The output Z^n of the attack channel will henceforth be referred to as the *forgery*.

It is now assumed and that the legitimate decoder has access to Z^n, rather than Y^n (in addition, of course, to K^n). Thus, in requirement no. 4, the decoder is redefined again, this time, as a mapping $g_n : \mathcal{Z}^n \times \mathcal{K}^n \to \hat{\mathcal{U}}^N$ such that $\hat{U}^N = g_n(Z^n, K^n)$ satisfies the distortion constraint (6). As for the equivocation requirements, the conditioning will now be on both Y^n and Z^n, i.e.,

$$H(U^N|Y^n, Z^n) \geq Nh \quad \text{and} \quad H(\hat{U}^N|Y^n, Z^n) \geq Nh', \tag{9}$$

as if the attacker and the eavesdropper are the same party (or if they cooperate), then s/he may access both. In fact, for the equivocation of U^N, the conditioning on Z^n is immaterial since $U^N \to Y^n \to Z^n$ is always a Markov chain, but it is not clear that Z^n is superfluous for the equivocation w.r.t. $\hat{U}^N$ since Z^n is one of the inputs to the decoder whose output is $\hat{U}^N$. Nonetheless, for the sake of uniformity and convenience (in the proof), we keep the conditioning on Z^n in both equivocation criteria.

Redefining block codes and achievable quintuples (D, D', R_C, h, h') according to the modified requirements in the same spirit, we now have the following coding theorem, which is substantially different from Theorems 1 and 2:

Theorem 3 *A quintuple (D, D', R_c, h, h') is achievable if and only if there exist RV's V and Y such that $P_{KXVYZ}(k, x, v, y, z) = P_X(x)P_K(k)P_{VY|KX}(v, y|k, x)P_{Z|Y}(z|y)$, where the*

alphabet size of V is bounded by $|\mathcal{V}| \leq |\mathcal{K}| \cdot |\mathcal{X}| \cdot |\mathcal{Y}| + 1$, and such that the following conditions are all satisfied:

(a) $h \leq H(K|Y)/\lambda + H(U) - R_U(D')$.

(b) $h' \leq H(K|Y)/\lambda$.

(c) $\lambda R_U(D') \leq I(V;Z|K) - I(V;X|K)$.

(d) $R_c \geq \lambda R_U(D') + I(X;Y,V|K) + I(K;Y)$.

(e) $D \geq Ed(X,Y)$.

First, observe that here, unlike in Theorems 1 and 2, it is no longer true that the encryption and the embedding/compression are decoupled. Note that now, although K is still assumed independent of X, it may, in general, depend on Y. On the negative side, this dependence causes a reduction in the equivocation of both the message source and its reconstruction, and therefore $H(K|Y)$ replaces $H(K)$ in conditions (a) and (b). On the positive side, on the other hand, this dependence introduces new degrees of freedom in enhancing the tradeoffs between the embedding performance (condition (c)) and the compressibility (condition (d)). At first glance, it may appear intuitive that the best choice of RV's would be to keep (V,Y) independent of K: The expression $I(V;Z|K) - I(V;X|K)$ should certainly be maximized for such a pair (V,Y) since K conveys irrelevant additional SI, due to the independence between X and K and the conditional independence between Z and K given Y. It is not clear, however, that such a choice of (V,Y) would also be best for the compressibility condition (d). In other words, due to the *combination* of requirements, the dependence of (V,Y) on K may be needed, in order to obtain full generality of performance tradeoffs, and so, K now may have the additional role of symbolizing SI that is available to *both* encoder and legitimate decoder. In this sense, there is *no* longer a separation principle, in contrast to the attack–free case.

The achievability of Theorem 3 involves essentially the same stages as before (rate–distortion coding of U^N, followed by encryption, followed in turn by embedding), but this time, the embedding scheme is a conditional version of the one proposed in [9], where all codebooks depend on K^n, the SI given at both ends (see Fig. 2). An interesting point regarding the encryption is that one needs to generate, from K^n, essentially $nH(K|Y)$

random bits that are *independent* of Y^n (and Z^n), in order to protect the secrecy against an eavesdropper that observes Y^n and Z^n. Clearly, if Y^n was given in advance to the encrypting unit, then the compressed bitstring of an optimal lossless source code that compresses K^n, given Y^n as SI, would have this property (as if there was any dependence, then this bitstring could have been further compressed, which is a contradiction). However, such a source code cannot be implemented since Y^n itself is in turn generated from the encrypted message, i.e., *after* the encryption. In other words, this would have required a circular mechanism, which may not be feasible. A simple remedy is then to use a *Slepian–Wolf encoder* [12], that generates $nH(K|Y)$ bits that are essentially independent of Y^n (due to the same consideration), without the need to access the vector Y^n to be generated. For more details, the reader is referred to the proof of the direct part (Section 6).

Observe that in the absence of attack (i.e., $Z = Y$), Theorem 2 is obtained as a special case of Theorem 3 by choosing $V = Y$ and letting both be independent of K, a choice which is simultaneously the best for conditions (a)–(d) of Theorem 3. To see this, note the following simple inequalities: In conditions (a) and (b), $H(K|Y) \leq H(K)$. In condition (c), by setting $Z = Y$, we have

$$\begin{aligned} I(V;Y|K) - I(V;X|K) &\leq I(V;X,Y|K) - I(V;X|K) \\ &= I(V;Y|X,K) \\ &\leq H(Y|X,K) \\ &\leq H(Y|X). \end{aligned} \tag{10}$$

Finally in condition (d), clearly, $I(K;Y) \geq 0$ and since X is independent of K, then $I(X;Y,V|K) = I(X;Y,V,K) \geq I(X;Y)$. Thus, for $Z = Y$, the achievable region of Theorem 3 is a subset of the one given in Theorem 2. However, since all these inequalities become equalites at the same time by choosing $V = Y$ and letting both be independent of K, the two regions are identical in the attack–free case.

Returning now to Theorem 3, as we observed, K^n is now involved not only in the role of a cipher key, but also as SI available at both encoder and decoder. Two important points are now in order, in view of this fact.

First, one may argue that, actually, there is no real reason to assume that K^n is necessarily independent of X^n (see also [11]). If the user has control of the mechanism of generating the key, then s/he might implement, in general, a channel $P_{K^n|X^n}(k^n|x^n)$ using

the available randomness resources, and taking (partial) advantage of the randomness of the covertext. Let us assume that this channel is stationary and memoryless, i.e.,

$$P_{K^n|X^n}(k^n|x^n) = \prod_{t=1}^{n} P_{K|X}(k_t|x_t) \tag{11}$$

with the single–letter transition probabilities $\{P_{K|X}(k|x)\ x \in \mathcal{X},\ k \in \mathcal{K}\}$ left as a degree of freedom for design. While so far, we assumed that K was independent of X, the other extreme is, of course, $K = X$ (corresponding to private watermarking). Note, however, that in the attack–free case, in the absence of the compressibility requirement no. 2 (say, $R_c = \infty$), no optimality is lost by assuming that K is independent of X, since the only inequality where we have used the independence assumption, in the previous paragraph, corresponds to condition (d).

The second point is that in Theorems 1–3, so far, we have defined the compressibility of the stegotext in terms of $H(Y^n)$, which is suitable when the decompression of Y^n is *public*, i.e., without access to K^n. The legitimate decoder in our model, on the other hand, has access to the SI K^n, which may depend on Y^n. In this context, it then makes sense to measure the compressibility of the stegotext also in a *private* regime, i.e., in terms of the *conditional* entropy, $H(Y^n|K^n)$.

Our last (and most general) version of the coding theorem below takes these two points in to account. Specifically, let us impose, in requirement no. 2, an additional inequality,

$$H(Y^n|K^n) \leq nR_c', \tag{12}$$

where R_c' is a prescribed constant, and let us redefine accordingly the block codes and the achievable region in terms of six–tuples $(D, D', R_c, R_c', h, h')$. We now have the following result:

Theorem 4 *A six–tuple $(D, D', R_c, R_c', h, h')$ is achievable if and only if there exist RV's V and Y such that $P_{KXVYZ}(k, x, v, y, z) = P_{XK}(x, k)P_{VY|KX}(v, y|k, x)P_{Z|Y}(z|y)$, where the alphabet size of V is bounded by $|\mathcal{V}| \leq |\mathcal{K}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|+1$, and such that the following conditions are all satisfied:*

(a) $h \leq H(K|Y)/\lambda + H(U) - R_U(D')$.

(b) $h' \leq H(K|Y)/\lambda$.

(c) $\lambda R_U(D') \leq I(V;Z|K) - I(V;X|K)$.

(d) $R_c \geq \lambda R_U(D') + I(X;Y,V|K) + I(K;Y)$.

(e) $R'_c \geq \lambda R_U(D') + I(X;Y,V|K)$.

(f) $D \geq Ed(X,Y)$.

Note that the additional condition, (e), is similar to condition (d) except for the term $I(K;Y)$. Also, in the joint PMF of (K,X,V,Y,Z) we are no longer assuming that K and X are independent. It should be pointed out that in the presence of the new requirement regarding $H(Y^n|K^n)$, it is more clear now that introducing dependence of (V,Y) upon K is reasonable, in general. In the case $K = X$, that was mentioned earlier, the term $I(V;X|K)$, in condition (c), and the term $I(X;Y,V|K)$, in conditions (e) and (f), both vanish. Thus, both embedding performance and compression performance improve, like in private watermarking.

5 Proof of the Converse Part of Theorem 4

Let an $(n, \lambda, D+\epsilon, D'+\epsilon, R_c+\epsilon, R'_c+\epsilon, h-\epsilon, h'-\epsilon)$ code be given. First, from the requirement $H(Y^n|K^n) \leq n(R'_c+\epsilon)$, we have:

$$\begin{aligned} n(R'_c+\epsilon) &\geq H(Y^n|K^n) && (13)\\ &= H(Y^n|U^N,K^n) + I(U^N;Y^n|K^n) \\ &\geq H(Y^n|U^N,K^n) + I(U^N;Z^n|K^n) \\ &= H(Y^n|U^N,K^n) + I(U^N;Z^n,K^n) && (14) \end{aligned}$$

where the second inequality comes from the data processing theorem ($U^N \to Y^n \to Z^n$ is a Markov chain given K^n) and the last equality comes from the chain rule and the fact that U^N and K^n are independent. Define $\tilde{V}_t = (X_{t+1}^n, U^N, K^{t-1}, Z^{t-1})$, J – as a uniform RV over $\{1,\ldots,n\}$, $X = X_J$, $K = K_J$, $Y = Y_J$, $V' = \tilde{V}_J$, and $V = (\tilde{V}_J, J) = (V', J)$. Now, the first term on the right–most side of eq. (14) is further lower bounded in the following

manner.

$$
\begin{aligned}
H(Y^n|U^N,K^n) &\geq I(X^n;Y^n|U^N,K^n) \\
&= I(X^n;Y^n,U^N,K^n) - I(X^n;U^N,K^n) \\
&= \sum_{t=1}^{n} I(X_t;Y^n,U^N,K^n|X_{t+1}^n) - I(X^n;K^n) \qquad (15)\\
&= \sum_{t=1}^{n} I(X_t;Y^n,U^N,K^n,X_{t+1}^n) - nI(X;K) \qquad (16)\\
&\geq \sum_{t=1}^{n} I(X_t;K_t,Y_t,U^N,K^{t-1},Z^{t-1},X_{t+1}^n) - nI(X;K) \qquad (17)\\
&= \sum_{t=1}^{n} I(X_t;K_t,Y_t,\tilde{V}_t) - nI(X;K) \\
&= n[I(X;K,Y,V'|J) - I(X;K)] \\
&= n[I(X;K,Y,V',J) - I(X;K)] \qquad (18)\\
&= nI(X;Y,V|K) \qquad (19)
\end{aligned}
$$

where (15) is due to the chain rule and fact that (X^n,K^n) is independent of U^N (hence $U^N \rightarrow K^n \rightarrow X^n$ is trivially a Markov chain), (16) is due to the memorylessness of $\{(X_t,K_t)\}$, (17) is due to the data processing theorem, and (18) follows from the fact that $\{X_t\}$ is stationary and so, $X = X_J$ is independent of J. The second term on the right–most side of eq. (14) is in turn lower bounded following essentially the same ideas as in the proof of the converse to the rate–distortion coding theorem (see, e.g., [3]):

$$
\begin{aligned}
I(U^N;Z^n,K^n) &= H(U^N) - H(U^N|Z^n,K^n) \\
&= \sum_{i=1}^{N}[H(U_i) - H(U_i|U^{i-1},Z^n,K^n)] \\
&= \sum_{i=1}^{N} I(U_i;U^{i-1},Z^n,K^n) \\
&\geq \sum_{i=1}^{N} I(U_i;[g_n(Z^n,K^n)]_i) \\
&\geq \sum_{i=1}^{N} R_U(Ed'(U_i,[g_n(Z^n,K^n)]_i)) \\
&\geq NR_U\left(\frac{1}{N}\sum_{i=1}^{N} Ed'(U_i,[g_n(Z^n,K^n)]_i)\right) \\
&\geq NR_U(D'+\epsilon), \qquad (20)
\end{aligned}
$$

where $[g_n(Z^n, K^n)]_i$ denotes the i-th component projection of $g_n(Z^n, K^n)$, i.e., $\hat{U}_i$ as a function of (Z^n, K^n). Combining eqs. (14), (19), and (20), we get

$$n(R_c' + \epsilon) \geq NR_U(D' + \epsilon) + nI(X; Y, V|K). \tag{21}$$

Dividing by n, we get

$$R_c' + \epsilon \geq \lambda R_U(D' + \epsilon) + I(X; Y, V|K). \tag{22}$$

Using the arbitrariness of ϵ together with the continuity of $R_U(\cdot)$, we get condition (e) of Theorem 4.

Condition (d) is derived in the very same manner except that the starting point is the inequality $n(R_c + \epsilon) \geq H(Y^n)$, and when $H(Y^n)$ is further bounded from below, in analogy to the chain of inequalities (14), there is an additional term, $I(K^n; Y^n)$, that is in turn lower bounded in the following manner:

$$\begin{aligned} I(K^n; Y^n) &\geq \sum_{t=1}^{n} I(K_t; Y_t) \\ &= nI(K; Y|J) \\ &= n[H(K|J) - H(K|J, Y)] \\ &\geq n[H(K) - H(K|Y)] \\ &= nI(K; Y), \end{aligned} \tag{23}$$

where the first inequality is because of the memorylessness of $\{K_t\}$, and the second inequality comes from the facts that conditioning reduces entropy (in the second term) and that K is independent of J (again, due to the stationarity of $\{K_t\}$). This gives the additional term, $I(K; Y)$, in condition (d).

Condition (c) is obtained as follows:

$$\begin{aligned} NR_U(D' + \epsilon) &\leq I(U^N; K^n, Z^n) \\ &= I(U^N; K^n, Z^n) - I(U^N; K^n, X^n) \\ &\leq \sum_{t=1}^{n} [I(\tilde{V}_t; K_t, Z_t) - I(\tilde{V}_t; K_t, X_t)] \qquad (24) \\ &= n[I(V'; K, Z|J) - I(V'; K, X|J)] \\ &\leq n[I(V', J; K, Z) - I(V', J; K, X)] \qquad (25) \\ &= n[I(V; K, Z) - I(V; K, X)] \\ &= n[I(V; Z|K) - I(V; X|K)], \qquad (26) \end{aligned}$$

where the first inequality is (20), the first equality is due to the independence between U^N and (K^n, X^n), the second inequality is an application of [5, Lemma 4], the third inequality is due to the fact that $I(K, Z; J) \geq 0$ and $I(K, X; J) = 0$ (due to the stationarity of $\{(K_t, X_t)\}$), and the last equality is obtained by adding and subtracting $I(V; K)$. Again, since this is true for every $\epsilon > 0$, it holds also for $\epsilon = 0$, due to continuity.

As for condition (f), we have:

$$D + \epsilon \geq \frac{1}{n} \sum_{t=1}^{n} Ed(X_t, Y_t) = Ed(X, Y), \tag{27}$$

and we use once again the arbitrariness of ϵ. Regarding condition (b), we have:

$$\begin{aligned}
nH(K|Y) &\geq nH(K|Y, J) \\
&= \sum_{t=1}^{n} H(K_t|Y_t) \\
&\geq \sum_{t=1}^{n} H(K_t|K^{t-1}, Y^n) \\
&= H(K^n|Y^n) \\
&= H(K^n|Y^n, Z^n) \\
&\geq I(K^n; \hat{U}^N|Y^n, Z^n) \\
&= H(\hat{U}^N|Y^n, Z^n) - H(\hat{U}^N|Y^n, Z^n, K^n) \\
&= H(\hat{U}^N|Y^n, Z^n) \\
&\geq N(h' - \epsilon),
\end{aligned} \tag{28}$$

where the last equality is due to the fact that $\hat{U}^N$ is, by definition, a function of (Z^n, K^n), and the last inequality is by the hypothesis that the code achieves an equivocation of at least $N(h' - \epsilon)$. Dividing by N and taking the limit $\epsilon \to 0$, leads to $h' \leq H(K|Y)/\lambda$, which is condition (b). Finally, to prove condition (a), consider the inequality $nH(K|Y) \geq H(\hat{U}^N|Y^n, Z^n)$, that we have just proved, and proceed as follows (see also [14]):

$$\begin{aligned}
nH(K|Y) &\geq H(\hat{U}^N|Y^n, Z^n) \\
&\geq H(\hat{U}^N|Y^n, Z^n) + N(h - \epsilon) - H(U^N|Y^n, Z^n) \\
&= N(h - \epsilon) - H(U^N) + I(U^N; Y^n, Z^n) - \\
&\quad I(\hat{U}^N; Y^n, Z^n) + I(\hat{U}^N; U^N) + H(\hat{U}^N|U^N) \\
&\geq N[h - \epsilon - H(U) + R_U(D' + \epsilon)] + \\
&\quad [I(U^N; Y^n, Z^n) - I(\hat{U}^N; Y^n, Z^n) + H(\hat{U}^N|U^N)],
\end{aligned} \tag{29}$$

where the second inequality follows from the hypothesis that the code satisfies $H(U^N|Y^n, Z^n) \geq N(h-\epsilon)$, and the third inequality is due to the memorylessness of $\{U_i\}$, the hypothesis that $\sum_{i=1}^N Ed'(U_i, \hat{U}_i) \leq N(D'+\epsilon)$, and the converse to the rate–distortion coding theorem. Now, to see that the second bracketted term is non–negative, we have the following chain of inequalities:

$$\begin{aligned} & I(U^N;Y^n,Z^n) - I(\hat{U}^N;Y^n,Z^n) + H(\hat{U}^N|U^N) \\ = \; & I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n|\hat{U}^N) + H(\hat{U}^N|U^N) \\ \geq \; & I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n|U^N,\hat{U}^N) + H(\hat{U}^N|U^N) \\ = \; & I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n,\hat{U}^N|U^N) \\ \geq \; & I(U^N;Y^n,Z^n) - H(Y^n,Z^n) + H(Y^n,Z^n|U^N) \\ = \; & 0. \end{aligned} \tag{30}$$

Combining this with eq. (29), we have

$$nH(K|Y) \geq N[h - \epsilon - H(U) + R_U(D'+\epsilon)]. \tag{31}$$

Dividing again by N, and letting ϵ vanish, we obtain $h \leq H(K|Y)/\lambda + H(U) - R_U(D')$, which completes the proof of condition (a).

To complete the proof of the converse part, it remains to show that the alphabet size of V can be reduced to $|\mathcal{K}| \cdot |\mathcal{X}| \cdot |\mathcal{Y}| + 1$. To this end, we extend the proof of the parallel argument in [9] by using the support lemma (cf. [4]), which is based on Carathéodory's theorem. According to this lemma, given J real valued continuous functionals f_j, $j = 1, ..., J$ on the set $\mathcal{P}(\mathcal{X})$ of probability distributions over the alphabets $\mathcal{X}$, and given any probability measure μ on the Borel σ-algebra of $\mathcal{P}(\mathcal{X})$, there exist J elements $Q_1, ..., Q_J$ of $\mathcal{P}(\mathcal{X})$ and J non-negative reals, $\alpha_1, ..., \alpha_J$, such that $\sum_{j=1}^J \alpha_j = 1$ and for every $j = 1, ..., J$

$$\int_{\mathcal{P}(\mathcal{X})} f_j(Q)\mu(dQ) = \sum_{i=1}^J \alpha_i f_j(Q_i). \tag{32}$$

Before we actually apply the support lemma, we first rewrite the relevant mutual informations of Theorem 4 in a more convenient form for the use of this lemma. First, observe that

$$\begin{aligned} I(V;Z|K) - I(V;X|K) &= H(Z|K) - H(Z|V,K) - H(X|K) + H(X|V,K) \\ &= H(Z|K) - H(X|K) + H(K,X|V) - H(K,Z|V). \end{aligned} \tag{33}$$

and

$$\begin{aligned} I(X;Y,V|K) &= I(X;V|K) + I(X;Y|V,K) \quad (34)\\ &= H(X|K) - H(X|V,K) + H(X|V,K) - H(X|V,Y,K)\\ &= H(X|K) - H(X|V,Y,K)\\ &= H(X|K) - H(K,X,Y|V) + H(K,Y|V). \quad (35)\end{aligned}$$

For a given joint distribution of (K,X,Y), and given $P_{Z|Y}$, $H(Z|K)$ and $H(X|K)$ are both given and unaffected by V. Therefore, in order to preserve prescribed values of $I(V;Z|K) - I(V;X|K)$ and $I(X;V,Y|K)$, it is sufficient the preserve the associated values $H(K,X|V) - H(K,Z|V)$ and $H(K,X,Y|V) - H(K,Y|V)$. Let us define then the following functionals of a generic distribution Q over $\mathcal{K}\times\mathcal{X}\times\mathcal{Y}$, where $\mathcal{K}\times\mathcal{X}\times\mathcal{Y}$ is assumed, without loss of generality, to be $\{1,2,...,m\}$, $m = |\mathcal{K}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|$:

$$f_i(Q) = Q(k,x,y), \quad i \stackrel{\Delta}{=} (k,x,y) = 1,...,m-1 \quad (36)$$

$$f_m(Q) = \sum_{k,x,y} Q(k,x,y) \sum_z P_{Z|Y}(z|y) \log \frac{\sum_{x,y} Q(k,x,y) P_{Z|Y}(z|y)}{Q(k,x)}. \quad (37)$$

Next define

$$f_{m+1}(Q) = \sum_{k,x,y} Q(k,x,y) \log \frac{Q(k,y)}{Q(k,x,y)}. \quad (38)$$

Applying now the support lemma, we find that there exists a random variable V (jointly distributed with (K,X,Y)), whose alphabet size is $|\mathcal{V}| = m+1 = |\mathcal{K}|\cdot|\mathcal{X}|\cdot|\mathcal{Y}|+1$ and it satisfies simultaneously:

$$\sum_v \Pr\{V=v\} f_i(P(\cdot|v)) = P_{KXY}(k,x,y), \ \ i = 1,...,m-1, \quad (39)$$

$$\sum_v \Pr\{V=v\} f_m(P(\cdot|v)) = H(K,X|V) - H(K,Z|V), \quad (40)$$

and

$$\sum_u \Pr\{V=v\} f_{m+1}(P(\cdot|v)) = H(K,X,Y|V) - H(K,Y|V). \quad (41)$$

It should be pointed out that this random variable maintains the prescribed distortion level $Ed(X,Y)$ since P_{XY} is preserved. By the same token, $H(K|Y)$ and $I(K;Y)$, which depend only on P_{KY}, are preserved as well. This completes the proof of the converse part of Theorem 4.

6 Proof of the Direct Part of Theorem 4

In this section, we show that if there exist RV's (V, Y) that satisfy the conditions of Theorem 4, then for every $\epsilon > 0$, there is a sufficiently large n for which $(n, \lambda, D+\epsilon, D'+\epsilon, R_c+\epsilon, R'_c+\epsilon, h-\epsilon, h'-\epsilon)$ codes exist. The main core of the proof is strongly based on a straightforward extension of the proof of the direct part of [9] to the case of additional SI present at both and decoder. Nonetheless, for the sake of completeness, the full details are provided here. It should be pointed out that for the attack–free case, an analogous extension can easily be offered to the direct part of [8].

We first digress to establish some additional notation conventions associated with the method of types [4]. For a given generic finite–alphabet random variable (RV) $A \in \mathcal{A}$ (or a vector of RV's taking on values in $\mathcal{A}$), and a vector $a^\ell \in \mathcal{A}^l$ (ℓ – positive integer), the empirical probability mass function (EPMF) is a vector $P_{a^\ell} = \{P_{a^l}(a'),\ a' \in \mathcal{A}\}$, where $P_{a^\ell}(a')$ is the relative frequency of the letter $a' \in \mathcal{A}$ in the vector a^ℓ. Given $\delta > 0$, let us denote the set of all δ-typical sequences of length ℓ by $T^\delta_{P_A}$, or by T^δ_A (if there is no ambiguity regarding the PMF that governs A), i.e., T^δ_A is the set of the sequences $a^\ell \in \mathcal{A}^\ell$ such that

$$(1-\delta)P_A(a') \leq P_{a^\ell}(a') \leq (1+\delta)P_A(a') \tag{42}$$

for every $a' \in \mathcal{A}$. For sufficiently large ℓ, the size of $T^\delta_{P_A}$ is well–known [4] to be bounded by

$$2^{\ell[(1-\delta)H(A)-\delta]} \leq |T^\delta_{P_A}| \leq 2^{\ell(1+\delta)H(A)}. \tag{43}$$

It is also well–known (by the weak law of large numbers) that:

$$\Pr\left\{A^\ell \notin T^\delta_A\right\} \leq \delta \tag{44}$$

for all ℓ sufficiently large. For a given generic channel $P_{B|A}(b|a)$ and for each $a^\ell \in T^\delta_A$, the set of all sequences b^l that are jointly δ-typical with a^ℓ, will be denoted by $T^\delta_{P_{B|A}}(a^\ell)$, or by $T^\delta_{B|A}(a^\ell)$ if there is no ambiguity, i.e., $T^\delta_{B|A}(a^\ell)$ is the set of all b^ℓ such that:

$$(1-\delta)P_{a^\ell}(a')P_{B|A}(b'|a') \leq P_{a^\ell b^\ell}(a', b') \leq (1+\delta)P_{a^\ell}(a')P_{B|A}(b'|a'), \tag{45}$$

for all $a' \in \mathcal{A}, b' \in \mathcal{B}$, where $P_{a^\ell b^\ell}(a', b')$ denotes the fraction of occurrences of the pair (a', b') in (a^ℓ, b^ℓ). Similarly as in eq. (42), for all sufficiently large ℓ and $a^\ell \in T^\delta_A$, the size of

$T^{\delta}_{B|A}(a^{\ell})$ is bounded as follows:

$$2^{\ell[(1-\delta)H(B|A)-\delta]} \leq |T^{\delta}_{B|A}(a^{\ell})| \leq 2^{\ell(1+\delta)H(B|A)}. \tag{46}$$

Finally, observe that for all $a^{\ell} \in T^{\delta}_{A}$ and $b^{\ell} \in T_{B|A}(b^{\ell})$, the distortion $d(a^{\ell}, b^{\ell}) = \sum_{j=1}^{\ell} d(a_j, b_j)$ is upper bounded by:

$$d(a^{\ell}, b^{\ell}) \leq \ell(1+\delta)^2 \sum_{a',b'} P_A P_{B|A}(b'|a') d(a', b') \stackrel{\Delta}{=} \ell(1+\delta)^2 E d(A, B). \tag{47}$$

Let (K, X, V, Y, Z) be a given random vector that satisfies the conditions of Theorem 4. We now describe the mechanisms of random code selection and the encoding and decoding operations. Fix δ such that $2\delta + \max\{2 \cdot \exp^{-2^{n\delta}} + 2^{-n\delta}, \delta^2\} \leq \epsilon$.

Generation of a rate–distortion code:
Apply the type–covering lemma [4] and construct a rate–distortion codebook that covers T^{δ}_{U} within distortion $N(D'+\epsilon)$ w.r.t. d', using $2^{NR_U(D')}$ codewords.

Generation of the encrypting bitstream:
For every $k^n \in T^{\delta}_{K}$, randomly select an index in the set $\{0, 1, \ldots, 2^{n[H(K|Y)+\delta]}-1\}$ with a uniform distribution. Denote by $s^J(k^n) = (s_1(k^n), \ldots, s_J(k^n))$, $s_j(k^n) \in \{0,1\}$, $j = 1, \ldots, J$, the binary string of length $J = n[H(K|Y)+\delta]$ that represents this index. (Note that $s^J(k^n)$ can be interpreted as the output of the Slepian–Wolf encoder for K^n, where Y^n plays the role of SI at the decoder [12].)

Generation of an auxiliary embedding code:
We first construct an auxiliary code capable of embedding $2^{NR_U(D')}$ watermarks by a random selection technique. First, $M_1 = 2^{nR_1}$, $R_1 \leq I(V;Z|K) - \epsilon_3 - \delta$, sequences $\{V^n(i, k^n)\}$, $i \in \{1, \ldots, M_1\}$, are drawn independently from $T^{\delta}_{V|K}(k^n)$ for every $k^n \in T^{\delta}_{K}$. For every such k^n, let us denote the set of these sequences by $\mathcal{C}(k^n)$. The elements of $\mathcal{C}(k^n)$ are evenly distributed among $M_U \stackrel{\Delta}{=} 2^{NR_U(D')}$ bins, each of size $M_2 = 2^{nR_2}$, $R_2 \geq I(X;V|K) + \epsilon_1 + \delta$ (this is possible thanks to condition (c) of Theorem 4, provided that the inequality therein is strict). A different (encrypted) message of length $L = NR_U(D') = n\lambda R_U(D')$ bits is attached to each bin, identifying a subcode that represents this message. We denote the codewords in bin number m ($m \in \{1, 2, \ldots, M_U\}$), by $\{V^n(m, j, k^n)\}$, $j \in \{1, 2, \ldots, M_2\}$.

Stegotext sequence generation:

For each auxiliary sequence (in the above auxiliary codebook of each δ–typical k^n), $V^n(m,j,k^n) = v^n$, a set of $M_3 \stackrel{\Delta}{=} 2^{nR_3}$, $R_3 \geq I(X;Y|V,K) + \epsilon_2 + \delta$, stegotext sequences $\{Y^n(j',v^n,k^n)\}$, $j' \in \{1,\ldots,M_3\}$, are independently drawn from $T^{\delta}_{Y|VK}(v^n,k^n)$. We denote this set by $\mathcal{C}(v^n,k^n)$.

Encoding:

Upon receiving a triple (u^N,x^n,k^n), the encoder acts as follows:

1. If $u^N \in T^{\delta}_U$, let $w^L = (w_1,\ldots,w_L)$, $w_i \in \{0,1\}$, $i = 1,\ldots,L$ be the binary representation of the index of the rate–distortion codeword for the message source. For $k^n \in T^{\delta}_K$, let $s^J(k^n) = (s_1(k^n),\ldots,s_J(k^n))$ denote binary representation string of the index of k^n. Let $\tilde{w}^L = (\tilde{w}_1,\ldots,\tilde{w}_L)$, where $\tilde{w}_j = w_j \oplus s_j(k^n)$, $j = 1,\ldots,J$, and $\tilde{w}_j = w_j$, $j = J+1,\ldots,L$, and where $\oplus$ denotes modulo 2 addition i.e., the XOR operation.[3] The binary vector $\tilde{w}^L$ is the (partially encrypted) message to be embedded. Let $m = \sum_{l=1}^{L} \tilde{w}_l 2^{l-1} + 1$ denote the index of this message. If $u^N \notin T^{\delta}_U$ or $k^n \notin T^{\delta}_K$, an arbitrary (error) message $\tilde{w}^L$ is generated (say, the all–zero message).

2. If $(k^n,x^n) \in T^{\delta}_{KX}$ find, in bin number m, the first j such that $V^n(m,j,k^n) = v^n$ is jointly typical, i.e., $(k^n,x^n,v^n) \in T^{\delta}_{KXV}$, and then find the first j' such that $Y^n(j',v^n,k^n) = y^n \in \mathcal{C}(v^n,k^n)$ is jointly typical, i.e., $(k^n,x^n,v^n,y^n) \in T^{\delta}_{KXVY}$. This vector y^n is chosen for transmission. If $(k^n,x^n) \notin T^{\delta}_{KX}$, or if there is no $V^n(m,j,k^n) = v^n$ and $Y^n(j',v^n,k^n) = y^n$ such that $(k^n,x^n,v^n,y^n) \in T^{\delta}_{KXVY}$, an arbitrary vector $y^n \in \mathcal{Y}^n$ is transmitted.

Decoding:

Upon receiving $Z^n = z^n$ and $K^n = k^n$, the decoder finds all sequences $\{v^n\}$ in $\mathcal{C}(k^n)$ such that $(k^n,v^n,z^n) \in T^{\delta}_{KVZ}$. If all $\{v^n\}$ found belong to the same bin, say, $\hat{m}$, then $\hat{m}$ is decoded as the embedded message, and then the binary representation vector $\hat{w}^L = (\hat{w}_1,\ldots,\hat{w}_L)$ corresponding to $\hat{m}$ is decrypted, again, by modulo 2 addition of its first J bits with $s(k^n)$. This decrypted binary L–vector is then mapped to the corresponding reproduction vector $\tilde{u}^N$ of the rate–distortion codebook for the message source. If there is no $v^n \in \mathcal{C}(k^n)$ such

[3]Note that since $H(K)$ is assumed smaller than $\lambda R_U(D')$, then so is $H(K|Y)$, and therefore $J \leq L$.

that $(k^n, v^n, z^n) \in T^\delta_{KVZ}$ or if there exist two or more bins that contain such a sequence, an error is declared.

We now turn to the performance analysis of this code in all relevant aspects. For each triple (k^n, x^n, u^N) and particular choices of the codes, the possible causes for incorrect watermark decoding are the following:

1. $(k^n, x^n, u^N) \notin T^\delta_{KX} \times T^\delta_U$. Let the probability of this event be defined as P_{e_1}.

2. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$, but in bin no. m there is no v^n s.t. $(k^n, x^n, v^n) \in T^\delta_{KXV}$. Let the probability of this event be defined as P_{e_2}.

3. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$ and in bin no. m there is v^n s.t. $(k^n, x^n, v^n) \in T^\delta_{KXV}$, but there is no $y^n \in \mathcal{C}(v^n, k^n)$ s.t. $(k^n, x^n, v^n, y^n) \in T^\delta_{KXVY}$. Let the probability of this event be defined as P_{e_3}.

4. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$ and in bin no. m there is v^n and $y^n \in \mathcal{C}(v^n, k^n)$ such that $(k^n, x^n, v^n, y^n) \in T^\delta_{KXVY}$, but $(k^n, v^n, z^n) \notin T^\delta_{KVZ}$. Let the probability of this event be defined as P_{e_4}.

5. $(k^n, x^n, u^N) \in T^\delta_{KX} \times T^\delta_U$ and in bin no. m there is v^n and $y^n \in \mathcal{C}(v^n, k^n)$ such that $(k^n, x^n, v^n, y^n) \in T^\delta_{KXVY}$, and $(k^n, v^n, z^n) \in T^\delta_{KVZ}$, but there exists another bin, say, no. $\tilde{m}$, that contains $\tilde{v}^n$ s.t. $(k^n, \tilde{v}^n, z^n) \in T^\delta_{KVZ}$. Let the probability of this event be defined as P_{e_5}.

If none of these events occur, the message $\tilde{w}^L$ (or, equivalently, m) is decoded correctly from z^n, the distortion constraint between x^n and y^n is within $n(D+\epsilon)$ (as follows from (47)), and the distortion between u^N and its rate–distortion codeword, $\tilde{u}^N = \hat{u}^N$, does not exceed $N(D'+\epsilon)$. Thus, requirements 1 and 4 (modified according to eq. (6), with $D'+\epsilon$ replacing D') are both satisfied. Therefore, we first prove that the probability for none of the events 1–5 to occur, tends to unity as $n \to \infty$.

The average probability of error P_e in decoding m is bounded by

$$P_e \leq \sum_{i=1}^{5} P_{e_i}. \tag{48}$$

The fact that $P_{e_1} \to 0$ follows immediately from (44). As for P_{e_2}, we have:

$$P_{e_2} \stackrel{\Delta}{=} \prod_{j=1}^{M_2} \Pr\{(k^n, x^n, V^n(m,j,k^n)) \notin T^{\delta}_{KXV}\}. \tag{49}$$

Now, by (43), for every j and every $(k^n, x^n) \in T^{\delta}_{KX}$:

$$\begin{aligned} \Pr\{V^n(m,j,k^n) \notin T^{\delta}_{V|KX}(k^n,x^n)\} &= 1 - \Pr\{V^n(m,j,k^n) \in T^{\delta}_{V|KX}(k^n,x^n)\} \\ &= 1 - \frac{|T^{\delta}_{V|KX}(k^n,x^n)|}{|T^{\delta}_{V|K}(k^n)|} \\ &\leq 1 - \frac{2^{n[(1-\delta)H(V|K,X)-\delta]}}{2^{n(1+\delta)H(V|K)}} \\ &= 1 - 2^{-n[I(X;V|K)+\epsilon_1]}, \end{aligned} \tag{50}$$

where

$$\epsilon_1 \stackrel{\Delta}{=} \delta[1 + H(V|K) + H(V|K,X)]. \tag{51}$$

Substitution of (50) into (49) provides us with the following upper-bound:

$$P_{e_2} \leq \left[1 - 2^{-n[I(X;V|K)+\epsilon_1]}\right]^{M_2} \leq \exp\left\{ - 2^{nR_2} \cdot 2^{-n[I(X;V|K)+\epsilon_1]} \right\} \to 0, \tag{52}$$

double–exponentially rapidly since $R_2 \geq I(X;V|K) + \epsilon_1 + \delta$. To estimate P_{e_3}, we repeat the same technique:

$$P_{e_3} \stackrel{\Delta}{=} \prod_{j'=1}^{M_3} \Pr\{(k^n, x^n, v^n, Y^n(j', v^n, k^n)) \notin T^{\delta}_{KXVY}\}. \tag{53}$$

Again, by the property of the typical sequences, for every j' and $(k^n, x^n, v^n) \in T^{\delta}_{KXV}$:

$$\Pr\{Y^n(j', v^n, k^n) \notin T^{\delta}_{Y|KXV}(k^n, x^n, v^n)\} \leq 1 - 2^{-n[I(X;Y|V,K)+\epsilon_2]}, \tag{54}$$

where

$$\epsilon_2 = \delta[1 + H(Y|K,V) + H(Y|K,X,V)], \tag{55}$$

and therefore, substitution of (54) into (53) gives

$$P_{e_3} \leq \left[1 - 2^{-n[I(X;Y|V,K)+\epsilon_2]}\right]^{M_3} \leq \exp\left\{ - 2^{nR_3} \cdot 2^{-n[I(X;Y|V,K)+\epsilon_2]} \right\} \to 0, \tag{56}$$

double–exponentially rapidly since $R_2 \geq I(X;Y|V,K) + \epsilon_2 + \delta$. The estimation of P_{e_4} is again based on properties of typical sequences. Since Z^n is the output of a memoryless channel

$P_{Z|Y}$ with input $y^n = Y^n(j', v^n, k^n)$ and by the assumption of this step $(k^n, x^n, v^n, y^n) \in T^{\delta}_{KXVY}$, from (44) and the Markov lemma [3, Lemma 14.8.1], we obtain

$$P_{e_4} = \Pr\{(k^n, x^n, v^n, y^n, Z^n) \notin T^{\delta}_{KXVYZ}\} \leq \delta, \tag{57}$$

and similarly to P_{e_1} can be made as small as desired by an appropriate choice of δ.

Finally, we estimate P_{e_5} as follows:

$$\begin{aligned} P_{e_5} &= \Pr\{\exists \tilde{m} \neq m : (k^n, V^n(\tilde{m}, j, k^n), z^n) \in T^{\delta}_{KVZ}\} && (58)\\ &\leq \sum_{\tilde{m}\neq m,\ j\in\{1,2,\ldots,M_1\}} \Pr\{(k^n, V^n(\tilde{m}, j, k^n), z^n) \in T^{\delta}_{KVZ}\} \\ &\leq (2^{NR_U(D')} - 1)2^{nR_2} \Pr\{(k^n, V^n(\tilde{m}, j, k^n), z^n) \in T^{\delta}_{KVZ}\} \\ &\leq 2^{nR_1} 2^{-n[I(V;Z|K)-\epsilon_3]}, && (59) \end{aligned}$$

where

$$\epsilon_3 = \delta[1 + H(V|K) + H(V|Z, K)]. \tag{60}$$

Now, since $R_1 \leq I(V; Z|K) - \epsilon_3 - \delta$, $P_{e_5} \to 0$. Since $P_{e_i} \to 0$ for $i = 1, \ldots, 5$, their sum tends to zero as well, implying that there exist at least one choice of an auxiliary code and related stegotext codes that give rise to the reliable decoding of $\tilde{W}^L$.

Now, let us denote by N_c the total number of composite sequences in a codebook that corresponds to a δ–typical k^n. Then,

$$\begin{aligned} N_c &= M_U \cdot M_2 \cdot M_3 \\ &\leq 2^{n[\lambda R_U(D') + I(X;V|K) + I(X;Y|V,K) + \epsilon_1 + \epsilon_2]} \\ &= 2^{n[\lambda R_U(D') + I(X;Y,V|K) + \epsilon_1 + \epsilon_2]}. && (61) \end{aligned}$$

Thus,

$$\begin{aligned} H(Y^n|K^n) &\leq \log N_c \\ &= n[\lambda R_U(D') + I(X; Y, V|K) + \epsilon_1 + \epsilon_2] \\ &\leq n(R'_c + \epsilon_1 + \epsilon_2), && (62) \end{aligned}$$

where in the last inequality we have used condition (e). For sufficiently small values of δ, $\epsilon_1 + \epsilon_2 \leq \epsilon$ and so, the compressibility requirement in the presence of K^n is satisfied.

It remains to prove the achievability of R_c and of the equivocation levels, h and h'. Although these parts of the proof are not difficult, their exact descriptions are lengthy and tedious. Therefore, we give here the outline only.

The achievability of R_c can be shown as follows: Suppose that we have already randomly selected a codebook for one representative member $\hat{k}^n$ of each type class $T^0_{Q_K} \subset T^\delta_K$ using the mechanism described above.[4] Now, consider the set of all permutations from $\hat{k}^n$ to every other member of $T^0_{Q_K}$. The auxiliary codebook and the stegotext codebooks for every other key sequence, $k^n \in T^0_{Q_K}$, will be obtained by permuting all (auxiliary and stegotext) codewords of those corresponding to $\hat{k}^n$ according to the same permutation that leads from $\hat{k}^n$ to k^n (thus preserving all the necessary joint typicality properties). Now, in the *union* of all stegotext codebooks, corresponding to all typical key sequences, each codeword will appear exponentially[5] at least $2^{nH(K|Y)}$ times, which is the number of permutations of $\hat{k}^n$ which leave a given stegotext codeword y^n unaltered. Since the total number of codewords for all key sequences (including repetitions) is exponentially $2^{nH(K)} \cdot 2^{n[\lambda R_U(D')+I(X;Y,V|K)]}$, and each codeword appears altogether exponentially at least $2^{nH(K|Y)}$ times, the number of *distinct* codewords in the union of all codebooks is exponentially at most:

$$\frac{2^{nH(K)} \cdot 2^{n[\lambda R_U(D')+I(X;Y,V|K)]}}{2^{nH(K|Y)}} = 2^{n[\lambda R_U(D')+I(X;Y,V|K)+I(K;Y)]}. \tag{63}$$

Thus, the number of bits required for public compression of Y^n without the key (and hence also $H(Y^n)$), using the union of all stegotext codebooks corresponding to all $k^n \in T^\delta_K$, is essentially upper bounded by the logarithm of the right–hand side of eq. (63), i.e., by $n[\lambda R_U(D_1)+I(X;Y,V|K)+I(K;Y)]$, which in turn is upper bounded by nR_c, by condition (d) of Theorem 4. Thus, the public compressibility requirement is satisfied as well.

Before we proceed to evaluate the equivocation levels, an important comment is in order in the context of public compression (and a similar comment will apply to private compression): Note that a straightforward (and not necessary optimal) method for public compression of Y^n is simply according to its index within T^δ_Y, which requires about $nH(Y)$ bits. On the other hand, the converse theorem tells us that the compressed representation of Y^n cannot be much shorter than $n[\lambda R_U(D') + I(X;Y,V|K) + I(K;Y)]$ bits (cf. the

[4] In this context, $T^0_{Q_K}$ means $T^\delta_{Q_K}$ with $\delta = 0$, with Q_K exhausting all PMF's that are within δ close to P_K, in accordance to the definition of δ–types.

[5] Here and below, we use the the term "exponentially" in the sense that the exponential rates of the expressions under discussion are asymptotically exact as n grows without bound and expressions that vanish as $\delta \to 0$ can be neglected.

necessity of condition (d) of Theorem 4). Thus, contradiction between these two facts is avoided only if

$$\lambda R_U(D') + I(X;Y,V|K) + I(K;Y) \leq H(Y), \tag{64}$$

or, equivalently,

$$\lambda R_U(D') + I(X;Y,V|K) \leq H(Y|K). \tag{65}$$

This means that any achievable point $(D, D', R_c, R_c', h, h')$ corresponds to a choice of random variables (K, X, Y, V) that must inherently satisfy eq. (65). This observation will now help us also in estimating the equivocation levels.

Consider first the equivocation w.r.t. the reproduction, for which we have the following chain of inequalities:

$$\begin{aligned}
Nh' &\leq nH(K|Y) && (66)\\
&= nH(K) - nI(K;Y)\\
&= H(K^n) - nI(K;Y)\\
&= H(K^n|Y^n,Z^n) + I(K^n;Y^n,Z^n) - nI(K;Y)\\
&= H(K^n|Y^n,Z^n) + I(K^n;Y^n) - nI(K;Y) && (67)\\
&= H(K^n|Y^n,Z^n) + H(Y^n) - H(Y^n|K^n) - nI(K;Y)\\
&\leq H(K^n|Y^n,Z^n) + n[\lambda R_U(D') + I(X;Y,V|K) + I(K;Y) + \epsilon] -\\
&\quad -n[\lambda R_U(D'+\epsilon) + I(X;Y,V|K) - \epsilon] - nI(K;Y)\\
&= H(K^n|Y^n,Z^n) + n\lambda[R_U(D') - R_U(D'+\epsilon)] + n\epsilon\\
&\stackrel{\Delta}{=} H(K^n|Y^n,Z^n) + n\epsilon'\\
&= I(K^n;\hat{U}^N|Y^n,Z^n) + H(K^n|Y^n,Z^n,\hat{U}^N) + n\epsilon'\\
&\leq H(\hat{U}^N|Y^n,Z^n) + H(K^n|Y^n,Z^n,\hat{U}^N) + n\epsilon', && (68)
\end{aligned}$$

where (66) is based on condition (b), (67) follows from the fact that $K^n \to Y^n \to Z^n$ is a Markov chain, (68) is due to the sufficiency of condition (d) (that we have just proved) and the necessity of condition (e), and ϵ' vanishes as $\epsilon \to 0$ due to the continuity of $R_U(\cdot)$. Comparing the left–most side and the right–most side of the above chain of inequalities, we see that to prove that $H(\hat{U}^N|Y^n,Z^n)$ is essentially at least as large as Nh', it remains to show that $H(K^n|Y^n,Z^n,\hat{U}^N)$ is small, say, less than $n\epsilon'$ for large n. This will be the case

if we show that given $(Y^n, Z^n, \hat{U}^N)$, we can actually find K^n with high probability, or more precisely, that the number of possible candidates of K^n, in the presence of $(Y^n, Z^n, \hat{U}^N)$, is sub–exponential in n. Now, let us suppose that the inequality in (65) is strict (otherwise, we can slightly increase the allowable distortion level D' and thus reduce $R_U(D')$). Now, as described earlier, for a given typical k^n, we randomly select $2^{n[\lambda R_U(D')+I(X;Y,V|K)]}$ stegotext vectors in $T^{\delta}_{Y|K}$. Let us first show that the probability of having more than one occurrence of a specific sequence y^n in the stegotext codebook of k^n, tends to zero as $n \to \infty$. The probability of obtaining y^n in a single random selection is given by

$$\Pr\{Y^n(j', V^n(m, j, k^n), k^n) = y^n\} = \frac{|T^{\delta}_{V|KY}(k^n, y^n)|}{|T^{\delta}_{V|K}(k^n)|} \cdot \frac{1}{|T^{\delta}_{Y|KV}(k^n, v^n)|} \tag{69}$$

where the first factor is the probability of having a $V^n(m, j, k^n) = v^n$ that is typical with y^n and k^n (a necessary condition for this v^n to generate the given y^n), and the second factor is the probability of selecting a given y^n in the random selection of the steogtext code. The exponential order of (69) is $2^{-nH(Y|K)}$. Since the total size of the stegotext codebook for k^n is of the exponential order of $2^{n[\lambda R_U(D')+I(X;Y,V|K)]}$, which is *less* than $2^{nH(Y|K)}$, the probability of the union of the events $\cup_{m,j,j'}\{Y^n(j', V^n(m, j, k^n), k^n) = y^n\}$ still vanishes. Now, as k^n exhausts $T^0_{Q_{K|Y}}(y^n)$ (and the codewords undergo the same permutations as k^n), the permuted stegotext codebooks have the same vector y^n at the same location (because these permutations leave y^n unaltered as explained above) and hence the same bin number, in the stegotext codebook. Other key sequences within T^{δ}_K cannot include y^n since they are not jointly typical with y^n. It follows then that y^n appears essentially only once and in the same bin in each codebook corresponding to $T_{K|Y}(y^n)$ and only in those codebooks.[6] Thus, Y^n alone essentially uniquely determines $\tilde{W}^L$. Now, by XORing the first J bits of $\tilde{W}^L$ with those of W^L (which in turn is a function of the given $\hat{U}^N$), one obtains $S^J(K^n)$, from which with high probability, we can reconstruct K^n in the presence of Y^n, using the Slepian–Wolf decoder.

Finally, for the equivocation w.r.t. the original message source, we have an additional equivocation of $N[H(U) - R_U(D')]$ bits that are missing from the complete (lossless) description of U^N, and which are essentially independent of the $NR_U(D')$ bits that are present

[6]Note that the union of all $T^0_{Q_K}$ within T^{δ}_K may give at most a polynomial multiplicity of y^n in all codebooks.

(again, due to compressibility considerations). More specifically, we have the following:

$$\begin{aligned} H(U^N|Y^n,Z^n) &= H(\hat{U}^N|Y^n,Z^n)+H(U^N|Y^n,Z^n)-H(\hat{U}^N|Y^n,Z^n) \\ &\geq nH(K|Y)-2n\epsilon'+H(U^N|Y^n,Z^n)-H(\hat{U}^N|Y^n,Z^n) \\ &= nH(K|Y)+H(U^N)-I(U^N;\hat{U}^N)-I(U^N;Y^n,Z^n)- \\ &\quad H(\hat{U}^N|U^N)+I(\hat{U}^N;Y^n,Z^n)-2n\epsilon' \\ &\geq nH(K|Y)+NH(U)-NR_U(D')-2\epsilon']- \\ &\quad [I(U^N;Y^n,Z^n)+H(\hat{U}^N|U^N)-I(\hat{U}^N;Y^n,Z^n)], \end{aligned} \tag{70}$$

where first inequality is due to the fact that $H(\hat{U}^N|Y^n,Z^n)\geq n[H(K|Y)-2\epsilon']$, that we have just shown, and the second inequality is due to the memorylessness of $\{U_i\}$ and the fact that $\hat{U}^N$ is the output of an essentially optimum rate–distortion code for U^N. Now, the second bracketted expression on the right–most side is the same as in eq. (30), where in the case of this specific scheme, both inequalities in (30) become equalities, i.e., this expression vanishes. This is because in our scheme, $U^N\to\hat{U}^N\to(Y^n,Z^n)$ is a Markov chain (and so, the first inequality of (30) is tight) and because $H(\hat{U}^N|U^N,Y^n,Z^n)\leq H(\hat{U}^N|U^N)=0$ (as $\hat{U}^N$ is a deterministic function of U^N), which makes the second inequality of (30) tight. As a result, we have

$$\begin{aligned} H(U^N|Y^n,Z^n) &\geq N[H(K|Y)/\lambda+H(U)-R_U(D_1)-2\epsilon'/\lambda] \\ &\geq N[h+R_U(D_1)-H(U)+H(U)-R_U(D_1)-2\epsilon/\lambda] \\ &= N(h-2\epsilon'/\lambda), \end{aligned} \tag{71}$$

where we have used condition (a). This completes the proof of the direct part.

7 Acknowledgement

The author would like to thank Dr. Yossi Steinberg for interesting discussions.

References

[1] A. Adelsbach, S. Katzenbeisser, and A.-R. Sadeghi, "Cryptography meets watermarking: detecting watermarks with minimal or zero knowledge disclosure," preprint 2002. Available on–line at [www-krypt.cs.uni-sb.de/download/papers]

[2] S. C. Cheung and D. K. W. Chiu, "A watermark infrastructure for enterprise document management," *Proc. 36th Hawaii International Conference on System Sciences (HICSS'03)*, Hawaii, 2003.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

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

[5] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Information and Control*, vol. 9, no. 1, pp. 19-31, 1980.

[6] A. Jayawardena, B. Murison, and P. Lenders, "Embedding multiresolution binary images into multiresolution watermark channels in wavelet domain," preprint 2000. Available on–line at [www.tsi.enst.fr/∼maitre/tatouage/icassp00/articles].

[7] K. Kuroda, M. Nishigaki, M. Soga, A. Takubo, and I. Nakamura, "A digital watermark using public–key cryptography for open algorithm," *Proc. ICITA 2002*. Also, available on–line at [http://charybdis.mit.csu.edu.au/∼mantolov/CD/ICITA2002/papers/131-21.pdf].

[8] A. Maor and N. Merhav, "On joint information embedding and lossy compression," submitted to *IEEE Trans. Inform. Theory*, July 2003. Available on–line at [www.ee.technion.ac.il/people/merhav].

[9] A. Maor and N. Merhav, "On joint information embedding and lossy compression in the presence of a stationary memoryless attack channel," submitted to *IEEE Trans. Inform. Theory*, January 2004. Available on–line at [www.ee.technion.ac.il/people/merhav].

[10] N. Merhav and S. Shamai (Shitz), "On joint source–channel coding for the Wyner–Ziv source and the Gel'fand–Pinsker channel," *IEEE Trans. Inform. Theory*, vol. 49, no. 11, pp. 2844–2855, November 2003.

[11] P. Moulin and J. A. O'Sullivan, "Information–theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 563–593, March 2003.

[12] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT–19, pp. 471–480, 1973.

[13] M. Steinder, S. Iren, and P. D. Amer, "Progressively authenticated image transmission," preprint 1999. Available on–line at [www.cis.udel.edu /amer/PEL/poc/pdf/milcom99-steiner.pdf].

[14] H. Yamamoto, "Rate–distortion theory for the Shannon cipher system," *IEEE Trans. Inform. Theory*, vol. 43, no. 3, pp. 827–835, May 1997.

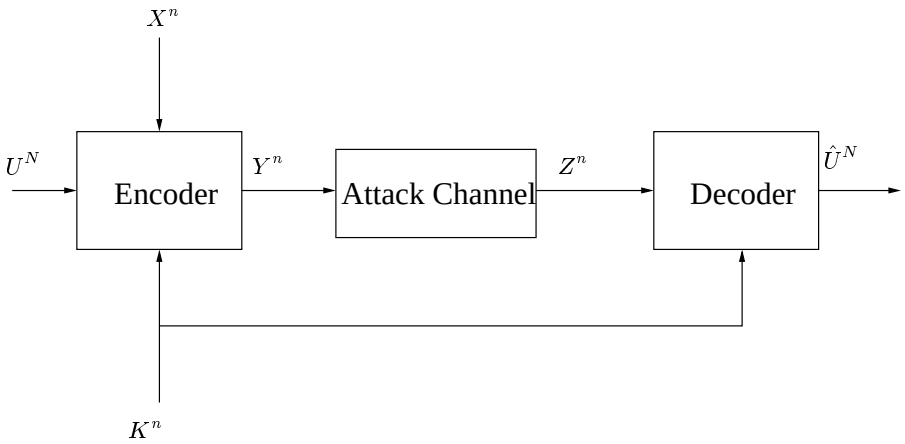

Figure 1: A generic watermarking/encryption scheme.

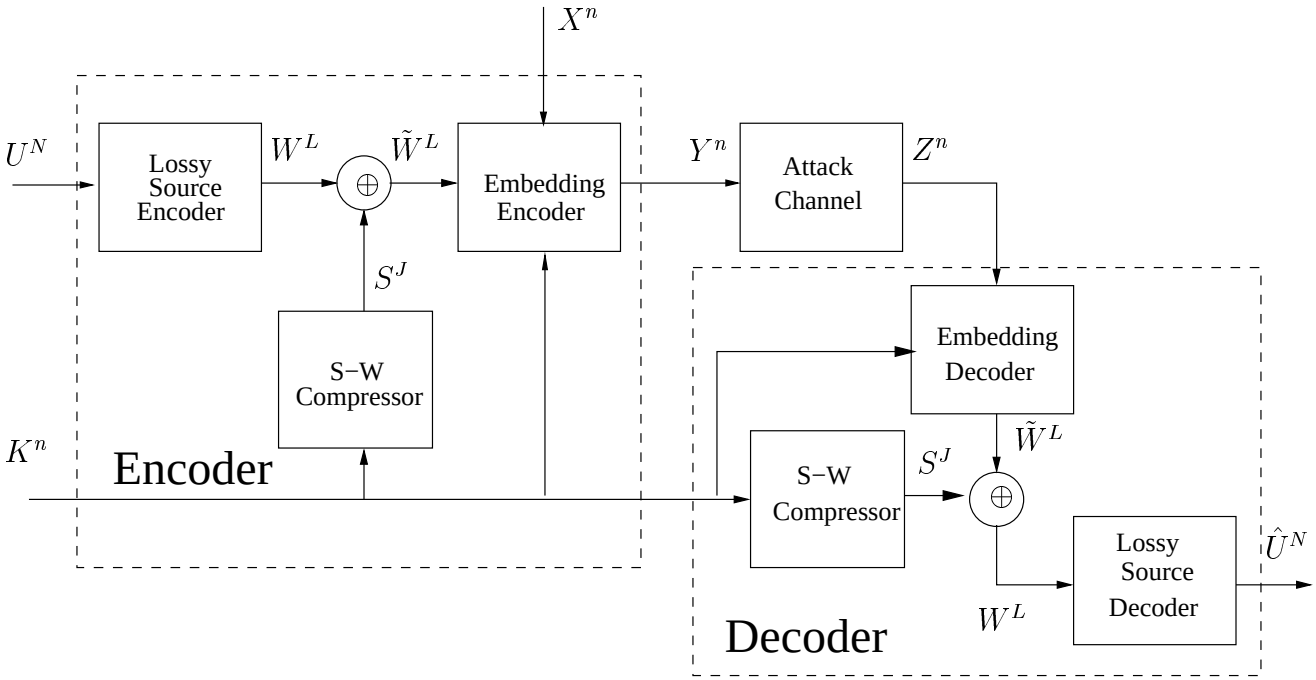

Figure 2: The proposed watermarking/encryption scheme (general case).