On Causal and Semicausal Codes for Joint Information Embedding and Source Coding

Neri Merhav*
Department of Electrical Engineering
Technion – Israel Institute of Technology
Haifa 32000, Israel
merhav@ee.technion.ac.il

Erik Ordentlich
Hewlett-Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304, USA
eord@hpl.hp.com

March 1, 2004

Abstract

A source of random message bits is to be embedded into a covertext modeled as a discrete memoryless source (DMS), resulting in a stegotext from which the embedded bits should be recoverable. A causal code for such a scenario consists of an encoder that generates the stegotext as a causal function of the message bits and the covertext, and a decoder that reproduces the message bits as a causal function of the stegotext. A semicausal code, on the other hand, has an encoder that is causal only with respect to the covertext, and not necessarily with respect to the message, and has a possibly noncausal decoder. We analyze the possible tradeoffs among: (a) the distortion between the stegotext and the covertext, (b) the compressibility of the stegotext, and (c) the rate at which random bits are embedded, that are achievable with causal and semicausal codes, with and without attacks on the stegotext. We also study causal and semicausal codes for the private version of the above scenario in which the decoder has access to the covertext. Connections are made with the causal rate–distortion function of Neuhoff and Gilbert [10], as well as the problem of channel coding with causal side information at the transmitter analyzed by Shannon [11]. For example, the optimal tradeoffs among the three quantities above for causal codes are shown to be achievable by time sharing a small number of scalar or symbol–by–symbol encoders and decoders, paralleling the main result of [10].

1 Introduction

We study the problem of joint lossy compression and information embedding under various causality restrictions on the encoder and decoder. Specifically, let $X_1, X_2, \ldots \sim P_X$ be a discrete memoryless covertext, whose elements take on values in a finite alphabet $\mathcal{X}$, and let $U_1, U_2, \ldots$ be an independent stream of purely random message bits. A scheme for joint compression and embedding with embedding rate R_e, in full generality, consists of an encoder that maps $U^{\lceil nR_e \rceil} \stackrel{\triangle}{=} U_1, \ldots, U_{\lceil nR_e \rceil}$ and $X^n \stackrel{\triangle}{=} X_1, \ldots, X_n$ into a stegotext $\hat{X}^n \stackrel{\triangle}{=} \hat{X}_1, \ldots, \hat{X}_n$, taking values in another finite alphabet $\hat{\mathcal{X}}$, and a

*Work done while visiting Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA.

corresponding decoder that maps $\hat{X}^n$ into an estimate $\hat{U}^{\lceil nR_e \rceil} \stackrel{\triangle}{=} \hat{U}_1, \ldots, \hat{U}_{\lceil nR_e \rceil}$ of the message bits. Given a distortion measure $d_1 : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, the distortion, compressibility, and reliability, of the embedding are respectively given by $\sum_{t=1}^n d_1(X_t, \hat{X}_t)$, $H(\hat{X}^n)$ (the joint entropy of $\hat{X}^n$), and the error in the decoder's estimate of the message bits, the latter being measured by word–error probability or bit–error rate. The joint compression and embedding problem is to understand the tradeoffs that exist among the embedding rate, distortion, compressibility, and reliability. Rate–distortion theory is at one extreme of this problem, corresponding to an embedding rate of zero, and for which, hence, the reliability is irrelevant.

Our focus in this work is on causal and semicausal joint lossy compression and embedding codes. A causal code, described in greater detail in Section 3, generates the stegotext by applying a sequence of causal functions to both X^n and $U^{\lceil nR_e \rceil}$ and generates $\hat{U}^{\lceil nR_e \rceil}$ by applying a sequence of causal functions to $\hat{X}^n$. A semicausal code, on the other hand, is required to causally generate the stegotext only with respect to the covertext, but not with respect to the message bits. The corresponding decoder is not required to be causal at all. The purely noncausal version of the problem of joint compression and embedding (noncausal with respect to covertext as well) has already been treated in [6] (see also [5, 7, 13]).

In addition to causal codes being a natural direction in which to reduce complexity relative to the noncausal case, we are also motivated to consider causal codes for joint compression and embedding by the work of Neuhoff and Gilbert [10], who analyzed the problem of causal lossy compression. In considering memoryless sources, Neuhoff and Gilbert obtained the elegant result that while a causal lossy compression scheme can generate the reconstructed sequence as an arbitrary function of the past source symbols, the past is actually of no value at all, and that an optimal scheme consists of time–sharing no more than two scalar quantizers. One goal of this work is to check if a similar result holds when the embedding rate is nonzero. Indeed, as detailed in Section 3, we find, under a fairly general model of causality with respect to the message bits, and perfect reliability, that scalar embedding operations are optimal. We find, however, that a time sharing of up to three such scalar embedding operations may be required. While the setting of semicausal codes has no counterpart in [10], it nevertheless constitutes another meaningful trade–off between implementation complexity and performance. We find that optimal, asymptotically reliable, semicausal codes essentially convey the message bits through a code consisting of sequences of scalar quantizers, interpreted more generally, as mappings between $\mathcal{X}$ and $\hat{\mathcal{X}}$. This bears a strong similarity to Shannon's capacity achieving scheme for channels with causal side information at the transmitter, a connection which is explored in greater detail below.

In the tradition of other information theoretic works on data embedding ([1, 9, 12] and references therein), we also expand the causal and semicausal joint compression and embedding settings above to include attacks. In systems with attacks, the decoder no longer observes the stegotext, but, instead, observes a "forgery" $Y_1, Y_2, \ldots$ generated by a distorting agent acting on the stegotext. Our results in this direction are as follows. For the case of known time invariant memoryless attacks we characterize optimal causal and semicausal codes for joint compression and embedding. The semicausal codes analysis again focuses on asymptotically reliable codes, while in the treatment of causal codes we find optimal codes subject to a constraint on the bit–error rate of the decoded message, since perfect reliability is, in general, not possible. The results remain essentially identical to the attack–free case with relevant information quantities taking the attack channel parameters into account. In the setting of causal codes, due to the addition of the bit–error rate constraint, a fourth scalar embedder may be required for the optimal time sharing scheme.

We also consider the case of unknown attacks subject to a constraint on a specified additive distortion measure between the forgery and the stegotext. In the causal setting we define a game setup in which a joint compressor and embedder and an attacker select constraint satisfying strategies, respectively, to minimize and maximize the end–to–end bit–error rate in the decoded message. It is shown that if the attacker satisfies a certain technical distortion constraint (which includes constraints previously formulated in the literature), and the embedder is allowed to randomize, that a saddle point solution exists for all n, consisting of a random scalar embedding code and a time invariant memoryless attack. The end–to–end bit–error rate at the saddle point is characterized. In the semicausal setting we consider block–memoryless attacks similar to [9], and analyze a game in which the joint compressor and embedder tries to maximize the reliable embedding rate, while the attacker tries to minimize it. Using tools from the theory of compound channels (such as maximum mutual information decoding) again a saddle point solution is shown to exist, in which a sequence of codes selected according to a memoryless distribution and a time invariant memoryless attack are, respectively, optimal. The saddle point embedding rate is characterized.

In a final set of results, we also characterize optimal causal and semicausal codes for the private versions of the above scenarios, in which the decoder also has access to the covertext X^n when decoding the embedded message.

The paper is organized as follows. Some overriding notation is explained in Section 2. All results for causal codes, with and without attacks, including the minimax results and private versions are presented in Section 3. The analagous set of results for semicausal codes are then presented in Section 4.

2 Notation

As is customary in the literature, given a symbol x and integers $m < n$, we use x_m^n to denote the subsequence $x_m, x_{m+1}, \ldots, x_n$. For $m = 1$ or $m = -\infty$ the subscript m will usually be omitted. Random variables will be denoted in upper case, while their realizations, generally in lower case. A calligraphic font will usually denote an alphabet or a set. The cardinality of a finite set $\mathcal{X}$ will be denoted by $|\mathcal{X}|$. $Ef(X, Y, \ldots)$ will denote the expectation of $f(\cdot)$ with respect to the random variables $X, Y, \ldots$. The expression $P_{X,Y,\ldots}$ will denote the joint probability mass function (pmf) of the random variables $X, Y, \ldots$ That is, $P_{X,Y,\ldots}(x, y, \ldots)$ is the probability that $X = x, Y = y, \ldots$ Occasionally, the joint pmf of a set of random variables will be denoted using a symbol other than P, but will have the same subscript convention. The expression $P_{Y|X}$ will denote the conditional pmf of Y given X. Conditional pmfs associated with channels or attacks will generally be denoted in the same way, but with Q in place of P. We adopt the usual notation for the standard information theoretic quantities [2] with all implicit logarithms in such quantities taken to the base 2.

3 Causal codes

3.1 Attack–free systems: perfect watermark reconstruction

Assuming the setting of the introduction, a (n, R_e) *causal code for joint lossy source coding and information embedding* operates as follows: Along every block of n covertext symbols, say, $(X_1, \ldots, X_n)$, the corresponding block of information bits to be embedded, $(U_1, \ldots, U_{\lceil nR_e \rceil})$, is parsed into n (possibly, variable–length) phrases $(W_1, W_2, \ldots, W_n) \stackrel{\triangle}{=} (U_1^{m_1}, U_{m_1+1}^{m_2}, \ldots, U_{m_{n-1}+1}^{\lceil nR_e \rceil})$, where $r_t = m_t - m_{t-1}$, $t = 1, \ldots, n$ (with $m_0 \stackrel{\triangle}{=} 0$, $m_n \stackrel{\triangle}{=} \lceil nR_e \rceil$), are non-negative[1] integers, that are all limited by a certain number $r_{\max} \in [R_e, \infty)$, and that sum to $\lceil nR_e \rceil$. Then the reproduction block is generated according to:

$$\hat{X}_t = f_t(X^t, W^t), \quad t = 1, \ldots, n, \tag{1}$$

[1]The value $r_t = 0$ is also allowed.

and $(\hat{X}_1, \ldots, \hat{X}_n)$ is entropy–coded. It is required that at time t, $t = 1, \ldots, n$, W_t would be recoverable from $\hat{X}^t$. Stated in another way, there is an inverse function g_t satisfying $W_t = g_t(\hat{X}^t)$, where $\hat{X}_t$ are obtained as above. Given a disortion measure $d_1 : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, we say that a compression ratio R_c is (D_1, R_e)–achievable with causal codes if, for all n sufficiently large, there exist (n, R_e) causal codes (consisting of $\{r_t, f_t(\cdot), g_t(\cdot)\}$) satisfying $(1/n)H(\hat{X}^n) \leq R_c$, $(1/n)\sum_{i=1}^n Ed_1(X_t, \hat{X}_t) \leq D_1$, and $g_t(\hat{X}^t) = W_t$ for all t with probability one.

Define the following OPTA function of D_1 and $r \in \{0, 1, \ldots, r_{\max}\}$:

$$R(D_1, r) = \min H(f(X, W^{(r)})), \tag{2}$$

where $W^{(r)}$ is an RV, uniformly distributed over $\{0,1\}^r$, independently of $X \sim P_X$, and the minimum is over all $f : \mathcal{X} \times \{0,1\}^r \to \hat{\mathcal{X}}$ such that $Ed_1(X, f(X, W^{(r)})) \leq D_1$ and such that there exists[2] a mapping $g : \hat{X} \to \{0,1\}^r$ for which $g(f(X, W^{(r)})) = W^{(r)}$ with probability one.

Let $\bar{R}(D_1, R_e)$ be the lower convex envelope (LCE) of $R(D_1, r)$ in both arguments. Specifically, enumerating all $n_r \stackrel{\triangle}{=} |\hat{\mathcal{X}}|^{2^r \cdot |\mathcal{X}|}$ functions $f : \mathcal{X} \times \{0,1\}^r \to \hat{\mathcal{X}}$, for all allowed values of r, as $\{f^{(r,j)}, r = 0, 1, \ldots, r_{\max},\ j = 1, 2, \ldots, n_r\}$, the function $\bar{R}(D_1, R_e)$ is defined as follows:

$$\bar{R}(D_1, R_e) = \inf \sum_{r=0}^{r_{\max}} \sum_{j=1}^{n_r} \alpha(r,j) H(f^{(r,j)}(X, W^{(r)})) \tag{3}$$

where the infimum is with respect to (w.r.t.) $\{\alpha(r,j)\}$ subject to the following constraints:

$$\begin{aligned}
\alpha(r,j) &\geq 0, \quad \forall\ r = 0, 1, \ldots, r_{\max},\ j = 1, 2, \ldots, n_r, \\
\sum_{r=0}^{r_{\max}} \sum_{j=1}^{n_r} \alpha(r,j) &= 1, \\
\sum_{r=0}^{r_{\max}} \sum_{j=1}^{n_r} \alpha(r,j) Ed_1(X, f^{(r,j)}(X, W^{(r)})) &\leq D_1, \\
\sum_{r=0}^{r_{\max}} \sum_{j=1}^{n_r} \alpha(r,j) r &\geq R_e.
\end{aligned} \tag{4}$$

The main result of this subsection is the following theorem.

Theorem 3.1 *The infimum of compression ratios that are (D_1, R_e)-achievable with causal codes is given by $\bar{R}(D_1, R_e)$.*

[2]This is the case iff the ranges of $f(\cdot, w)$ are disjoint for various values of $w \in \{0,1\}^r$.

Proof.

Converse. Let an arbitrary (n, R_e) causal code, with normalized distortion and block entropy not exceeding D_1 and R_c respectively, and encoding and decoding functions f_t and g_t, be given. Let $\tilde{F}_t$ be a random variable taking values in the space of functions mapping $\mathcal{X} \times \{0,1\}^{r_t} \to \hat{\mathcal{X}}$ determined by $f_t(\cdot, X^{t-1}, \cdot, W^{t-1})$, with the indeterminates corresponding to the values of x_t and w_t, so that $\hat{X}_t = \tilde{F}_t(X_t, W_t)$. Note also that $\tilde{F}_t$ is a function of (X^{t-1}, W^{t-1}). Similarly, let $\tilde{G}_t$ be a random variable taking values in the space of functions mapping $\hat{\mathcal{X}}$ to $\{0,1\}^{r_t}$ defined by $g_t(\cdot, \hat{X}^{t-1})$, so that $W_t = \tilde{G}_t(\hat{X}_t) = \tilde{G}_t(\tilde{F}_t(X_t, W_t))$. Note that $\tilde{G}_t$ is a function of $\hat{X}^{t-1}$, and hence, in turn, a function of (X^{t-1}, W^{t-1}). The compression ratio of the block $(\hat{X}_1, \ldots, \hat{X}_n)$ is lower bounded as follows:

$$\begin{aligned}
R_c \geq \frac{1}{n} H(\hat{X}_1^n) &= \frac{1}{n} \sum_{t=1}^{n} H(\hat{X}_t | \hat{X}^{t-1}) \\
&\geq \frac{1}{n} \sum_{t=1}^{n} H(\hat{X}_t | X^{t-1}, W^{t-1}) \\
&= \frac{1}{n} \sum_{t=1}^{n} H(\hat{X}_t | \tilde{F}_t, \tilde{G}_t, X^{t-1}, W^{t-1}) \\
&= \frac{1}{n} \sum_{t=1}^{n} H(\hat{X}_t | \tilde{F}_t, \tilde{G}_t), \qquad (5)
\end{aligned}$$

where we have used the fact that $\hat{X}^{t-1}$ is a function of X^{t-1}, W^{t-1}, due to the causality of the encoder, and the above definitions and properties of $\tilde{F}_t$ and $\tilde{G}_t$.

Note that W_t and X_t are independent of each other and of $\tilde{F}_t$ and $\tilde{G}_t$, with $X_t \sim P_X$ and W_t uniformly distributed on $\{0,1\}^{r_t}$. Additionally, since $W_t = \tilde{G}_t(\hat{X}_t) = \tilde{G}_t(\tilde{F}_t(X_t, W_t))$ with probability one, it follows that for every (f, g) in the support of $(\tilde{F}_t, \tilde{G}_t)$, f falls within the class of functions over which $H(f(X, W^r))$ is minimized in the definition of $R(D_1, r)$, with D_1 replaced by $Ed_1(X, f(X_t, W_t))$ and r replaced by r_t. Therefore, letting $D_f = Ed_1(X, f(X, W^{(r_t)}))$, for (f, g) in the support of $\tilde{F}_t, \tilde{G}_t$,

$$H(\hat{X}_t | (\tilde{F}_t, \tilde{G}_t) = (f, g)) \geq R(D_f, r_t) \geq \bar{R}(D_f, r_t), \qquad (6)$$

so that

$$\begin{aligned}
H(\hat{X}_t | \tilde{F}_t, \tilde{G}_t) &\geq E\bar{R}(D_{\tilde{F}_t}, r_t) \\
&\geq \bar{R}(ED_{\tilde{F}_t}, r_t) \\
&= \bar{R}(Ed_1(X_t, \hat{X}_t), r_t), \qquad (7)
\end{aligned}$$

where we have used the convexity of $\bar{R}$ and Jensen's inequality, along with the fact that $ED_{\tilde{F}_t} = Ed_1(X_t, \hat{X}_t)$. Finally,

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} H(\hat{X}_t|\tilde{F}_t, \tilde{G}_t) &\geq \frac{1}{n}\sum_{t=1}^{n} \bar{R}(Ed_1(X_t, \hat{X}_t), r_t) \\
&\geq \bar{R}\left(\frac{1}{n}\sum_{t=1}^{n} Ed_1(X_t, \hat{X}_t), \frac{1}{n}\sum_{t=1}^{n} r_t\right) \\
&\geq \bar{R}(D_1, \lceil nR_e\rceil/n). \qquad (8)
\end{aligned}$$

where the second inequality follows from the convexity of $\bar{R}(\cdot,\cdot)$ jointly in both arguments, and the third inequality follows from the hypothesis that $\frac{1}{n}\sum_{t=1}^{n} Ed_1(X_t, \hat{X}_t) \leq D_1$ and the non-increasing monotonicity of $\bar{R}(\cdot, \lceil nR_e\rceil/n)$. The converse follows by continuity of $\bar{R}(D_1, \cdot)$ and since $\lim_{n\to\infty}\lceil nR_e\rceil/n = R_e$.

Direct. Referring to the definition of $\bar{R}(D_1, R_e)$, we readily see that it is attainable by time–sharing the mappings $f^{(r,j)}$ and the corresponding bit-stream phrase lengths according to the minimizing vector of weights $\{\alpha(r,j)\}$. Moreover, since the minimization associated with $\bar{R}(D_1, R_e)$ is a linear programming problem on the simplex, with two additional constraints, it is easy to show that the minimum can be attained by a vector $\{\alpha(r,j)\}$ for which no more than three components are non-zero. In other words, optimal time-sharing can be implemented with three encoders. This completes the proof of the theorem. □

Discussion. Note that since $W^{(r)}$ must be recoverable from $f(X, W^{(r)})$ (cf. the definition of $R(D_1, r)$), then

$$\begin{aligned}
H(f(X, W^{(r)})) &= H(W, f(X, W^{(r)})) \\
&= H(W^{(r)}) + H(f(X, W^{(r)})|W^{(r)}) \\
&= r + \frac{1}{2^r}\sum_{w^{(r)}} H(f(X, w^{(r)})|W^{(r)} = w^{(r)}) \\
&= r + \frac{1}{2^r}\sum_{w^{(r)}} H(f(X, w^{(r)})). \qquad (9)
\end{aligned}$$

Taking the LCE of both sides, the first term is lower bounded by R_e while the second term gives time sharing among different causal source codes $f_{w^{(r)}}(x) = f(x, w^{(r)})$. The second term is lower bounded by the Neuhoff-Gilbert causal rate–distortion function $r_c(D_1)$ [10]. That is, $\bar{R}(D_1, R_e) \geq R_e + r_c(D_1)$.

3.2 Memoryless attack channels

Next, we extend our setting to include a memoryless attack channel $Q : \hat{\mathcal{X}} \to \mathcal{Y}$, operating on the reproduction sequence and producing an output ('forgery') sequence $\{Y_t\}$. As before, the sequence of messages is decoded sequentially, but this time from $\{Y_t\}$. Obviously, here we cannot, in general, expect error–free reconstruction of $\{U_i\}$. Optimum tradeoffs are, therefore, sought now among the following criteria: the compression rate R_c corresponding to entropy coding of $\{\hat{X}_t\}$, the distortion level D_1 between $\{X_t\}$ and $\{\hat{X}_t\}$ w.r.t. the distortion measure d_1, and the distortion D_2, w.r.t. another distortion measure d_2, between $\{U_i\}$ and $\{\hat{U}_i\}$, the bitstream estimated from $\{Y_t\}$. The (n, R_e) causal encoder for joint embedding and data compression is of the same structure as before. The only difference is that here the estimator of W_t is given by a causal function of the channel output, i.e., $\hat{W}_t = g_t(Y^t)$. Of course, if Q is the identity channel and $D_2 = 0$, we are back to the previous case. In analogy to the attack–free case, we say that a compression ratio R_c is (D_1, D_2, R_e)–achievable with causal codes if for all n sufficiently large, there exist (n, R_e) causal codes satisfying $(1/n)H(\hat{X}^n) \le R_c$, $(1/n)\sum_{i=1}^n Ed_1(X_t, \hat{X}_t) \le D_1$, and $(1/\lceil nR_e \rceil)\sum_{i=1}^{\lceil nR_e \rceil} Ed_2(U_i, \hat{U}_i) \le D_2$.

We now define a new OPTA function, similarly as before, as follows:

$$R(D_1, D_2, r) = \min H(f(X, W^{(r)})), \tag{10}$$

where X and $W^{(r)}$ are distributed as above and the minimum is over all $f : \mathcal{X} \times \{0,1\}^r \to \hat{\mathcal{X}}$ such that

$$Ed_1(X, f(X, W^{(r)})) \le D_1 \quad \text{and} \quad \min_{g:\mathcal{Y}\to\{0,1\}^r} Ed_2(W^{(r)}, g(Y)) \le D_2,$$

where $P_{Y|X,W^{(r)}}(y|x,w) = Q_{Y|\hat{X}}(y|f(x,w))$, and (with a slight abuse of notation) for $w = (u_1, \ldots, u_r)$ and $w' = (u_1', \ldots, u_r')$,

$$d_2(w, w') \stackrel{\triangle}{=} \sum_{i=1}^r d_2(u_i, u_i'), \tag{11}$$

with the convention that the summation over an empty set (when $r = 0$) is defined as zero.

Let $\bar{R}(D_1, D_2, R_e)$ be the LCE of $R(D_1, D_2, r)$ in all three arguments, which is defined similarly as $\bar{R}(D_1, R_e)$ but with the additional constraint

$$\sum_{r=0}^{r_{\max}} \sum_{j=1}^{n_r} \alpha(r,j) \cdot \min_g 2^{-r} \sum_{w\in\{0,1\}^r} \sum_{x,y} P_X(x) Q_{Y|\hat{X}}(y|f^{(r,j)}(x,w)) d_2(w, g(y)) \le D_2.$$

The following theorem is the main result of this subsection.

Theorem 3.2 *The infimum of all compression ratios that are (D_1, D_2, R_e)–achievable with causal codes subject to memoryless attacks distributed according to $Q_{Y|\hat{X}}$ is given by $\bar{R}(D_1, R_e D_2, R_e)$.*

Proof.

Converse. Let an arbitrary causal code, with embedding rate R_e and normalized distortion levels and block entropy not exceeding D_1, D_2, and R_c, respectively, be given. Define $\tilde{F}_t$ and $\tilde{G}_t$ based on the given encoder and decoder functions as in the proof of the attack free case, noting that the decoder is now based on Y^t instead of $\hat{X}^t$. The compression ratio of the block $(\hat{X}_1, \ldots, \hat{X}_n)$ is, in a manner similar to the attack free case, lower bounded as

$$R_c \geq \frac{1}{n} H(\hat{X}_1^n) \geq \frac{1}{n} \sum_{t=1}^{n} H(\hat{X}_t | X^{t-1}, W^{t-1}, Y^{t-1}) = \frac{1}{n} \sum_{t=1}^{n} H(\hat{X}_t | \tilde{F}_t, \tilde{G}_t). \tag{12}$$

Let $D_{1,f} = Ed_1(X, f(X, W^{(r_t)}))$ and $D_{2,f,g} = Ed_2(W^{(r_t)}, g(Y))$, where in the latter expression $P_{Y|W^{(r_t)},X}(y|w,x) = Q_{Y|\hat{X}}(y|f(x,w))$, and in both expressions W and X are independent of each other with W uniform on $\{0,1\}^{r_t}$ and $X \sim P_X$. Reasoning as in the attack free case,

$$H(\hat{X}_t | (\tilde{F}_t, \tilde{G}_t) = (f,g)) \geq R(D_{1,f}, D_{2,f,g}, r_t) \geq \bar{R}(D_{1,f}, D_{2,f,g}, r_t). \tag{13}$$

The joint convexity of $\bar{R}(\cdot,\cdot,\cdot)$ in all arguments, the non-increasing monotonicity of $\bar{R}(\cdot,\cdot,R_e)$ in both arguments, and the assumed distortion of the given causal coder, imply, by reasoning closely parallelling the attack free case, that

$$\frac{1}{n} \sum_{t=1}^{n} H(\hat{X}_t | \tilde{F}_t, \tilde{G}_t) \geq \bar{R}(D_1, \lceil nR_e \rceil D_2/n, \lceil nR_e \rceil/n). \tag{14}$$

The argument uses the fact that $\frac{1}{n}\sum_{i=1}^{n} ED_{2,\tilde{F}_t,\tilde{G}_t} \leq \lceil nR_e \rceil D_2/n$, which comes from the end-to-end distortion constraint. The converse follows by continuity, as in the attack free case.

Direct. The direct part is, again, by time–sharing – this time four encoders suffice.

Discussion. One might consider the private version of the problem, where the decoder has also causal access to the covertext $\{X_t\}$ in addition to the channel output. The results remain essentially the same, the only difference being that in the single letter expression, the minimization over g is now over all functions from $\mathcal{X} \times \mathcal{Y}$ to $\{0,1\}^r$, and so the D_2 distortion constraint, in the definition of $R(D_1, D_2, r)$, becomes $\min_g Ed_2(W^{(r)}, g(X,Y)) \leq D_2$.

3.3 Minimax and maximin watermarking games

In the previous subsection, we assumed a fixed memoryless attack channel Q. Suppose now that the channel is induced by a hostile party that tries to maximize the distortion in the reconstruction of $\{U_i\}$. We would like to find a possibly randomized causal embedding code that thwarts the efforts of the hostile party, which, we assume, benefits from full knowledge of the randomized code.

The actions of the hostile party are constrained to not distort the stegotext excessively. Our attack constraint takes the following form. Let $\mathcal{M}_{\mathcal{Y}^n|\hat{\mathcal{X}}^n}$ denote the space of all conditional probability distributions on $\mathcal{Y}^n$ given elements in $\hat{\mathcal{X}}^n$. Given a block length n and a distortion measure $d_3 : \hat{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$ define

$$\mathcal{Q}^n(D_3) \triangleq \Big\{ Q_{Y^n|\hat{X}^n} \in \mathcal{M}_{\mathcal{Y}^n|\hat{\mathcal{X}}^n} : \forall \text{ i.i.d. } \hat{X}^n, \forall\, \hat{x} \in \hat{\mathcal{X}},\ \frac{1}{n}\sum_{t=1}^{n} Ed_3(\hat{x}, Y_t) \le D_3, \\ \text{where } P_{Y_t}(y) = E(Q_{Y_t|\hat{X}^n}(y|\hat{X}^{t-1}, \hat{x}, \hat{X}^n_{t+1})) \Big\}. \quad (15)$$

In this subsection, we restrict attacks $Q_{Y^n|\hat{X}^n}$ to those that belong to $\mathcal{Q}^n(D_3)$. The set $\mathcal{Q}^n(D_3)$ includes all channels that satisfy

$$E(d_3(\hat{X}_t, Y_t)|\hat{X}^n = \hat{x}^n) \le D_3 \text{ for all } t \in \{1, \ldots, n\} \text{ and all } \hat{x}^n \in \hat{\mathcal{X}}^n. \quad (16)$$

A similar peak distortion constraint on the attacker is assumed in [8]. If d_3 is a difference distortion measure then the set $\mathcal{Q}^n(D_3)$ also includes additive attacks where $Y^n = Z^n + \hat{X}^n$ where Z^n is independent of $\hat{X}^n$ and satisfies

$$E\left(\frac{1}{n}\sum_{t=1}^{n} d_3(Z_t)\right) \le D_3.$$

Finally, note that $\mathcal{Q}^n(D_3)$ contains all DMC's with component channels belonging to $\mathcal{Q}^1(D_3)$. We will see below that the worst case attack in $\mathcal{Q}^n(D_3)$ is, in fact, such a DMC.

Given that the attack channel is not known ahead of time, it makes sense to consider randomized embedding strategies. Specifically, we allow the transmitter and receiver to select a causal code at random, independently of the covertext and subject to constraints on the average embedding rate, average compressibility rate and average distortion. Let R^n, F^n, and G^n be random variables denoting the phrase lengths, encoding functions, and decoding functions arising from the random code selection and let $\mathbf{C}^n = [R^n, F^n, G^n]$. Given distortion and rate constraints D_1, R_c, and R_e, let $\Theta^n(D_1, R_c, R_e)$ denote the set of all probability distributions $\theta_{\mathbf{C}^n}$ on causal codes satisfying $(1/n)\sum_{t=1}^{n} ER_t \ge R_e$,

$(1/n)H(\hat{X}^n|\mathbf{C}^n) \leq R_c$, and $(1/n)\sum_{t=1}^n Ed_1(X_t, \hat{X}_t) \leq D_1$. The conditional entropy in the average compressibility constraint (R_c) reflects the fact that the compressor and decompressor have full knowledge of the embedding code selected.

Formalizing the introductory remarks above, the goal of this subsection is to analyze the two quantities

$$D^n_{\text{minimax}}(D_1, R_c, R_e, D_3) \stackrel{\triangle}{=} \min_{\theta_{\mathbf{C}^n} \in \Theta^n(D_1, R_c, R_e)} \max_{Q_{Y^n|\hat{X}^n} \in \mathcal{Q}^n(D_3)} \frac{1}{n} \sum_{i=t}^{n} Ed_2(W_t, \hat{W}_t) \tag{17}$$

and

$$D^n_{\text{maximin}}(D_1, R_c, R_e, D_3) \stackrel{\triangle}{=} \max_{Q_{Y^n|\hat{X}^n} \in \mathcal{Q}^n(D_3)} \min_{\theta_{\mathbf{C}^n} \in \Theta^n(D_1, R_c, R_e)} \frac{1}{n} \sum_{t=1}^{n} Ed_2(W_t, \hat{W}_t), \tag{18}$$

where the expectation, in both definitions, is with respect to the distribution induced by the n-block attack channel $Q_{Y^n|\hat{X}^n}$ and the operation of a causal code randomly selected according to $\theta^n_{\mathbf{C}^n}$, independently of the data $\{U_i\}, \{X_t\}$. The random variables W_t and $\hat{W}_t$ and the extension of the distortion measure d_2 to operate on these random variables are as defined in the previous subsections (see (11)). Note that the end–to–end distortion is now normalized by n rather than $\lceil nR_e \rceil$. Also note that, in contrast to the previous subjections, the compressibility rate is now part of the constraints on the encoder and decoder, while the objective function is the end–to–end distortion.

The minimax distortion $D^n_{\text{minimax}}(D_1, R_c, R_e, D_3)$ corresponds to a game in which the transmitter and receiver jointly select, at random and independently of the data, a causal code, while the attacker, with knowledge of only the distribution used to select the causal code and not the particular code selected, chooses an attack from $\mathcal{Q}^n(D_3)$ to maximize the expected end-to-end distortion in the decoded embedded bits. Operationally, the joint selection of the causal encoder and decoder can be accomplished via a randomly chosen secret key revealed to both the transmitter and the receiver, but not to the attacker. The maximin distortion $D^n_{\text{maximin}}(D_1, R_c, R_e, D_3)$ corresponds to the dual situation in which the attacker goes first and the causal code selection is carried out with full knowledge of the selected attack. In both cases compression and decompression of $\hat{X}^n$ is also allowed to depend on the secret key.

We next define the main quantities characterizing the behavior of $D^n_{\text{minimax}}(D_1, R_c, R_e, D_3)$ and $D^n_{\text{maximin}}(D_1, R_c, R_e, D_3)$. Let $\mathcal{F}_r$ denote the set of mappings $f : \mathcal{X} \times \{0,1\}^r :\to \hat{\mathcal{X}}$ and let $\mathcal{G}_r$ denote the set of mappings $g : \mathcal{Y} \to \{0,1\}^r$. Let $\Theta(D_1, R_c, R_e)$ denote the set of probability distributions θ on

$\{(r, f, g) : r \in 0, 1, \ldots, r_{\max}, f \in \mathcal{F}_r, g \in \mathcal{G}_r\}$ that satisfy

$$\begin{aligned} Ed_1(X, F(X, W)) &\leq D_1 \\ H(F(X, W)|F, R) &\leq R_c \\ E(R) &\geq R_e, \end{aligned} \tag{19}$$

where the joint distribution of (X, W, R, F, G) is given by

$$P^{\theta}_{X,W,F,R,G}(x, w, r, f, g) = \theta(r, f, g)P_X(x)2^{-r}1(w \in \{0, 1\}^r) \tag{20}$$

For $\theta \in \Theta(D_1, R_c, R_e)$ and $Q \in \mathcal{Q}^1(D_3)$ define

$$\Gamma(\theta, Q) \stackrel{\triangle}{=} Ed_2(W, \hat{W}) \tag{21}$$

where $d_2(w, w')$ is as defined in (11) and the expectation is with respect to the joint distribution of $(W, \hat{W})$ induced by the following joint distribution of $(X, W, R, F, G, Y, \hat{W})$:

$$P^{\theta,Q}_{X,W,F,R,G,Y,\hat{W}}(x, w, r, f, g, y, \hat{w}) = P^{\theta}_{X,W,F,R,G}(x, w, r, f, g)Q(y|f(x, w))1(\{\hat{w} = g(y)\}), \tag{22}$$

with $P^{\theta}_{X,W,F,R,G}$ defined in (20). Finally, define

$$\begin{aligned} D_2(D_1, R_c, R_e; Q) &\stackrel{\triangle}{=} \min_{\theta \in \Theta(D_1, R_c, R_e)} \Gamma(\theta, Q) &(23) \\ D_2(D_1, R_c, R_e, D_3) &\stackrel{\triangle}{=} \min_{\theta \in \Theta(D_1, R_c, R_e)} \max_{Q \in \mathcal{Q}^1(D_3)} \Gamma(\theta, Q). &(24) \end{aligned}$$

The following theorem characterizes the behavior of D^n_{minimax} and D^n_{maximin}.

Theorem 3.3 *For all n,*

$$D^n_{\text{minimax}}(D_1, R_c, R_e, D_3) = D^n_{\text{maximin}}(D_1, R_c, R_e, D_3) = D_2(D_1, R_c, R_e, D_3). \tag{25}$$

Theorem 3.3 and its proof indicate that the minimax and maximin games have the same value $D_2(D_1, R_c, R_e, D_3)$, which is attained, for each n, at a saddle point where both the embedding strategy and the attack are memoryless. Specifically, it is shown that the embedding code distribution achieving $D^n_{\text{minimax}}(D_1, R_c, R_e, D_3)$ selects a sequence of independent and identically distributed phrase lengths and scalar encoding and decoding functions with each component phrase length, encoder, and decoder triple distributed according to the pmf achieving $D_2(D_1, R_c, R_e, D_3)$ in (24). Similarly, the attack channel achieving $D^n_{\text{maximin}}(D_1, R_c, R_e, D_3)$ is shown to be a stationary DMC with component channels achieving the maximum over Q of $D_2(D_1, R_c, R_e; Q)$ defined by (23).

Proof. Since it is immediate that $D^n_{\text{maximin}}(D_1, R_c, R_e, D_3) \leq D^n_{\text{minimax}}(D_1, R_c, R_e, D_3)$, it suffices to show that

$$D^n_{\text{maximin}}(D_1, R_c, R_e, D_3) \geq D_2(D_1, R_c, R_e, D_3)$$

and

$$D^n_{\text{minimax}}(D_1, R_c, R_e, D_3) \leq D_2(D_1, R_c, R_e, D_3).$$

These are proved, respectively, in the converse and direct parts below.

Converse. Let Q^* maximize $D_2(D_1, R_c, R_e; Q)$, and let $[Q^*]^n$ denote that element of $\mathcal{Q}^n(D_3)$ corresponding to a DMC with component channels all equal to Q^*. Let $\theta^*_{\mathbf{C}^n}$ achieve the minimum in (18) when the outer maximization is omitted and Q is simply set to $[Q^*]^n$. Let R_t denote the random variable specifying the phrase-length parameter r used at index t, as selected according to $\theta^*_{\mathbf{C}^n}$. In analogy to the proofs of Theorems 3.1 and 3.2, let $\tilde{F}_t$ denote the random variable specifying the mapping $F_t(\cdot, X^{t-1}, \cdot, W^{t-1})$ sending $\mathcal{X} \times \{0,1\}^{R_t}$ to $\hat{\mathcal{X}}$, where, in this case, $F_t(\cdot)$ is itself a random variable distributed according to $\theta^*_{\mathbf{C}^n}$. Let $\tilde{G}_t$ denote the analogous random variable corresponding to the mapping $G_t(\cdot, Y^{t-1})$ sending $\mathcal{Y}$ to $\{0,1\}^{R_t}$. For a causal code selected according to $\theta^*_{\mathbf{C}^n}$ operating under attacks from $[Q^*]^n$, it then follows that

$$\begin{aligned} D^n_{\text{maximin}}(D_1, R_c, R_e, D_3) &\geq \frac{1}{n}\sum_{t=1}^{n} Ed_2(W_t, \hat{W}_t) \\ &\geq \frac{1}{n}\sum_{t=1}^{n} D_2(Ed_1(X_t, \hat{X}_t), H(\hat{X}_t|\tilde{F}_t, R_t), ER_t; Q^*), \end{aligned} \tag{26}$$

where expectations are with respect to distributions induced by $\theta^*_{\mathbf{C}^n}$, the operation of $\mathbf{C}^n$, $[Q^*]^n$, X^n, and W^n. Inequality (26) is justified as follows. Notice that the joint distribution of $(R_t, \tilde{F}_t, \tilde{G}_t)$, as induced by $\theta^*_{\mathbf{C}^n}$, belongs to $\Theta(Ed_1(X_t, \hat{X}_t), H(\hat{X}_t|\tilde{F}_t, R_t), ER_t)$ defined by (19), with $\hat{X}_t = \tilde{F}_t(X_t, W_t)$, and that the joint distribution of W_t and $\hat{W}_t$ is of the form underlying the definition of $\Gamma(\theta, Q^*)$ given by (21). Inequality (26) thus follows by applying the definition (23) of $D_2(D_1, R_c, R_e; Q)$ for each t.

Since $\theta^*_{\mathbf{C}^n}$ belongs to $\Theta^n(D_1, R_c, R_e)$, it follows that

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^n Ed_1(X_t, \hat{X}_t) &\leq D_1 \\ \frac{1}{n}\sum_{t=1}^n ER_t &\geq R_e, \end{aligned}$$

and, by an argument similar to the one leading to (5), that

$$\frac{1}{n}\sum_{t=1}^n H(\hat{X}_t|\tilde{F}_t, R_t) \leq \frac{1}{n}H(\hat{X}^n|\mathbf{C}^n) \leq R_c.$$

These facts, along with the convexity of $D_2(\cdot,\cdot,\cdot;Q^*)$ and the appropriate monotonicity of $D_2(\cdot,\cdot,\cdot;Q^*)$ in each argument (while fixing the remaining arguments), allow us to conclude, by applying Jensen's inequality to (26), that

$$D^n_{\text{maximin}}(D_1, R_c, R_e, D_3) \geq D_2(D_1, R_c, R_e; Q^*) = D_2(D_1, R_c, R_e, D_3), \tag{27}$$

where equality follows from the minimax theorem of convex analysis, the assumptions of which are clearly satisfied by $\Gamma(\theta, Q)$ and the respective convex constraints on θ and Q given by (23) and (24).

Direct. Let $\theta^*_{R,F,G} \in \Theta(D_1, R_c, R_e)$ achieve $D_2(D_1, R_c, R_e, D_3)$. Consider a random causal code $\mathbf{C}^n$ comprised of i.i.d. scalar or single–letter encoding and decoding functions based on i.i.d. message phrase lengths, with each component function/phrase length selected according to $\theta^*_{R,F,G}$. Letting R_t again denote the random phrase length at index t and $\tilde{F}_t$ and $\tilde{G}_t$ denote the corresponding encoder and decoder mappings determined respectively by $F_t(\cdot, X^{t-1}, W^{t-1})$ and $G_t(\cdot, Y^{t-1})$, we have that $\{(R_t, \tilde{F}_t, \tilde{G}_t)\}$ are i.i.d. with $(R_t, \tilde{F}_t, \tilde{G}_t) \sim \theta^*_{R,F,G}$. Let $\tilde{\theta}^*_{\mathbf{C}^n}$ denote this distribution on causal codes. Note that $\tilde{\theta}^*_{\mathbf{C}^n} \in \Theta^n(D_1, R_c, R_e)$ for all n.

Let $Q^*_{Y^n|\hat{X}^n}$ achieve the maximum in (17) when the outer minimization is omitted and $\theta_{\mathbf{C}^n}$ is set to $\tilde{\theta}^*_{\mathbf{C}^n}$. Consider the expected distortion induced by a causal code selected according $\tilde{\theta}^*_{\mathbf{C}^n}$ operating under the attack $Q^*_{Y^n|X^n}$. For each t,

$$\begin{aligned} Ed_2(W_t, \hat{W}_t) &= \sum_{\substack{r^n, \tilde{f}^n, x^n, \\ w^n, \tilde{g}^n, y}} d_2(w_t, \tilde{g}_t(y)) Q^*_{Y_t|\hat{X}^n}(y|\tilde{f}_1(x_1, w_1), \ldots, \tilde{f}_n(x_n, w_n)) \cdot \\ &\qquad \prod_{j=1}^n \frac{P_X(x_j)\theta^*_{R,F,G}(r_j, \tilde{f}_j, \tilde{g}_j)}{2^{-r_j}} \qquad (28) \\ &= \sum_{\substack{r,f,g, \\ x,w,y}} d_2(w, g(y)) \bar{Q}^*_{Y|\hat{X};t}(y|f(x,w)) P^{(\theta^*_{R,F,G})}_{X,W,F,R,G}(x, w, f, r, g), \qquad (29) \end{aligned}$$

where $P^{(\theta^*_{R,F,G})}_{X,W,F,R,G}$ is given by (20), with $\theta = \theta^*_{R,F,G}$, and

$$\bar{Q}^*_{Y|\hat{X};t}(y|\hat{x}) \stackrel{\triangle}{=} \sum_{\substack{r^{t-1},r^n_{t+1},\tilde{f}^{t-1},\tilde{f}^n_{t+1}\\ x^{t-1},x^n_{t+1},w^{t-1},w^n_{t+1}}} \Big[Q^*_{Y_t|\hat{X}^n}(y|\tilde{f}_1(x_1,w_1),\ldots,\tilde{f}_{t-1}(x_{t-1},w_{t-1}),\hat{x},\tilde{f}_{t+1}(x_{t+1},w_{t+1}),\ldots,\tilde{f}_n(x_n,w_n))\cdot \prod_{j\in\{1,\ldots,n\}\setminus t} P_X(x_j)\theta^*_{R,F}(r_j,\tilde{f}_j)2^{-r_j}\Big]. \quad (30)$$

The summations in (28) and (30) are assumed to be only over those values satisfying the conditions $w_j \in \{0,1\}^{r_j}$ and $\tilde{f}_j \in \mathcal{F}_{r_j}$.

Let

$$\bar{Q}^*_{Y|\hat{X}}(y|\hat{x}) \stackrel{\triangle}{=} \frac{1}{n}\sum_{t=1}^{n} \bar{Q}^*_{Y|\hat{X};t}(y|f(x,w)).$$

We then have that

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n} Ed_2(W_t,\hat{W}_t) &= \sum_{\substack{r,f,g,\\ x,w,y}} d_2(w,g(y))\bar{Q}^*_{Y|\hat{X}}(y|f(x,w))P^{(\theta^*_{R,F,G})}_{X,W,F,R,G}(x,w,f,r,g) \\ &= \Gamma(\theta^*_{R,F,G},\bar{Q}^*_{Y|\hat{X}}), \end{aligned} \quad (31)$$

which follows from the definition of $\Gamma(\cdot,\cdot)$ given by (21).

We note, for use below, that, for all n,

$$\bar{Q}^*_{Y|\hat{X}} \in \mathcal{Q}^1(D_3).$$

This can be seen as follows. Since, by assumption, $Q^*_{Y^n|\hat{X}^n} \in \mathcal{Q}^n(D_3)$, it satisfies, for all $\hat{x}$,

$$\begin{aligned} D_3 &\geq \frac{1}{n}\sum_{t=1}^{n}\sum_{y} d_3(\hat{x},y)E(Q^*_{Y_t|\hat{X}^n}(y|\hat{X}^{t-1},\hat{x},\hat{X}^n_{t+1})) \\ &= \sum_{y} d_3(\hat{x},y)\frac{1}{n}\sum_{t=1}^{n} E(Q^*_{Y_t|\hat{X}^n}(y|\hat{X}^{t-1},\hat{x},\hat{X}^n_{t+1})) \\ &= \sum_{y} d_3(\hat{x},y)\frac{1}{n}\sum_{t=1}^{n} \bar{Q}^*_{Y|\hat{X};t}(y|\hat{x}) \\ &= \sum_{y} d_3(\hat{x},y)\bar{Q}^*_{Y|X}(y|\hat{x}), \end{aligned}$$

where $\hat{X}^n = \tilde{F}_1(X_1,W_1),\ldots,\tilde{F}_n(X_n,W_n)$ with $\{(X_t,W_t,\tilde{F}_t,R_t)\}$ i.i.d. as above.

Consolidating the above, we arrive at the following chain of inequalities:

$$D^n_{\text{minimax}}(D_1, R_c, R_e, D_3) \overset{(a)}{\leq} \frac{1}{n}\sum_{t=1}^{n} Ed_2(W_t, \hat{W}_t) \tag{32}$$

$$\overset{(b)}{=} \Gamma(\theta^*_{R,F,G;n}, \bar{Q}^*_{Y|\hat{X}}) \tag{33}$$

$$\overset{(c)}{\leq} D_2(D_1, R_c, R_e, D_3), \tag{34}$$

where (a) follows since $\theta^*_{\mathbf{C}^n} \in \Theta^n(D_1, R_c, R_e)$ and from the manner in which $Q^*_{Y^n|\hat{X}^n}$, with respect to which the expected distortion is computed, was chosen above, (b) – from (31), and (c) – from the fact, noted above, that $\bar{Q}^*_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)$, the definition of $D_2(D_1, R_c, R_e, D_3)$, and the fact that $\theta^*_{R,F,G}$ achieves it. □

Discussion. Referring again to the private game, a result similar to Theorem 3.3 also holds, where, in this case, the transmitter and receiver can jointly select, at random, a private causal code, in which the decoding is allowed to also depend causally on $\{X_t\}$. One need only modify the definition of $\mathcal{G}_r$ above to now denote the set of mappings $g : \mathcal{Y} \times \mathcal{X} \to \{0,1\}^r$, replace $1(\hat{w} = g(y))$ with $1(\hat{w} = g(y,x))$ in the definition of $P^{\theta,Q}_{X,W,F,R,G,Y,\hat{W}}$, and propagate these changes into the definition of $D_2(D_1, R_c, R_e, D_3)$ to obtain the private causal code analogue of Theorem 3.3.

4 Semicausal codes

In this section we partially relax the causality constraints of the previous section and require that the embedding be causal only with respect to the covertext and not necessarily the message. The decoder may be noncausal. While the optimal schemes of the previous section consist of scalar operations both at the encoder and decoder, it turns out that the optimal schemes in the semicausal setting consist of "codebooks" of sequences of scalar embedding functions which are used, in a channel coding like fashion, to signal a message to the decoder.

Let us describe more formally the semicausal compressed embedding setting. A message W uniformly distributed over the alphabet $\{1, \ldots, M\}$ is to be communicated by appropriately perturbing a sequence of samples $X^n = X_1, \ldots, X_n \in \mathcal{X}^n$ from a covertext modeled as a DMS into a stegotext $\hat{X}^n = \hat{X}_1, \ldots, \hat{X}_n \in \hat{\mathcal{X}}$. The encoder of a (M, n) semicausal code achieves this perturbation via a sequence of functions $f_t(x^t, w) : \mathcal{X}^t \times \{1, \ldots, M\} \to \hat{\mathcal{X}}$ such that $\hat{X}_t = f_t(X^t, W)$ for $t = 1, \ldots, n$. The adjective "semicausal" refers to the fact that the causality of the perturbation is required only with respect to the covertext sequence but not the message. A corresponding decoder, which is *not* required to be causal, generates an estimate of the message $\hat{W}$ by observing $\hat{X}^n$, or, in the case of attacks, which we consider in Subsections 4.3 and 4.4 below, by observing a "forgery" Y^n which is $\hat{X}^n$ subjected to a distorting effect or attack. In all cases, it is desired that the probability of a decoding error $P_e = \Pr(\hat{W} \neq W)$ tend to 0 as the operating block length n increases. The following additional distortion and rate constraints are imposed on the system. As in the causal case, given a single letter distortion measure $d_1(x, \hat{x})$ we require that the encoder mappings have an expected average distortion $E((1/n)\sum_{t=1}^{n} d_1(X_t, \hat{X}_t))$ no larger than D_1. We also require that $(1/n)H(\hat{X}^n)$ be no larger than R_c which would allow the stegotext to be compressed at an asymptotic rate of R_c.

In the following subsections we consider attack free systems, systems subject to known time invariant memoryless attacks, and finally systems subject to unknown blockwise time invariant and memoryless attacks.

4.1 Attack free systems

Here we consider the attack free case in which the decoder observes $\hat{X}^n$ directly or, more specifically, is a mapping $g : \hat{\mathcal{X}}^n \to \{1, \ldots, M\}$ with $\hat{W} = g(\hat{X}^n)$. An embedding rate R_e is (D_1, R_c)–achievable with semicausal codes if and only if there exists a sequence of $(2^{\lceil nR_e \rceil}, n)$ semicausal codes such that: the error probability P_e tends to zero, the expected average distortion satisfies

$$E\left(\frac{1}{n}\sum_{t=1}^{n} d_1(X_t, \hat{X}_t)\right) \leq D_1,$$

and the entropy of $\hat{X}^n$ satisfies

$$\frac{1}{n}H(\hat{X}^n) \leq R_c.$$

Let $\mathcal{F}$ denote the set of all functions with domain $\mathcal{X}$ and range $\hat{\mathcal{X}}$. The set of embedding rates that are (D_1, R_c)–achievable with semicausal codes will be expressed in terms of random variables F taking values in $\mathcal{F}$.

For a discrete memoryless covertext $\{X_t\}$ with marginal probability distribution function $P_X(x) = \Pr(X_t = x)$, let

$$C(D_1, R_c) \stackrel{\triangle}{=} \max_{F: Ed_1(X, F(X)) \leq D_1} \min[I(F; F(X)), R_c - H(F(X)|F)], \tag{35}$$

where the maximization is over random variables F that are independent of X and take values in $\mathcal{F}$. The following theorem provides a single letter characterization of the semicausal (D_1, R_c) embedding capacity defined as the supremum of embedding rates that are (D_1, R_c)–achievable with semicausal codes.

Theorem 4.1 *The semicausal (D_1, R_c) embedding capacity is given by $C(D_1, R_c)$.*

The direct part of Theorem 4.1 is proved by considering an embedding scheme in which the message W is mapped to a codeword over an alphabet of scalar embedding functions and interpreting the embedding process, in which the codeword symbols are evaluated on the corresponding covertext symbols, as passing the codeword through an appropriate DMC. The corresponding decoder is simply a channel decoder for the resulting DMC. As noted in greater detail in the discussion following the proof below, this embedding scheme bears a strong similarity to Shannon's coding scheme for channels with causal side information at the transmitter.

Proof.

Converse. Let R_e be an achievable rate. There is a sequence of $(\lceil 2^{nR_e} \rceil, n)$ semicausal embedding codes of block length n having error probability no larger than δ_n with $\lim_{n\to\infty} \delta_n = 0$, expected average distortion no larger than D_1, and $(1/n)H(\hat{X}^n) \leq R_c$. As in the proof of Theorem 3.1, the random variables $\tilde{F}_t$ appearing below take values in $\mathcal{F}$ with $\tilde{F}_t = f_t(\cdot, X^{t-1}, W)$. It follows therefore that $\hat{X}_t = \tilde{F}_t(X_t)$ and that $\tilde{F}_t$ is a function of (X^{t-1}, W). For each n, the following chain of inequalities holds for $\delta'_n = h(\delta_n)/n + R_e\delta_n$, where $h(\cdot)$ is the binary entropy function.

$$
\begin{aligned}
nR_e &\leq H(W) && (36)\\
&= H(W) - H(W|\hat{W}) + H(W|\hat{W}) && (37)\\
&\overset{(a)}{\leq} I(W;\hat{W}) + n\delta'_n && (38)\\
&\overset{(b)}{\leq} I(W;\hat{X}^n) + n\delta'_n && (39)\\
&= \sum_{t=1}^{n} I(W;\hat{X}_t|\hat{X}^{t-1}) + n\delta'_n && (40)\\
&\overset{(c)}{\leq} \sum_{t=1}^{n} H(\hat{X}_t) - \sum_{t=1}^{n} H(\hat{X}_t|\hat{X}^{t-1}, W) + n\delta'_n && (41)\\
&\overset{(d)}{\leq} \sum_{t=1}^{n} H(\hat{X}_t) - \sum_{t=1}^{n} H(\hat{X}_t|X^{t-1}, \hat{X}^{t-1}, W) + n\delta'_n && (42)\\
&\overset{(e)}{=} \sum_{t=1}^{n} H(\hat{X}_t) - \sum_{t=1}^{n} H(\hat{X}_t|X^{t-1}, W) + n\delta'_n && (43)\\
&\overset{(f)}{=} \sum_{t=1}^{n} H(\hat{X}_t) - \sum_{t=1}^{n} H(\hat{X}_t|\tilde{F}_t, X^{t-1}, W) + n\delta'_n && (44)\\
&\overset{(g)}{=} \sum_{t=1}^{n} H(\hat{X}_t) - \sum_{t=1}^{n} H(\hat{X}_t|\tilde{F}_t) + n\delta'_n && (45)\\
&= \sum_{t=1}^{n} I(\tilde{F}_t; \tilde{F}_t(X_t)) + n\delta'_n && (46)\\
&\overset{(h)}{=} nI(F;F(X)|J) + n\delta'_n && (47)\\
&= n(H(F(X)|J) - H(F(X)|F,J)) + n\delta'_n && (48)\\
&\overset{(i)}{\leq} n(H(F(X)) - H(F(X)|F)) + n\delta'_n && (49)\\
&= nI(F;F(X)) + n\delta'_n, && (50)
\end{aligned}
$$

where J, introduced in (h), is uniformly distributed in $\{1,\ldots,n\}$ and is independent of all other random variables, and where $X \triangleq X_J$ and $F \triangleq \tilde{F}_J$. Note, for use below, that X so defined is also distributed according to P_X, and further that X and F are independent. Inequality (a) follows from Fano's inequality, (b) – from the data processing inequality, (c) and (d) – since conditioning reduces entropy, (e) – since $\hat{X}^{t-1}$ is a function of (X^{t-1}, W), (f) – since $\tilde{F}_t$ is a function of (X^{t-1}, W), (g) – since $\hat{X}_t = \tilde{F}_t(X_t)$ and X_t is independent of (X^{t-1}, W), and (i) – since conditioning reduces entropy, and since $F(X) \to F \to J$ are readily seen to form a Markov chain.

The following chain of inequalities also holds.

$$
\begin{aligned}
nR_c &\geq H(\hat{X}^n) && (51)\\
&= H(\hat{X}^n) + H(W|\hat{W}) - H(W|\hat{W}) && (52)\\
&\overset{(a)}{\geq} H(\hat{X}^n) + H(W|\hat{W}) - n\delta'_n && (53)\\
&\overset{(b)}{\geq} H(\hat{X}^n) + H(W|\hat{W}, \hat{X}^n) - n\delta'_n && (54)\\
&\overset{(c)}{=} H(\hat{X}^n) + H(W|\hat{X}^n) - n\delta'_n && (55)\\
&= H(W) + H(\hat{X}^n|W) - n\delta'_n && (56)\\
&\overset{(d)}{\geq} nR_e + \sum_{t=1}^{n} H(\hat{X}_t|\hat{X}^{t-1}, W) - n\delta'_n && (57)\\
&\overset{(e)}{\geq} nR_e + \sum_{t=1}^{n} H(\hat{X}_t|X^{t-1}, W) - n\delta'_n && (58)\\
&\overset{(f)}{=} nR_e + \sum_{t=1}^{n} H(\hat{X}_t|\tilde{F}_t) - n\delta'_n && (59)\\
&= nR_e + \sum_{t=1}^{n} H(\tilde{F}_t(X_t)|\tilde{F}_t) - n\delta'_n && (60)\\
&= nR_e + nH(F(X)|F, J) - n\delta'_n && (61)\\
&\overset{(g)}{=} nR_e + nH(F(X)|F) - n\delta'_n, && (62)
\end{aligned}
$$

where (J, X, F) are as defined above. Inequality (a) again follows from Fano's inequality, (b) – since conditioning reduces entropy, (c) – since $W \to \hat{X}^n \to \hat{W}$ is a Markov chain, (d) – by the chain rule and since, by assumption, $H(W) \geq nR_e$, (e) – by the same reasoning as in (c) through (e) above, and (f) – by the same reasoning as (f) and (g) above. The last equality (g) again follows since $F(X) \to F \to J$ form a Markov chain.

Note that $Ed_1(X, F(X))$ is the expected average distortion incurred by the embedding so that, by assumption, $Ed_1(X, F(X)) \leq D_1$. Thus, we have found a random variable F which is independent of $X \sim P_X$ and satisfies

$$
\begin{aligned}
R_e &\leq \min[I(F; F(X)), R_c - H(F(X)|F)] + \delta'_n, \\
&\leq C(D_1, R_c) + \delta'_n && (63)
\end{aligned}
$$

where (63) follows from the definition of $C(D_1, R_c)$.

Since n can be chosen to make δ'_n arbitrarily small

$$R_e \leq C(D_1, R_c),$$

thereby proving the converse.

Direct. Let F^* achieve the maximum in (35) and let

$$R_e < \min[I(F^*; F^*(X)), R_c - H(F^*(X)|F^*)].$$

Consider the DMC with inputs $f \in \mathcal{F}$, and output $\hat{X} = f(X)$. The channel transition matrix is $P_{\hat{X}|F}(\hat{x}|f) = \Pr(f(X) = \hat{x})$. Consider also the generalized power constraints determined by the instantaneous powers $P_H(f) = H(f(X))$ and $P_{d_1}(f) = Ed_1(X, f(X))$. Note that $EP_H(F) = H(F(X)|F)$ and $EP_{d_1}(F) = Ed_1(X, F(X))$, for F and X independent.

We can now follow the standard achievability arguments (c.f. e.g. [2] pp. 244–245, [3] pp. 108–110) for power constrained channels to establish the existence of a sequence of $(\lceil 2^{nR_e} \rceil, n)$ channel codes with probability of error $P_{e,n} \leq \delta_n$ and with all codewords $[\tilde{f}_1(W), \ldots, \tilde{f}_n(W)]$ satisfying

$$\frac{1}{n}\sum_{i=1}^{n} P_H(\tilde{f}_i(W)) \leq EP_H(F^*) + \delta_n = H(F^*(X)|F^*) + \delta_n \tag{64}$$

and

$$\frac{1}{n}\sum_{t=1}^{n} P_{d_1}(\tilde{f}_t(W)) \leq EP_{d_1}(F^*) \leq D_1, \tag{65}$$

for some $\delta_n \to 0$ and n sufficiently large,.

These channel codes will now be shown to determine a suitable sequence of rate R_e semicausal (D_1, R_c) embedding codes. Specifically, for a message $W \in \{1, \ldots, \lceil 2^{nR_e} \rceil\}$ select the corresponding channel codeword $\mathbf{f}^n(W) \stackrel{\triangle}{=} [\tilde{f}_1(W), \ldots, \tilde{f}_n(W)] \in \mathcal{F}^n$ and let the semicausally obtained stegotext be $\hat{X}_t = \tilde{f}_t(W)(X_t) \stackrel{\triangle}{=} f_t(X_t, W)$. Thus the block of stegotext is precisely the output of the above DMC when the codeword $[\tilde{f}_1(W), \ldots, \tilde{f}_n(W)]$ is transmitted. Since this sequence of channel codes is arbitrarily reliable, a decoder applying a corresponding optimal channel decoder to the block of stegotext will recover the embedded message W with error probability $P_{e,n} \to 0$.

It remains to show that the resulting semicausal embedding codes satisfy the distortion and compressibility constaints. That the former is satisfied follows by (65), since $P_{d_1}(f) = Ed_1(X, f(X))$.

Similarly

$$\begin{aligned}
\frac{1}{n}H(\hat{X}^n) &\le \frac{1}{n}H(W,\hat{X}^n) \\
&= \frac{1}{n}H(W) + \frac{1}{n}H(\hat{X}^n|W) \\
&\overset{(a)}{\le} R_e + \frac{1}{n}\sum_{t=1}^{n} H(f_t(X_t,W)|W) && (66)\\
&\overset{(b)}{=} R_e + \frac{1}{n}E\left[\sum_{t=1}^{n} P_H(\tilde{f}_t(W))\right] && (67)\\
&\overset{(c)}{\le} R_e + EP_H(F^*) + \delta_n && (68)\\
&\overset{(d)}{=} R_e + H(F^*(X)|F^*) + \delta_n && (69)\\
&\overset{(e)}{<} R_c, && (70)
\end{aligned}$$

for sufficiently large n, where (a) follows from the fact that $\hat{X}_t = f_t(X_t, W)$ and the independence of the X_t, (b) through (d) follow from (64), and (e) follows, for sufficiently large n, since by assumption $R_e < R_c - H(F^*(X)|F^*)$. □

Discussion. Semicausal embedding is closely related to communication over channels with causal side information at the transmitter, as analyzed by Shannon in [11]. Specifically, we can interpret the covertext $X_1, \ldots, X_n$ as side information that is causally available to a transmitter trying to communicate over a clean channel, with input/output alphabet $\hat{\mathcal{X}}$, that is subject to a side information dependent "power" constraint induced by the distortion constraint, $(1/n)\sum_{t=1}^{n} Ed_1(X_t, \hat{X}_t) \le D_1$. The capacity of such a system can be inferred from the results of [11] to be the right hand side of (35), but without the term $R_c - H(F(X)|F)$. This latter term arises from the compressibility constraint on $\hat{X}^n$ which appears to have no counterpart in Shannon's problem [11].

If $I(F^*; F^*(X)) \le R_c - H(F^*(X)|F^*)$ then $H(F^*) \le R_c$. In this case, the compressibility constraint is especially easy to satisfy. In particular, for a code in which the empirical distribution of code symbols in all codewords is close to the distribution of F^* it is not hard to see that modeling $\hat{X}^n$ as an i.i.d. source with $\hat{X}_i$ distributed as $F^*(X)$ achieves an average code–length of approximately $H(F^*(X)) \le R_c$. Thus we see that in this case "memoryless" compression is optimal.

As in [6], an alternative characterization of the performance limits of semicausal (D_1, R_c) embedding can be obtained by minimizing the entropy of the stegotext subject to a lower bound on the embedding rate. From this point of view, an entropy rate R_c is (D_1, R_e)–achievable with semicausal codes if and only if R_e is (D_1, R_c)–achievable with semicausal codes. Let $R_c^*(D_1, R_e)$ be the infimum of all such (D_1, R_e)–achievable embedding entropy rates. Then it follows from (35) and Theorem 4.1 that

$$R_c^*(D_1, R_e) = R_e + \min_{\substack{F: Ed_1(X, F(X)) \leq D_1, \\ I(F; F(X)) \geq R_e}} H(F(X)|F). \tag{71}$$

For sufficiently small R_e the constraint $I(F; F(X)) \geq R_e$ becomes inactive in which case

$$\begin{aligned} R_c^*(D_1, R_e) &= R_e + \min_{F: Ed_1(X, F(X)) \leq D_1} H(F(X)|F) \\ &= R_e + r_c(D_1), \end{aligned} \tag{72}$$

where $r_c(D_1)$ is the Neuhoff–Gilbert causal rate–distortion function [10]. This is similar to the expression obtained in [6] for the noncausal version of $R_c^*(D_1, R_e)$, with $r_c(D_1)$ replacing $R(D_1)$, the rate–distortion function. Let

$$R_e^* = \max_{\substack{F: Ed_1(X, F(X)) \leq D_1, \\ H(F(X)|F) = r_c(D_1)}} I(F; F(X)). \tag{73}$$

Note that (72) holds for $R_e \leq R_e^*$. Thus we see, paralleling [6], that for $R_e \leq R_e^*$ embedding is "free of charge" in the sense that (72) would be the rate incurred by compressing the covertext using a Neuhoff-Gilbert code and appending the resulting compressed bitstream to a binary representation of the message W.

Additionally, it can be shown that

$$\min_{\substack{F: Ed_1(X, F(X)) \leq D_1, \\ I(F; F(X)) \geq R_e}} H(F(X)|F),$$

is a convex function of R_e for fixed D_1, which, in turn, implies that the constraint $I(F; F(X)) \geq R_e$ in the definition of $R_c(D_1, R_e)$ is met with equality for $R_e > R_e^*$. Therefore,

$$R_c^*(D_1, R_e) = \min_{F: Ed_1(X, F(X)) \leq D_1} H(F(X)), \tag{74}$$

for $R_e > R_e^*$, again in analogy with the properties of the noncausal $R_c^*(D_1, R_e)$ derived in [6].

4.2 Private embedding

We now modify the scenario of the previous subsection by giving the decompressor *and* the embedding decoder free access to the covertext sequence X^n. Specifically, the decoder is now of the form g :

$(\hat{\mathcal{X}}^n, \mathcal{X}^n) \to \{1, \ldots, M\}$ with $\hat{W} = g(\hat{X}^n, X^n)$, and we now require only that $(1/n)H(\hat{X}^n|X^n)$ be no larger than R_c. Adapting the definitions of a (D_1, R_c)–achievable semicausal embedding rate and the corresponding embedding capacity to the private setting, we have the following analogues of (35) and Theorem 4.1.

For a discrete memoryless covertext $\{X_t\}$ with marginal probability distribution function $P_X(x) = \Pr(X_t = x)$, let

$$C_{pr}(D_1, R_c) \triangleq \max_{F: Ed_1(X,F(X)) \leq D_1} \min[H(F(X)|X), R_c] \tag{75}$$

where F is independent of X and takes values in $\mathcal{F}$.

Theorem 4.2 *The semicausal (D_1, R_c) embedding capacity in the private setting is $C_{pr}(D_1, R_c)$.*

Proof.

Converse. A chain of inequalities paralleling (36) through (50) in the proof of Theorem 4.1, but with $\hat{X}_t$ replaced throughout with $(X_t, \hat{X}_t)$ leads to the corresponding upper bound $I(F; F(X), X)$ on R_e, with F and X independent and $Ed_1(X, F(X)) \leq D_1$. However,

$$\begin{aligned} I(F; F(X), X) &= H(F(X), X) - H(F(X), X|F) \\ &\stackrel{(a)}{=} H(F(X), X) - H(X) && (76) \\ &= H(F(X)|X), && (77) \end{aligned}$$

where (a) follows from the independence of X and F. In analogy to the second chain of inequalities in the proof of Theorem 4.1, we have

$$\begin{aligned} nR_c &\geq H(\hat{X}^n|X^n) \\ &\stackrel{(a)}{\geq} H(\hat{X}^n|X^n) + H(W|X^n, \hat{X}^n) - n\delta'_n && (78) \\ &= H(W|X^n) + H(\hat{X}^n|X^n, W) - n\delta'_n \\ &\stackrel{(b)}{\geq} nR_e - n\delta'_n && (79) \end{aligned}$$

where (a) follows from Fano's inequality, and (b) since $H(W|X^n) = H(W) \geq nR_e$ and since $\hat{X}^n$ is a function of X^n and W. We conclude that R_e also satisfies $R_e \leq R_c$, again since $\lim_{n\to\infty} \delta'_n = 0$.

Direct. The channel coding proof of the direct part of Theorem 4.1 also applies here, except that the relevant channel output is now $(X, F(X))$. The distortion constraint is satisfied in a similar manner, and $H(\hat{X}^n|X^n) \leq H(\hat{X}^n, W|X^n) = H(W) \leq nR_e$, since $\hat{X}^n$ is a function of W and X^n. $\square$

Discussion. The scenario in which the decoder $g(\cdot)$ is a function of $\hat{X}^n$ and X^n but the compressibility constraint reverts to $H(\hat{X}^n) \leq R_c$ may also be of interest. This would correspond to a three party interaction in which there are two types of receivers. One is a public receiver that should be able to decompress $\hat{X}^n$ without knowledge of X^n, but is not required to be able to decode the embedded message. The other is a private receiver like that considered above. Such a public/private system appears to be considerably more difficult to analyze, and is left for future work. The difficulty stems from the fact that the line of reasoning used to arrive at (79), and the analagous bounds in the converse proofs of the earlier theorems that account for the effect of the compressibility constraint, relies heavily on the receiver being required to decode the embedded message. This, however, is not the case for the public receiver.

4.3 Time invariant memoryless attacks

We now consider a scenario in which the decoder, instead of observing $\hat{X}^n$ directly, sees only a "forgery" $Y^n \in \mathcal{Y}^n$ corresponding to $\hat{X}^n$ corrupted by a time invariant DMC. Let $Q_{Y|\hat{X}}$ be the conditional output distribution for this channel, which is assumed to be fixed in advance and known to the transmitter and receiver. The definitions of a (D_1, R_c)–achievable embedding rate with semicausal codes and of the resulting embedding capacity carry over from the attack free case. Theorem 4.3 below characterizes the semicausal (D_1, R_c) embedding capacity under a known time invariant memoryless attack.

For a discrete memoryless covertext $\{X_t\}$ with marginal probability distribution function $P_X(x) = \Pr(X_t = x)$, and a DMC with conditional output distribution $Q_{Y|\hat{X}}$ let

$$C(D_1, R_c; Q_{Y|\hat{X}}) \stackrel{\triangle}{=} \max_{F: Ed_1(X, F(X)) \leq D_1} \min[I(F;Y), R_c - H(F(X)|F)] \tag{80}$$

where the joint distribution of X, F, and Y is given by

$$P_{X,F,Y}(x, f, y) = P_X(x) P_F(f) Q_{Y|\hat{X}}(y|f(x)) \tag{81}$$

and F takes values in $\mathcal{F}$ defined as the set of functions $f : \mathcal{X} \to \hat{\mathcal{X}}$.

Theorem 4.3 *The semicausal (D_1, R_c) embedding capacity under under time invariant discrete memoryless attacks distributed according to $Q_{Y|\hat{X}}$ is given by $C(D_1, R_c; Q_{Y|\hat{X}})$.*

The proof of this theorem is a straightforward extension of the proof of Theorem 4.1 above (which is a special case). We omit the details.

4.4 Informed blockwise memoryless attacks

Next we allow the attacker to select a constrained block based attack strategy to maximize the decoding error probability based on full knowledge of the semicausal embedding code and decoding algorithm. The particular attack chosen is assumed to be unknown to the encoder and decoder. The attacker is constrained to be memoryless from one block to the next, to use the same strategy in each block, and to avoid excessively distorting the reconstruction signal. Specifically, for $n = mk$, the conditional distribution of the "forgery" Y^n induced by a k–block attack on a reconstruction signal $\hat{X}^n = \hat{x}^n$ is

$$Q_{Y^n|\hat{X}^n}(y^n|\hat{x}^n) = \prod_{j=0}^{m-1} Q_{Y^k|\hat{X}^k}(y_{jk+1}^{(j+1)k}|\hat{x}_{jk+1}^{(j+1)k}), \tag{82}$$

where $Q_{Y^k|\hat{X}^k}$ characterizes the blockwise operation of the k–block attacker. Given a per–letter distortion measure $d_3(\hat{x}, y)$ between the reconstruction and forgery, we require that the per–block attack $Q_{Y^k|\hat{X}^k}$ belong to the set $\mathcal{Q}^k(D_3)$, where $\mathcal{Q}^k(D_3)$ is defined by (15) in Section 3.3.

The above model is intended to cover attackers with limited resources that are constrained to operate on small chunks of the stegotext at any one time, and do so without retaining state information from one chunk to the next. Many signal processing algorithms are essentially of this form, as are popular lossy compression algorithms such as JPEG and MPEG. We note that a similar blockwise memoryless attack model is also assumed in [9].

An important aspect of the above blockwise memoryless model is that even though the encoder and decoder are assumed to be ignorant of the particular blockwise attack selected, the traditional notions of channel capacity remain relevant thanks to the existence of universal channel decoding algorithms such as the maximum mutual information (MMI) decoding algorithm for (blockwise) memoryless channels. The restriction of blockwise attacks to belong to $\mathcal{Q}^k(D_3)$ is somewhat technical. It results in a large class of attacks that still allows a single letter saddle–point expression characterizing embedding rates that are possible under the present scenario, as well its dual, treated in Subsection 4.5 below, in which the attacker tries to minimize the maximum embedding rate, where the embedding code is selected with full knowledge of the attack.

We say that an embedding rate R_e is (D_1, R_c)–achievable with semicausal codes under attacks from $\mathcal{Q}^k(D_3)$ if there exists a sequence of $(\lceil 2^{nR_e} \rceil, n)$ semicausal encoder and decoder pairs such that the maximum error probability P_e induced by *all* k–block attacks in $\mathcal{Q}^k(D_3)$, applied according to (82), tends to zero. The semicausal restriction on the encoder along with the compressibility and distortion constraints are retained from the attack free case. Let

$$C(D_1, R_c, D_3) \stackrel{\triangle}{=} \max_{F: Ed_1(X, F(X)) \le D_1} \min_{Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)} \min[I(F;Y), R_c - H(F(X)|F)], \tag{83}$$

where the joint distribution of X, F, and Y is given by (81). Define the semicausal (D_1, R_c) embedding capacity with respect to $\mathcal{Q}^k(D_3)$ as the supremum of all embedding rates R_e that are (D_1, R_c)–achievable with semicausal codes under attacks from $\mathcal{Q}^k(D_3)$.

Theorem 4.4 *The semicausal (D_1, R_c) embedding capacity with respect to $\mathcal{Q}^k(D_3)$ is given by $C(D_1, R_c, D_3)$, independently of k.*

Proof.

Converse. Suppose the embedding rate R_e is (D_1, R_c)–achievable with semicausal codes under attacks from $\mathcal{Q}^k(D_3)$. Then, by definition, there is a sequence of $(\lceil 2^{nR_e} \rceil, n)$ semicausal embedding codes of block length n having error probability no larger than δ_n satisfying $\lim_{n\to\infty} \delta_n = 0$ for every attack channel in $\mathcal{Q}^k(D_3)$ applied according to (82), expected average distortion no larger than D_1, and $(1/n)H(\hat{X}^n) \le R_c$. The error probability bound δ_n holds, in particular, for every time invariant DMC with transition matrix $Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)$. Therefore, we can apply reasoning similar to that used in the converse proof of Theorem 4.1 to obtain, for each n, a random variable F taking values in $\mathcal{F}$, having a distribution that does not depend on $Q_{Y|\hat{X}}$ (just on the code and source distribution), and satisfying $Ed_1(X, F(X)) \le D_1$ and

$$R_e \le \min[I(F;Y), R_c - H(F(X)|F)] + \delta'_n \tag{84}$$

with $P_{XFY}(x, f, y) = P_X(x)P_F(f)Q_{Y|\hat{X}}(y|f(x))$, for all channels $Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)$, and δ'_n defined as in the converse proof of Theorem 4.1. Therefore,

$$\begin{aligned} R_e &\le \min_{Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)} \min[I(F;Y), R_c - H(F(X)|F)] + \delta'_n \\ &\le C(D_1, R_c, D_3) + \delta'_n, \end{aligned} \tag{85}$$

and since n can be chosen to make δ'_n arbitrarily small, the converse is proved.

Direct. We will use the channel coding approach of the proof of the direct part of Theorem 4.3, except that here, since a single encoder/decoder pair must be reliable under any attack in $\mathcal{Q}^k(D_3)$ a more suitable model is that of the compound DMC ([3] Ch. 2, Sec. 5). The compound DMC of interest is defined on k–blocks of symbols and consists of the channels $Q_{Y^k|F^k}$ determined by cascading the (random) mapping $F^k \to [F^k(X^k) = \hat{X}^k]$ (assuming components of X^k are i.i.d. $\sim P_X$) with $Q_{Y^k|\hat{X}^k} \in \mathcal{Q}^k(D_3)$ where the t–th component of F^k takes values in the set of functions $f : \mathcal{X}^t \to \hat{\mathcal{X}}$ and the t–th component of $F^k(X^k)$ is $F_t(X^t)$.

Let F^* achieve the maximum in (83) and define $[F^*]^k = F_1^*, \ldots, F_k^*$ to be i.i.d. with components F_t^* having the same distribution as F^*. Additionally, let the t–th component of $[F^*]^k(X^k)$ be $F_t^*(X_t)$. Let R_e satisfy

$$kR_e < \min_{Q_{Y^k|\hat{X}^k} \in \mathcal{Q}^k(D_3)} \min[I([F^*]^k; Y^k), kR_c - H([F^*]^k(X^k)|[F^*]^k)]. \tag{86}$$

Then the arguments behind Corollary 5.10 of [3] (principally Theorem 5.2) imply the existence of a sequence of $(\lceil 2^{mkR_e} \rceil, mk)$ codes having the following properties for the above compound DMC. Under MMI decoding (of the empirical distribution of k–blocks), the probability of error vanishes uniformly for all attacks/channels in the compound DMC family. The codewords, on a symbol–wise basis, satisfy the generalized power constraints (64) and (65) with the expectations evaluated with respect to the present distribution of F^*.

Following the second half of the proof of the direct part of Theorem 4.1, we see that the above sequence of channel codes for the above compound DMC determines a sequence of semicausal $(\lceil 2^{nR_e} \rceil, n)$ embedding codes, which, when decoding under the corresponding MMI channel decoder, yield a vanishing error probability for every attack in $\mathcal{Q}^k(D_3)$. That the resulting sequence of embedding codes satisfies the distortion and compressibility constraints follows from the same reasoning as in the proof of Theorem 4.1.

To complete the direct part, it suffices to show that

$$kC(D_1, R_c, D_3) = \min_{Q_{Y^k|\hat{X}^k} \in \mathcal{Q}^k(D_3)} \min[I([F^*]^k; Y^k), kR_c - H([F^*]^k(X^k)|[F^*]^k)], \tag{87}$$

since then any $R_e < C(D_1, R_c, D_3)$ also satisfies (86).

To see (87) first note that $kR_c - H([F^*]^k(X^k)|[F^*]^k)$ is independent of $Q_{Y^k|\hat{X}^k}$ and satisfies

$$kR_c - H([F^*]^k(X^k)|[F^*]^k) = k(R_c - H(F^*(X)|F^*)), \tag{88}$$

since the components of $[F^*]^k$, X^k, and $[F^*]^k(X^k)$ are i.i.d. Additionally, for a fixed $Q_{Y^k|\hat{X}^k} \in \mathcal{Q}^k(D_3)$, we have the following chain of inequalities (similar to the converse proof in the rate–distortion coding theorem).

$$\begin{aligned}
I([F^*]^k;Y^k) &= H([F^*]^k) - H([F^*]^k|Y^k) \\
&\overset{(a)}{=} \sum_{t=1}^{k} H(F_t^*) - H([F^*]^k|Y^k) && (89)\\
&\overset{(b)}{\geq} \sum_{t=1}^{k} I(F_t^*;Y_t) && (90)\\
&\overset{(c)}{\geq} kI(F^*;Y) && (91)\\
&\overset{(d)}{\geq} k \min_{Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)} I(F^*;Y), && (92)
\end{aligned}$$

where the joint distribution of F^*, X, and Y satisfies

$$\begin{aligned}
P_{F^*,X,Y}(f,x,y) &= \frac{1}{k}\sum_{t=1}^{k} P_{F_t^*,X,Y_t}(f,x,y) \\
&= \frac{1}{k}\sum_{t=1}^{k} P_{F^*}(f)P_X(x)P_{Y_t|F_t^*,X_t}(y|f,x) \\
&= P_{F^*}(f)P_X(x)\frac{1}{k}\sum_{t=1}^{k} P_{Y_t|F_t^*,X_t}(y|f,x), && (93)
\end{aligned}$$

and (a) follows from the fact that the components of $[F^*]^k$ are independent, (b) – chain rule and removing the conditioning on $([F^*]^{t-1}, Y_1^{t-1}, Y_{t+1}^k)$, (c) – since the F_t^* all have the same distribution as F^* and since $I(F;Y)$ is convex in $\{P_{Y|F}(\cdot|\cdot)\}$ for fixed $P_F(\cdot)$. To justify (d) it suffices to show that

$$\begin{aligned}
P_{Y|F,^*X}(y|f,x) &= \frac{1}{k}\sum_{t=1}^{k} P_{Y_t|F_t^*X}(y|f,x) && (94)\\
&= Q_{Y|\hat{X}}(y|f(x)) && (95)
\end{aligned}$$

for some $Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)$, which in turn would follow from showing that

$$P_{Y_t|F_t^*,X}(y|f,x) = Q_{Y_t|\hat{X}_t}(y|f(x))$$

for some $Q_{Y_t|\hat{X}_t} \in \mathcal{Q}^1(D_3)$. The latter is established by

$$\begin{aligned}
&P_{Y_t|F_t^*,X_t}(y|f,x) \\
&= \frac{\sum_{y^k,x^k,\tilde{f}^k:(y_t,\tilde{f}_t,x_t)=(y,f,x)} Q_{Y^k|\hat{X}^k}(y^k|\tilde{f}^k(x^k)) \prod_{j\in\{1,\ldots,k\}} P_{F^*}(\tilde{f}_j)P_X(x_j)}{P_{F^*}(f)P_X(x)} \\
&= \sum_{y^k,x^k,\tilde{f}^k:(y_t,\tilde{f}_t,x_t)=(y,f,x)} Q_{Y^k|\hat{X}^k}(y^k|\tilde{f}^k(x^k)) \prod_{j\in\{1,\ldots,k\}\setminus t} P_{F^*}(\tilde{f}_j)P_X(x_j) && (96)\\
&\triangleq Q_{Y_t|\hat{X}_t}(y|f(x)), && (97)
\end{aligned}$$

where

$$\tilde{f}^k(x^k) \stackrel{\triangle}{=} \tilde{f}_1(x_1), \ldots, \tilde{f}_k(x_k),$$

and (97) follows since (96) depends on f and x only through $f(x)$. Additionally, since $Q_{Y^k|\hat{X}^k} \in \mathcal{Q}^k(D_3)$, explicitly computing the expectation of $d_3(f(x), Y_t)$ according to (96), combined with the definition (15) of $\mathcal{Q}^k(D_3)$ (with $P_{\hat{X}^n}$ in the definition of $\mathcal{Q}^k(D_3)$ induced by $\hat{X}_j = \tilde{F}_j(X_j)$ with $\tilde{F}_j, X_j$ i.i.d. according to $P_{F^*}P_X$), shows that $Q_{Y_t|\hat{X}_t} \in \mathcal{Q}^1(D_3)$.

That the right side of equation (87) exceeds the left now follows by combining (88) and (92), and from the fact that (92) holds for all $Q_{Y^k|\hat{X}^k} \in \mathcal{Q}^k(D_3)$. Equality follows by noting that the left side of (87) corresponds to the right side with the outer minimization restricted to stationary DMCs with component channels in $\tilde{\mathcal{Q}}^1(D_3)$. □

Discussion. A different attack model considered in [12] is to require that

$$\Pr\left(\sum_{t=1}^{n} d_3(\hat{X}_t, Y_t) > nD_3 \,\middle|\, \hat{X}^n = \hat{x}^n\right) = 0 \text{ for all } \hat{x}^n \in \hat{\mathcal{X}}^n, \tag{98}$$

where n is the full block-length of the code. Adapting the above definitions of (D_1, R_c)–achievable embedding rates to this attack model, we conjecture (paralleling [12]) that the related embedding capacity is given by

$$C_{SM}(D_1, R_c, D_3) \stackrel{\triangle}{=} \max_{F: Ed_1(X, F(X)) \leq D_1} \quad \min_{Q_{Y|\hat{X}}: Ed_3(F(X), Y) \leq D_3} \min[I(F; Y), R_c - H(F(X)|F)], \tag{99}$$

where the joint distribution of X, F, and Y is given by (81). Note that the set over which $Q_{Y|\hat{X}}$ is minimized in (99) is larger than the set $\mathcal{Q}^1(D_3)$ appearing in (83), and, in general, depends on the distribution of F.

4.5 The minimax game

For completeness, we consider the minimax counterpart to the attack model of the previous subsection. In this setting the roles of attacker and embedder are reversed, in the sense that it is the embedder that is free to select the encoding and decoding strategies based on full knowledge of the k–block attack that will be used. In this case, the attacker should select the attack which minimizes the induced (D_1, R_c) embedding capacity. We refer to the resulting minimum as the minimax k–block (D_1, R_c) embedding capacity. For the class $\mathcal{Q}^k(D_3)$ defined by (15) the minimax k–block (D_1, R_c) embedding capacity is readily seen to be bounded from below by $C(D_1, R_c, D_3)$. Define

$$\Lambda(F, Q_{Y|\hat{X}}) = \min[I(F; Y), R_c - H(F(X)|F)].$$

Restricting attacks to be DMC type attacks selected from $\mathcal{Q}^1(D_3)$, in turn, establishes, via arguments similar to those used in the converse proof of Theorem 4.1, that the minimax (D_1, R_c) capacity is bounded from above by

$$\min_{Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)} \max_{F: Ed_1(X,F(X)) \leq D_1} \Lambda(F, Q_{Y|\hat{X}}).$$

Proposition 4.5 below shows that this quantity is equal to

$$\max_{F: Ed_1(X,F(X)) \leq D_1} \min_{Q_{Y|\hat{X}} \in \mathcal{Q}^1(D_3)} \Lambda(F, Q_{Y|\hat{X}}) = C(D_1, R_c, D_3),$$

thereby establishing, with the lower bound, that the minimax k–block (D_1, R_c) embedding capacity is also $C(D_1, R_c, D_3)$. It may seem that Proposition 4.5 should follow immediately from the minimax theorem of convex analysis, since the constraint sets involved are convex, and since both $I(F;Y)$ and $R_c - H(F(X)|F)$ have the required convexity/concavity properties in the variables over which the minimum and maximum are taken. The difficulty stems from the fact that the minimum of two convex functions is, in general, not convex.

Proposition 4.5 *For any convex family of channels $\mathcal{Q}$,*

$$\begin{aligned}
\max_{F: Ed_1(X,F(X)) \leq D_1} \min_{Q_{Y|\hat{X}} \in \mathcal{Q}} \Lambda(F, Q_{Y|\hat{X}}) &= \min_{Q_{Y|\hat{X}} \in \mathcal{Q}} \max_{F: Ed_1(X,F(X)) \leq D_1} \Lambda(F, Q_{Y|\hat{X}}) &\quad (100)\\
&= \Lambda(F^*, Q^*_{Y|\hat{X}}), &\quad (101)
\end{aligned}$$

where F^* and $Q^*_{Y|\hat{X}}$ respectively achieve the maximum and minimum in the left and right hand sides of (100).

Proof. Let

$$\Lambda(F, Q_{Y|\hat{X}}, \theta) = \theta I(F;Y) + (1-\theta)(R_c - H(F(X)|F)).$$

Note that $\Lambda(F, Q_{Y|\hat{X}}, \theta)$ is concave in F, convex in $Q_{Y|\hat{X}}$, and linear in θ, when the remaining two respective arguments are fixed. Then

$$\begin{aligned}
&\max_{F:Ed_1(X,F(X))\le D_1} \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \Lambda(F, Q_{Y|\hat{X}}) \\
&\quad = \max_{F:Ed_1(X,F(X))\le D_1} \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \min_{\theta\in[0,1]} \Lambda(F, Q_{Y|\hat{X}}, \theta) && (102)\\
&\quad = \max_{F:Ed_1(X,F(X))\le D_1} \min_{\theta\in[0,1]} \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \Lambda(F, Q_{Y|\hat{X}}, \theta) \\
&\quad \overset{(a)}{=} \min_{\theta\in[0,1]} \max_{F:Ed_1(X,F(X))\le D_1} \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \Lambda(F, Q_{Y|\hat{X}}, \theta) && (103)\\
&\quad \overset{(b)}{=} \min_{\theta\in[0,1]} \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \max_{F:Ed_1(X,F(X))\le D_1} \Lambda(F, Q_{Y|\hat{X}}, \theta) && (104)\\
&\quad = \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \min_{\theta\in[0,1]} \max_{F:Ed_1(X,F(X))\le D_1} \Lambda(F, Q_{Y|\hat{X}}, \theta) \\
&\quad \overset{(c)}{=} \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \max_{F:Ed_1(X,F(X))\le D_1} \min_{\theta\in[0,1]} \Lambda(F, Q_{Y|\hat{X}}, \theta) && (105)\\
&\quad = \min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \max_{F:Ed_1(X,F(X))\le D_1} \Lambda(F, Q_{Y|\hat{X}}), && (106)
\end{aligned}$$

where (a), (b), and (c) follow from repeated applications of the mini-max theorem of convex analysis, which is applicable since all of the sets and functions satisfy the assumptions of the theorem (sets are convex, functions are concave (resp. convex) in the argument over which the maximum (resp. minimum) is taken). Step (a), in particular, is justified since $\min_{Q_{Y|\hat{X}}\in\mathcal{Q}} \Lambda(F, Q_{Y|\hat{X}}, \theta)$ is still linear in θ (the minimizing Q depends only on F and not on θ) and also concave in F as the minimum of a set of concave functions is concave.

The saddlepoint claim (101) follows by noting that

$$\Lambda(F^*, Q^*_{Y|\hat{X}}) \le \max_F \Lambda(F, Q^*_{Y|\hat{X}}) = \min_{Q_{Y|\hat{X}}} \max_F \Lambda(F, Q_{Y|\hat{X}})$$

and

$$\Lambda(F^*, Q^*_{Y|\hat{X}}) \ge \min_{Q_{Y|\hat{X}}} \Lambda(F^*, Q_{Y|\hat{X}}) = \max_F \min_{Q_{Y|\hat{X}}} \Lambda(F, Q_{Y|\hat{X}}).$$

□

5 Future work

One direction for future work is to extend the semicausal (D_1, R_c) embedding rate analysis to the attack model considered in [12]. A conjecture for the resulting semicausal (D_1, R_c) capacity appears in

Section 4.4. Another direction is to extend the analysis of Section 4.2 of private semicausal embedding to a semi–private version, in which the decompressor and decoder are distinct parties with only the decoder having access to the covertext X^n.

References

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

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

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems,* Akadémiai Kiadó, Budapest, 1981.

[4] A. Dembo and O. Zeitouni *Large Deviations Techniques and Applications,* Jones and Bartlett Publishers, London, 1993.

[5] D. Karakos and A. Papamarcou, *A relationship between quantization and watermarking rates in the presence of additive Gaussian attacks*, *IEEE Trans. Inform. Theory*, August 2003.

[6] A. Maor and N. Merhav, *On joint information embedding and lossy compression,* submitted to *IEEE Trans. Inform. Theory,* July 2003.
See also `http://www.ee.technion.ac.il/people/merhav`.

[7] 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. See also `http://www.ee.technion.ac.il/people/merhav`.

[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, IEEE Trans. Inform. Theory,* vol. 49, no. 3, pp. 563–593, March 2003.

[10] D. Neuhoff and R. Gilbert, *Causal source codes, IEEE Trans. Inform. Theory,* vol. 28, no. 5, pp. 701–713, Sep. 1982.

[11] C. E. SHANNON, *Channels with side information at the transmitter, IBM Journal,* pp. 289–293, October, 1958.

[12] A. SOMEKH-BARUCH AND N. MERHAV, *On the capacity game of public watermarking systems,* to appear *IEEE Trans. Inform. Theory*, March 2004.

[13] F. WILLEMS AND T. KALKER, *Reversible embedding methods*, *Proc. 40th Allerton Conference on Communications Control and Computing*, Monticello, IL, October 2002.