Coding into a source: a direct inverse Rate-Distortion theorem

Mukul Agarwal, Anant Sahai, and Sanjoy Mitter

***Abstract*— Shannon proved that if we can transmit bits reliably at rates larger than the rate distortion function $R(D)$, then we can transmit this source to within a distortion D. We answer the converse question "If we can transmit a source to within a distortion D, can we transmit bits reliably at rates less than the rate distortion function?" in the affirmative. This can be viewed as a direct converse of the rate distortion theorem.**

I. Introduction

In [1], Shannon proved that if there is a channel with capacity $C > R(D)$, a source can be transmitted to within a distortion D reliably over this channel ($R(D)$ is the rate distortion function for the source) in two steps:

1) Suppose $C = R(D - \alpha)$. First, source code to within a distortion $(D - \frac{\alpha}{2})$ by using random codes. The source code has rate arbitrarily close to $R(D - \frac{\alpha}{2})$.
2) Transmit these bits reliably[1] over the channel.

The traditional converse to this separation theorem is proved using the data-processing inequality and shows that no other joint source-channel scheme can do any better.

We want to instead ask the converse question at the engineering level[2]: if there is a "black box" over which an iid source $X_i \sim p_X$ can be transmitted to within a distortion level D, can we do reliable communication of bits (in the Shannon sense) over this "black box" at rates less than $R(D)$?

If one assumes that the communication of X_i over the black box satisfies only an expected distortion constraint $Ed(X_i, \hat{X}_i) \leq D$, then we **cannot** guarantee reliable communication. The black box should be viewed as an attacker and the attacker can do anything that it wishes as long as it meets the expected distortion constraint.

Consider an equiprobable binary source $\{0, 1\}$ under the Hamming distortion. Suppose the black box is constrained to communicate this source to within an expected distortion of 0.25. A possible attacker could flip a fair coin once at the beginning of time. If it is heads, then it transmits the symbols perfectly for all time; if it is tails, it just transmits 0 for all time. It is then easy to see that one cannot do reliable communication over this attacker at any non-zero rate, whereas the rate-distortion $R(0.25) > 0$.

Thus, the expected distortion constraint is not sufficient. It turns out that a block distortion constraint is sufficient. If the attacker is such that[3]

$$\Pr\left(\frac{1}{n}\sum_{t=1}^{n} d(X_t, Y_t)\right) > D \to 0 \text{ as } n \to \infty \qquad (1)$$

it can be proved that reliable communication is possible over this attacker at all rates less than $R(D)$. This is the main theorem of this paper which is stated formally in Section III.

Following [2], one can draw an equivalence between all rate-distortion problems with a given value of $R(D)$. Consider the collection of all iid sources and corresponding distortion levels, (C_β, D_β) such that $R_{C_\beta}(D_\beta) = R_0$. If any one of these sources can be communicated over an attacker such that the block distortion criterion (1) holds, then all of them can be communicated to within a distortion level $(D_\beta + \delta)$ over this same attacker, for arbitrarily small positive δ. One way to show this is:

1) Source code one source to within the distortion level $D_\beta + \delta$ by using less than nR_0 bits.
2) Communicate these nR_0 bits reliably by embedding them into the source accepted by the attacker and recovering them from the distorted sequence.

In Section II, we state the precise formulation of the above problem. In Section III, we state our main theorem. In Section IV, we state the connection of the formulated problem to coding theory, arbitrarily varying channels and to watermarking with no covertext. In Section V, we prove the theorems stated in Section III and comment on them in Section VI. Section VII formulates a conditional version of the theorem and it is proved in Section VIII. In Section IX, we state the relation of this problem to watermarking. Section X shows how to generalize to the case of non-finite sources with difference distortion. Section XI, shows how the results can be easily extended to stationary ergodic sources that mix appropriately.

Because of space limitations, some of the details in the later sections are omitted. The full proofs can be found in [3].

II. Problem Formulation - Unconditional case

We start with some notation:

- $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\} \to$ finite set. $\mathcal{X}^\infty$ is the input space.
- $\mathcal{Y} = \{1, 2, \ldots, |\mathcal{Y}|\} \to$ finite set. $\mathcal{Y}^\infty$ is the output space.

Mukul Agarwal is an EECS student at MIT. The core of this work was performed largely while he was visiting Prof. Sahai at Wireless Foundations at UC Berkeley. magar@mit.edu

Anant Sahai is with Wireless Foundations in EECS at UC Berkeley sahai@eecs.berkeley.edu

Sanjoy Mitter is with LIDS in EECS at MIT mitter@mit.edu

[1] The bounded nature of the distortion function only becomes important if we are interested in end-to-end expected distortion. If all that is desired is for the probability of excess distortion to be arbitrarily small, then no such assumptions are needed.

[2] Fundamentally, we are asking whether reliable lossless communication is necessarily the right primitive that defines layering in a multipurpose communication system. Could lossy coding serve as an equally good primitive in principle?

[3] For simplicity of notation, the dependence of the attacker on block-length n is suppressed. To be precise, (1) should be interpreted as a family of attackers indexed by n such that the probability of excess distortion can be made as close to zero as desired by choosing an attacker with an appropriately large n. This parallels the existence result for channel coding.

- $p_X \rightarrow$ probability distribution on $\mathcal{X}$.
- $X_1^\infty \rightarrow$ iid sequence of random variables, each $X_i \sim p_X$.
- $d : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{R}$ is a non-negative valued function. We should think of $d(i,j)$ as the distortion between $i \in \mathcal{X}$, $j \in \mathcal{Y}$. The focus is on the average additive distortion on n-sequences, $\frac{1}{n} d_n(x_1^n, y_1^n) = \frac{1}{n} \sum_{t=1}^n d(x_t, y_t)$.
- The Attacker is a black box which takes in the input sequence $x_1^\infty \in \mathcal{X}_1^\infty$ and produces an output $y_1^\infty \in \mathcal{Y}_1^\infty$. y_1^∞ need not be a deterministic function of x_1^∞; it can be randomized.

  Note that the attacker is, in general, non-causal in the sense that it takes in the whole input sequence, looks at it, and produces an output sequence. The situation that the attacker looks at x_1^∞ and produces y_1^∞ is the most general possible. In practice, the attacker will only look at finite length sequences and produce an output; this is a special case of our definition.

  The attacker can also be viewed as a channel. We will use the words attacker and attack channel interchangeably.
- D-distortion attacker $\rightarrow$ If the input to the attacker is the random variable sequence X_1^∞ (defined above - each X_i iid p_X), the attacker produces the random variable sequence Y_1^∞. This results in a joint probability measure on (X_1^∞, Y_1^∞). Under this probability measure, there should exist some function $f(n)$ with $\lim_{n\rightarrow\infty} f(n) = 0$ so that:

$$\sup_t \Pr\left(\frac{1}{n}\sum_{u=t}^{t+n-1} d(X_u, Y_u) > D\right) \leq f(n) \qquad (2)$$

  The above equation says that the average distortion caused to long sequences is bounded by D with high probability, and this probability $\rightarrow 1$ at least as fast[4] as $1 - f(n)$ with increasing block lengths n uniformly over at which time this sliding block[5] is taken (hence, the name D-distortion attacker).

  Note that on an individual symbol level, the attacker is essentially unconstrained — for any X_t, the attacker can distort it really badly. It is only constrained over very long blocks.
- $p_X \pm \epsilon$ will denote the set of all probability measures q_X on X such that $|q_X(i) - p_X(i)| \leq \epsilon \forall i \in \mathcal{X}$.

As we can see, the rate-distortion problem when the input sequence is iid p_X is solved (in the sense of [2] by this attacker for distortion value D. The question we want to ask is, "Can we transmit bits reliably over this attacker in the Shannon sense, and if yes, at what rates?"

III. Main results - unconditional case

Theorem 1: Assuming that there is common randomness available at the transmitter and the receiver, all rates

$$R < R_X(D) \triangleq \inf_{\substack{X \sim p_X \\ Ed(X,Y) \leq D}} I(X;Y) \qquad (3)$$

are achievable over a D-distortion attack channel, and in fact, this can be done by using iid p_X random codes.

The above theorem says that we can solve the Shannon communication problem over a D-distortion attacker at all rates less than the rate distortion function, $R_X(D)$. We comment on the need for common randomness in Section VI after we prove the above theorem.

We also have a converse theorem:

Theorem 2: Rates larger than $R_X(D)$ can in general not be achieved over a D-distortion attacker.

After a few comments about this formulation in the next section, it is proved in the section after next.

IV. Connections to AVCs and Watermarking

We can view the attacker as a non-causal arbitrarily varying channel (AVC). The AVC is constrained in such a way that it distorts *most* input sequences to an average distortion less than or equal to D where "most" is according to the iid p_X measure over the input sequences. The question that we are asking is, "What is the capacity of this AVC?" The foundational papers on AVCs are the papers by Blackwell, Breiman and Thomasian, [4], [5]. [4] considers the case when the channel is a fixed DMC coming from a particular set, but unknown. [5] considers the case when the channel can vary arbitrarily, but is a DMC at each time, and comes from a particular set based on past history unlike in our case where the attack channel at each time does not come from a particular set, nor is it causal. Stiglitz [6] has the same setup as [5], but calculates error exponents. Csiszar and Narayan [7] uses a minimum distance decoding rule similar to the one that we will use, but it does not consider AVCs in the form that we do.

To the extent that minimum distance is the relevant idea, this work can also be considered a generalization of the original formulation of coding theory in [8] with the distortion measure generalizing the Hamming distance. In addition, the composition of the codewords is specified in advance. Fundamentally, Theorem 1 says that every rate-distortion problem is also associated with a coding theory problem.

This paper's formulation can also be viewed as a watermarking problem ([9]) with no covertext. The goal is to embed our data in the input to an attacker that acts within a distortion constraint. [10] by Somekh-Baruch and Merhav is the closest to our work. It allows for non-causal attackers and the definition of attacker is very similar to ours. But [10] does not use a minimum distortion decoding rule — they use another decoding rule which is superior in the sense that it achieves the best possible error exponent. We believe that proofs in [10], with slight modification, should be applicable

[4] No restrictions are made on how fast $f(n)$ tends to zero — just that we know how fast this probability goes to zero for this particular family of attackers so that we can pick an appropriate block-length for the code.

[5] The purpose of the sliding block is merely to reduce notation in stating the condition. All theorems will be proved within a single block of length n that is sufficiently long on its own. This can be repeated with disjoint blocks if a stream of data needs to be transmitted.

in our scenario too, but we use a different decoding rule (a variant of minimum distance decoding) since it is arguably more natural and achieves capacity. The distinction between the two papers is more significant in the conditional case.

V. Proofs - unconditional case

We first prove Theorem 1 stated in Section III and show that by using p_X random codes, we can transmit reliably (in the Shannon sense) at all rates $R < R_X(D)$ over the D-distortion attack channel.

Codebook Construction: Generate 2^{nR} codewords iid p_X. This is the codebook, which we denote by $\mathcal{C}$.

Decoding: Fix $\epsilon > 0$. Restrict attention to those codewords which are p_X-typical, that is, whose type lies in $p_X \pm \epsilon$ (recall the definition of $p_X \pm \epsilon$ in Section II: all q_X such that $|q_X(i) - p_X(i)| \leq \epsilon \forall i \in \mathcal{X}$).

Denote this restricted set of codewords by $\mathcal{C}_R$.

Let y_1^n denote the output of the attacker. If there is a unique p_X-typical x_1^n in the codebook which is at an average distortion less than or equal to D from the output sequence, declare that x_1^n was transmitted, else declare error.

We call our decoding rule the "ϵ-Nearest Typical Neighbor" decoding rule. The truly nearest neighbor decoding rule might be a bit more natural, but it is harder to analyze.

In what follows,

- x_1^n denotes the transmitted codeword.
- y_1^n denotes the received sequence (output of the attacker).
- z_1^n denotes a p_X typical codeword (that is, $z_1^n \in \mathcal{C}_R$) such that z_1^n is **NOT** transmitted.

The error event can be decomposed into 3 parts.

- $E_1 \rightarrow$ transmitted codeword atypical: $x_1^n \notin \mathcal{C}_R$.
- $E_2 \rightarrow$ Distortion caused by the attacker is not typical: $\frac{1}{n}\sum_{t=1}^{n} d(x_t, y_t) > D$.
- $E_3 \rightarrow$ a typical codeword which is not transmitted is at an average distortion less than or equal to D from the received sequence. Mathematically, $\exists z_1^n \in \mathcal{C}_R$ such that z_1^n is not transmitted and $\frac{1}{n}\sum_{t=1}^{n} d(x_t, y_t) \leq D$.

Clearly, $\Pr(\text{error}) \leq \Pr(E_1) + \Pr(E_2) + \Pr(E_3)$. By the weak law of large numbers, $\Pr(E_1) \rightarrow 0$ as $n \rightarrow \infty$. $\Pr(E_2) \rightarrow 0$ as $n \rightarrow \infty$ follows by the definition of D-distortion attacker (2). To upper bound $\Pr(E_3)$, we do a type-based calculation [11] on the probability of error for a given received sequence y_1^n.

In what follows, it will be helpful to remember that q will always denote probability measures with *observed* types, whereas p will always denote probability measures with *transmitted* types. Recall that the received sequence is y_1^n. Let the type of y_1^n be q_Y, that is, $\forall j \in \mathcal{Y}$, the number of j occurring in y_1^n is $nq_Y(j)$.

Sort the output to place all the $j \in \mathcal{Y}$ together, and correspondingly shuffle the positions in the codebook's codewords. This leads to no change in distortion between shuffled codewords and the sorted received sequence y_1^n.

Look at a generic shuffled codeword $z_1^n \in \mathcal{C}_R$ which is not transmitted. Over the chunk of length $nq_Y(j)$, let the type of the corresponding entries of z_1^n be $q_{X|Y=j}$. (See Figure 1)

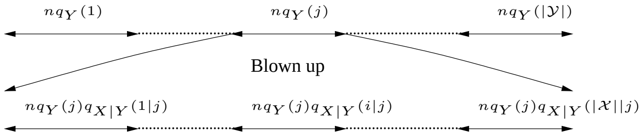

Fig. 1. The sorted received sequence y_1^n and the correspondingly shuffled codeword z_1^n illustrating the relevant types.

For the error event E_3,

1) z_1^n is typical, that is,
$$\sum_{j \in \mathcal{Y}} q_Y(j) q_{X|Y}(i|j) \in p_X \pm \epsilon \forall i \in \mathcal{X} \tag{4}$$
Denote $\sum_{j \in \mathcal{Y}} q_Y(j) q_{X|Y}(i|j)$ as $q_X(i)$. Thus,
$$q_X \in p_X \pm \epsilon \tag{5}$$

2) z_1^n is at an average distortion $\leq D$ from the received sequence y_1^n so
$$\sum_{i \in \mathcal{X}, j \in \mathcal{Y}} q_Y(j) q_{X|Y}(i|j) d(i,j) \leq D \tag{6}$$
Denote the distribution $q_Y(j)qX|Y(i|j)$ on $\mathcal{X} \times \mathcal{Y}$ by $q_{X,Y}(i,j)$. Thus,
$$E_{q_{XY}} d(X,Y) \leq D \tag{7}$$

Let us now bound the probability of event E_3.

First, the probability that over the chunk of length $nq_Y(j)$, the corresponding entries of Z_1^n have type $q_{X|Y=j}$ (recall that p_X is the generating distribution of codeword Z_1^n) is given by:
$$\leq 2^{-nq_Y(j) D(q_{X|Y=j} || p_X)} \tag{8}$$

Thus, the probability that over the whole block of length n, in the chunks $nq_Y(j)$, the corresponding entries of z_1^n have type $q_{X|Y=j}$, for all j
$$\begin{aligned} &\leq \Pi_{j \in \mathcal{Y}} 2^{-nq_Y(j) D(q_{X|Y=j}||p_X)} && (9)\\ &= 2^{-n \sum_{j \in \mathcal{Y}} q_Y(j) D(q_{X|Y=j}||p_X)} && (10)\\ &= 2^{-nD(q_{XY}||p_X q_Y)} && (11)\end{aligned}$$

It would be helpful to note the positions of where p occur and where q occur, in the above expression.

To bound the probability that z_1^n is at a distortion $\leq D$ from y_1^n, we have to sum the above probability over all possible types $q_{X|Y=j}, 1 \leq j \leq |\mathcal{Y}|$ such that conditions 1 and 2 above (equivalently, (5) and (7)) are satisfied.

Number of $q_{X|Y=j}$ types $\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}$. Also recall that number of non-transmitted codewords $|\mathcal{C}_R| \leq 2^{nR}$.

Putting all this together and using the union bound,
$$\begin{aligned} &\Pr(E_3 | \text{ type of } y_1^n \text{ is } q_Y) && (12)\\ &\quad \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{nR} 2^{-n \inf_{q_{XY} \in \mathcal{S}} D(q_{XY}||p_X q_Y)} \end{aligned}$$

where $\mathcal{S}$ denotes the set of types satisfying conditions 1 and 2 (equivalently, (5) and (7)), and is
$$\mathcal{S} = \left\{ q_{XY} : \begin{array}{c} q_X \in p_X \pm \epsilon \\ E_{q_{XY}} d(X,Y) \leq D \\ q_Y \text{ fixed} \end{array} \right\} \tag{13}$$

Now, q_Y, the type of the received sequence y_1^n is arbitrary. Thus, an easy way to bound $\Pr(E_3)$ is to just remove the q_Y fixed condition from the above definition of $\mathcal{S}$.

Thus finally,

$$\Pr(E_3) \leq (n+1)^{|\mathcal{X}|(|\mathcal{Y}|+1)} 2^{nR} 2^{-n \inf_{q_{XY} \in \mathcal{T}} D(q_{XY}||p_X q_Y)} \tag{14}$$

where $\mathcal{T}$ is the set

$$\mathcal{T} = \left\{ q_{XY} : \begin{array}{l} q_X \in p_X \pm \epsilon \\ E_{q_{XY}} d(X,Y) \leq D \end{array} \right\} \tag{15}$$

The only difference between the sets $\mathcal{S}$ and $\mathcal{T}$ is that the q_y fixed condition which exists in $\mathcal{S}$ has been removed in $\mathcal{T}$.

Since $(n+1)^{|\mathcal{X}|(|\mathcal{Y}|+1)}$ is a polynomial, $\Pr(E_3) \to 0$ as $n \to \infty$ if

$$R < \inf_{\substack{q_X \in p_X \pm \epsilon \\ Ed_{q_{XY}}(X,Y) \leq D}} D(q_{XY}||p_X q_Y) \tag{16}$$

Thus to prove Theorem 1, it suffices to prove that

$$\begin{aligned} \Theta_1 &\triangleq \lim_{\epsilon \to 0} \inf_{\substack{q_X \in p_X \pm \epsilon \\ E_{q_{XY}} d(X,Y) \leq D}} D(q_{XY}||p_X q_Y) \\ &= R_X(D) = \inf_{\substack{X \sim p_X \\ Ed(X,Y) \leq D}} I(X;Y) \\ &= \inf_{\substack{p_X \text{ fixed} \\ p_Y \text{ can vary} \\ E_{p_{XY}} d(X,Y) \leq D}} D(p_{XY}||p_X p_Y) \triangleq \Theta_2 \end{aligned} \tag{17}$$

The main difference between Θ_1 and Θ_2 (note the definitions of Θ_1 and Θ_2 in the above equation) is that:

- In Θ_1, we have $D(q_{XY}||p_X q_Y)$; $q_X \in p_X \pm \epsilon$
- In Θ_2, we have $D(p_{XY}||p_X p_Y)$

It is clear that Θ_1 has "more freedom" and hence, $\Theta_1 \leq \Theta_2$.

All we need to prove is that $\Theta_1 \geq \Theta_2$.

This we do with a simple trick:

$$\begin{aligned} D(q_{XY}||p_X q_Y) &= D(q_X||p_X) + D(q_{XY}||q_X q_Y) \\ &\geq D(q_{XY}||q_X q_Y) \end{aligned} \tag{18}$$

Thus,

$$\Theta_1 \geq \lim_{\epsilon \to 0} \inf_{\substack{q_X \in p_X \pm \epsilon \\ E_{q_{XY}} d(X,Y) \leq D}} D(q_{XY}||q_X q_Y) \tag{19}$$

So we only need to prove that

$$\begin{aligned} &\lim_{\epsilon \to 0} \inf_{\substack{q_X \in p_X \pm \epsilon \\ E_{q_{XY}} d(X,Y) \leq D}} D(q_{XY}||q_X q_Y) \\ &\geq \lim_{\epsilon \to 0} \inf_{\substack{p_X \text{ fixed} \\ Ed(X,Y) \leq D}} D(p_{XY}||p_X p_Y) \end{aligned} \tag{20}$$

This holds with equality, and follows from the continuity of the rate distortion function $R_X(D)$ in p_X and proves the direct theorem.

The sequence of choosing n, ϵ depending on the rate $R < R_X(D)$ and probability of error p_e is:

1) Choose ϵ small enough so that $R < \inf_{X \in p_X \pm \epsilon} R_X(D)$.
2) Choose n large enough so that the total probability of error from the events E_1, E_2 and E_3 adds up to a value less than p_e.

We now sketch the proof of the converse theorem, Theorem 2, that is, in general, we cannot transmit at rates larger than $R_X(D)$ over a D-distortion attacker. Another way of stating this is that if one tries to transmit at rates larger than $R_X(D)$, there is a D-distortion attacker such that we cannot transmit reliably over this attacker.

First, consider the case that we are restricted to using iid p_X random codes; we will remove this restriction later.

Let the rate at which we want to transmit, $R = R_X(D - \alpha) > R_X(D)$ for some $\alpha > 0$.

We will show that there is a D-distortion attacker which is a DMC for which error probability $\not\to 0$.

Look at all DMCs that produce an average distortion of $(D - \frac{\alpha}{2})$ between the input and output when input is p_X distributed.

$$C_{\text{worst}} = \inf_{\substack{X \sim p_X \\ Ed(X,Y) \leq (D - \frac{\alpha}{2})}} I(X;Y) \tag{21}$$

But this value is precisely $R_X(D - \frac{\alpha}{2})$. Also, any DMC that produces an average distortion of $(D - \frac{\alpha}{2})$ is a D-distortion attacker (follows from the weak law of large numbers). Thus, we have exhibited a DMC which is a D-distortion attacker and over which, we cannot reliably at rates larger than $R_X(D - \frac{\alpha}{2}) < R_X(D - \alpha) = R$.

To remove the assumption that we have to use p_X random codes, consider the following attacker:

Fix $\epsilon > 0$. The attacker looks at inputs of length n and if the input is not p_X typical (that is, the empirical type does not lie in $p_X \pm \epsilon$), the attacker will produce junk output, say the all 1 sequence, whereas if the input sequence is p_X-typical, the attacker will act like the above DMC. The attacker needs to keep increasing the length of sequences which it looks at and attacks, and correspondingly decrease ϵ. It is intuitively clear that if a codebook is chosen with a codeword which is not p_X-typical, the output of the attacker will give no positive rate information about what was transmitted, and hence, the encoder can not use such codewords to transmit reliably at rates larger than $R_X(D)$.

VI. Comments on the proof

If one compares the proofs of Shannon's channel coding theorem and the above, the two are quite similar in the error calculation for the event E_3, but there is one difference. In Shannon's theorem, proving that the average error probability over the ensemble of codes $\to 0$ implies that there exists a codebook for which the error probability $\to 0$ for every single message. This is not immediately true in our case because the attacker can use different strategies over different blocks.

Furthermore, if we were to use the same codebook over and over again, the input would no longer look iid p_x on very long sequences and the attacker would be free to just drive us to zero. Thus, the codebook has to be generated at least somewhat independently in each block of length n. This is where we use the assumption that there is common randomness available — using this common randomness, the transmitter and the receiver can generate the codebook again and again, independently.

However, the code as given requires an exponentially large amount of common randomness. This can easily be reduced to a polynomial (in the block-length n) amount of common randomness by using the following tricks:(details in [3])

- Simulate in advance whether the input block will be ϵ-typical or not. (Can use $O(\log n)$ bits) If it is atypical, just declare error no matter what message was sent.
- Make slight modifications to the proof to instead show the existence of deterministic codebooks with input types like $p_x \pm \epsilon$ that can be list-decoded to some possibly large, but constant, list-size l when facing a worst-case attacker inducing a distance D. This is done by patching the above proof with arguments analogous to those for Theorem 5.1 in [12]. The additional trick is just noticing that $I(X;Y) = H(Y) - H(Y|X)$ and that $2^{nH(Y)}$ is essentially the total number of output sequences[6] of type q_Y. When l is large enough, $\frac{l}{l+1}H(Y) - H(Y|X)$ is as close as desired[7] to $R_X(D)$.
- Once the deterministic codes are constant composition, a random permutation of the indices will make each of them behave as though they were drawn from the original iid p_x distribution conditioned on the empirical type being typical. This takes $O(n\log(n))$ commonly-random bits.
- By using the code at a rate slightly less than the rate of the code, the message can be padded with a randomly chosen hash of the true message. This takes at most another $O(n)$ commonly-random bits and allows the decoder to uniquely disambiguate the decoded lists with high probability by just rejecting messages whose hashes do not match up correctly.

VII. Theorem - Conditional Case

Until now, we assumed that the input to the attacker should be a p_X-iid sequence. Now, consider the case that the input is still an independently generated sequence but the distribution of X_i depends on an iid random variable sequence V_1^∞ that is revealed non-causally to all parties.

We state some notation to add to the notation previously.

- $\mathcal{V} \rightarrow= \{1, 2, \ldots |\mathcal{V}|\}$ is a finite set. A generic element of $\mathcal{V}$ will be denoted by s.
- $p_V \rightarrow$ probability distribution on $\mathcal{V}$.
- $V_1^\infty \rightarrow$ iid sequence of random variables generated p_V. In watermarking terms, this can be thought of as the "cover-story." We will talk about relations to watermarking in Section IX.
- $p_{X|V=s} \rightarrow$ If $V_i = s$, X_i is generated according to the distribution $p_{X|V=s}$, but independently of other X_j. The joint distribution on (V_i, X_i) will be denoted by p_{VX}
- Attacker → **We assume that V_i is known noncausally to the encoder, decoder and the attacker**.

The next theorem is a conditional version of the inverse rate-distortion theorem, Theorem 1.

Theorem 3: Assuming that there is common randomness available at the transmitter and the receiver, all rates

$$R < R_{X|V}(D) \triangleq \inf_{\substack{(V,X)\sim p_{VX}\\ Ed(X,Y)\le D}} I(X;Y|V) \tag{22}$$

are achievable over a D-distortion attack channel, and in fact, this can be done by using iid $p_{X|V}$ random codes.

We omit a converse theorem though the same arguments as above would give one.

VIII. Proofs - Conditional Case

The proof is very similar to the proof of the theorem in the unconditional case. Recall that V_1^∞ is known to the transmitter, receiver, and attacker.

Codebook Construction: Generate 2^{nR} codewords iid $p_{X|V}$. This is the codebook, which we denote by $\mathcal{C}$.

Decoding: Fix $\epsilon > 0$. Restrict attention to those codewords x_1^n such that (v_1^n, x_1^n) is p_{VX} typical, that is, whose type lies in $p_{VX} \pm \epsilon$.

Denote this restricted set of codewords by $\mathcal{C}_R$.

Note that if v_1^n is not typical, C_R will be empty. Thus:

- The definition of $\mathcal{C}_R$ implicitly assumes an error if v_1^n is not strongly typical.
- $\mathcal{C}_R$ depends on v_1^n, that is, the codewords of $\mathcal{C}$ which lie in $\mathcal{C}_R$ are different for different v_1^n.

Let y_1^n denote the output of the attacker. If there is a unique x_1^n in the restricted codebook which is at an average distortion less than or equal to D from the output sequence, declare that x_1^n was transmitted, else declare error. We call this the "ϵ-Nearest Conditionally Typical Neighbor" decoding rule.

In what follows, z_1^n will denote a non-transmitted codeword as before. As in the unconditional case, the error event consists of 3 parts:

- $E_1 \rightarrow (v_1^n, x_1^n)$ is not typical. This is a slight modification of E_1 in the unconditional case.
- $E_2 \rightarrow$ Distortion caused by the attacker is not typical, that is, transmitted codeword is at an average distortion larger than D from the received sequence. Mathematically, $\frac{1}{n}\sum_{t=1}^n d(x_t, y_t) > D$. This is exactly the same as in the unconditional case.
- $E_3 \rightarrow$ a typical codeword which was not transmitted is at an average distortion less than or equal to D from

[6] Rather than computing the probability of error, we are computing the expected number of D-balls that have at least $l+1$ codewords in them. For a given $l+1$ codeword positions, this is just the existing probability of collision raised to the $l+1$ power times the number of possible D-balls. The total number of such combinations is also no more than $2^{nR(l+1)}$.

[7] And so the expected total number of collisions is as small as we want and so there exists at least one deterministic codebook that has no such collisions at the l-list level.

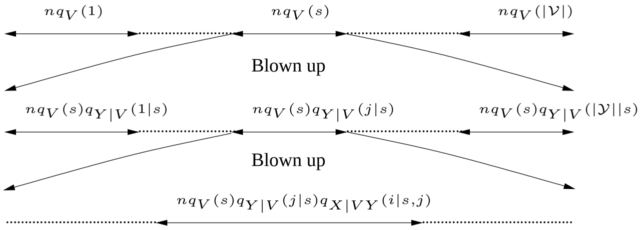

Fig. 2. The various types illustrated in the conditional rate-distortion case.

the received sequence. This is exactly the same as in the unconditional case.

$\Pr(\text{error}) \leq \Pr(E_1)+\Pr(E_2)+\Pr(E_3)$. $\Pr(E_1), \Pr(E_2) \to 0$ as in the unconditional case.

All we need to do is to upper bound $\Pr(E_3)$. As before, we do a method-of-types calculation on the probability of possible z_1^n that will cause an error for a given received sequence y_1^n.

The only essential difference between this proof and in the proof of the unconditional case is that we first do a sorting based on V and then proceed exactly the same as before, that is, do a sorting based on Y and then do a sorting based on X.

Let the type of v_1^n look like q_V. Sort, so that all t such that $V_t = s$ are together. Over the subsequence where $V_t = s$, let the type of the output produced by the attacker be $q_{Y|V=s}$. Again, do a sub-sorting such that all $Y_t = j$ are together in each subsequence of $V_t = s$. In this $(V_t = s, Y_t = j)$ subsequence, let the type of the subsequence of z_1^n (recall - z_1^n is a codeword which is NOT transmitted) look like $q_{X|Y=j,V=s}$. See Figure 2.

We now do the $\Pr(E_3)$ calculation.

First restrict attention to the subsequence $V_t = s$. Over this subsequence, we do exactly what we did in the unconditional case. It follows from the proof of the unconditional case that the probability that Z_1^n looks like $q_{X|V=s,Y=j}$ given that the y_1^n subsequence type looks like $q_{Y|V=s}$ is

$$\leq 2^{-nq_V(s)D(q_{XY|V=s}||p_{X|V=s}q_{Y|V=s})} \tag{23}$$

The probability that over the whole sequence, the Z_1^n type is $q_{X|V,Y}$ given that the Y type is $q_{Y|V}$

$$\begin{aligned} &\leq \quad 2^{-n\sum_{s\in\mathcal{V}} q_V(s)D(q_{XY|V=s}||p_{X|V=s}q_{Y|V=s})} &(24)\\ &= \quad D(q_{XY|V}||p_{X|V}q_{Y|V}|q_V) &(25)\end{aligned}$$

There are a polynomial number of q_{VXY} types, $\leq (n+1)^{|\mathcal{V}||\mathcal{X}||\mathcal{Y}|}$ and by argument similar to that in the unconditional case,

$$\begin{aligned} &Pr(E_3) \leq \\ &2^{nR}(n+1)^{|\mathcal{V}||\mathcal{X}|(|\mathcal{Y}|+1)}2^{-n\inf_{q_{VXY}\in\mathcal{R}} D(q_{XY|V}||p_{X|V}q_{Y|V}|q_V)} \end{aligned} \tag{26}$$

where the set $\mathcal{R}$ over which the above infimum is taken is:

1) (v_1^n, z_1^n) is typical, that is, $q_{VX} \in p_{VX} \pm \epsilon$.
2) z_1^n is at an average distortion $\leq D$ from the received sequence y_1^n, that is, $E_{q_{VXY}} d(X,Y) \leq D$

Thus,

$$\mathcal{R} = \left\{ q_{VXY} : \begin{array}{l} q_{VX} \in p_{VX} \pm \epsilon \\ E_{q_{VXY}} d(X,Y) \leq D \end{array} \right\} \tag{27}$$

It follows that we only need to prove that

$$\begin{aligned} \lim_{\epsilon\to 0} \inf_{\substack{q_{VX} \in p_{VX} \pm \epsilon \\ E_{q_{VXY}} d(X,Y) \leq D}} & D(q_{XY|V}||p_{X|V}q_{Y|V}|q_V) &(28)\\ = R_{X|V}(D) = \inf_{\substack{(V,X) \sim p_{VX} \\ Ed(X,Y) \leq D}} & I(X;Y|V) \\ = \inf_{\substack{p_{VX} \text{ fixed} \\ Ed_{p_{VXY}}(X,Y) \leq D}} & D(p_{XY|V}||p_{X|V}p_{Y|V}|p_V) \end{aligned}$$

The proof of this follows in almost the same way as in the unconditional case, just that we have to use the continuity of $R_{X|V}(D)$ in p_{VX} (in the unconditional case, we had used the continuity of $R_X(D)$ in p_X).

This proves the conditional theorem, Theorem 3.

IX. Relation to Watermarking

We can view this conditional problem as a watermarking problem with a coverstory[8]. In watermarking, the user is allowed to make some tolerable level of distortion to the covertext. We have a restriction of another kind, that is, if the coverstory entry is s, the input distribution should be $p_{X|V=s}$. Also, in watermarking, the covertext is not known to the attacker.[9] We assume that the covertext is known to the attacker. If one looks at (38) in the paper of Somekh-Baruch and Merhav [10], this is the reason for the Markov Chain condition $U \to X \to Y$. We do not have the Markov Chain condition $V \to X \to Y$ because the covertext V is known to the attacker.

X. Continuous alphabets

In this section, we consider the case when $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{V}$ are not necessarily finite discrete alphabets. We divide the problem into 6 cases:

1) $\mathcal{X}$ finite, $\mathcal{Y}$ finite, $\mathcal{V}$ not there.
2) $\mathcal{X}$ finite $\mathcal{Y}$ finite, $\mathcal{V}$ finite.
3) $\mathcal{X}$ non-finite, $\mathcal{Y}$ non-finite, $\mathcal{V}$ not there.
4) $\mathcal{X}$ non-finite, $\mathcal{Y}$ non-finite, $\mathcal{V}$ finite.
5) $\mathcal{X}$ finite, $\mathcal{Y}$ finite, $\mathcal{V}$ non-finite.
6) $\mathcal{X}$ non-finite, $\mathcal{Y}$ non-finite, $\mathcal{V}$ non-finite.

We will refer to these as Cases 1 through 6. Case 1 is the unconditional case covered in Theorem 1, Case 2 is the conditional case covered in Theorem 3. We now go on to the rest. The proofs will be based on quantization of the above sets and using ideas from the proofs of Theorem 1 and 3.

For Case 3, we need to prove that rates $< R_X(D)$ are achievable and for Cases 4,5,6, we need to prove that rates $< R_{X|V}(D)$ are achievable.

[8] To distinguish it from the "covertext" in traditional watermarking

[9] Since otherwise, presumably the attacker could just replace the input with the covertext itself. The same is not true if it is considered as a coverstory.

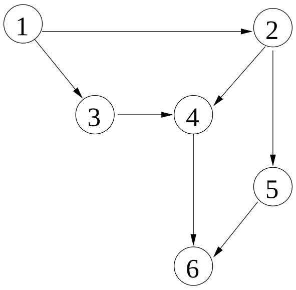

Fig. 3. Dependence graph for the proofs of the various cases

Figure 3 is a dependency graph of which proofs depend on which.

A. Compact support

We first tackle Case 3, that is, $\mathcal{X}, \mathcal{Y}$ are non-finite sets, and there is no $\mathcal{V}$. We first assume that $\mathcal{X}$ and $\mathcal{Y}$ are bounded subsets of $\mathcal{R}^\gamma$, for some positive integer γ. The case of unbounded support is addressed later.

We first state some notation:

- $\mathcal{X}, \mathcal{Y} \rightarrow$ **bounded** subsets of $\mathcal{R}^\gamma$.
- $a \rightarrow$ generic point in $\mathcal{X}$. We do not use x because of potential confusion with the transmitted sequence.
- $b \rightarrow$ generic point in $\mathcal{Y}$. We do not use y because of potential confusion with the received sequence.
- $d : \mathcal{R}^\gamma \times \mathcal{R}^\gamma \rightarrow \mathcal{R}$ is a difference distortion measure which is assumed to be **uniformly continuous** with respect to the Euclidean metric.
- D-distortion attacker $\rightarrow$ Same as before. If the input to the attacker is the random variable sequence X_1^∞ (X_i iid p_X) , the attacker produces the random variable sequence Y_1^∞. This results in a joint probability measure on (X_1^∞, Y_1^∞). Under this probability measure,

$$\sup_t \Pr\left(\frac{1}{n}\sum_{u=t}^{t+n-1} d(X_u, Y_u) > D\right) \rightarrow 0 \text{ as } n \ \rightarrow \infty \tag{29}$$

- $\mathcal{X}_\Delta, \mathcal{Y}_\Delta \rightarrow \Delta$-hypercube grid quantization of $\mathcal{X}, \mathcal{Y}$ respectively. The boundary of the hypercube can be put in any of the adjoining sets but not both. The quantization point is taken as the center of the hypercube.
- $a_\Delta \rightarrow$ Generic point of $\mathcal{X}_\Delta$. $a_\Delta \in \mathcal{X}_\Delta$ is obtained by quantizing $a \in \mathcal{X}$.
- $b_\Delta \rightarrow$ Generic point of $\mathcal{Y}_\Delta$. $b_\Delta \in \mathcal{Y}_\Delta$ is obtained by quantizing $b \in \mathcal{Y}$.
- $p_{X_\Delta} \rightarrow$ Probability distribution on $\mathcal{X}_\Delta$ obtained from the distribution p_X on $\mathcal{X}$ in the obvious way.

Note that since the difference distortion function is uniformly continuous and $\mathcal{X}, \mathcal{Y}$ are bounded, $\frac{1}{n}\sum_{t=1}^n d(x_{t\Delta}, y_{t\Delta}) \leq \frac{1}{n}\sum_{t=1}^n d(x_t, y_t) + g(\Delta) \forall (x_1^n, y_1^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ where $g(\Delta) \rightarrow 0$ as $\Delta \rightarrow 0$.

It follows that under the distribution governing $(X_{1\Delta}^n, Y_{1\Delta}^n)$ under the D-distortion attacker,

$$\sup_t \Pr\left(\frac{1}{n}\sum_{u=t}^{t+n-1} d(X_{u\Delta}, Y_{u\Delta}) > D + g(\Delta)\right) \rightarrow 0 \text{ as } n \rightarrow \infty \tag{30}$$

If we work in the quantized world, this suggests what the decoding rule should be.

Codebook Construction: Generate 2^{nR} codewords iid p_X. This is the codebook $\mathcal{C}$. Let $\mathcal{C}_\Delta$ denote the quantized codebook obtained by quantizing each codeword.

Decoding: Fix $\epsilon > 0$.

Restrict attention to those quantized codewords which are p_{X_Δ}-typical. Denote this restricted set of quantized codewords by $\mathcal{C}_R^\Delta$.

Let $y_{1\Delta}^n$ denote the quantized output of attacker. If there is a unique p_{X_Δ}-typical quantized codeword $x_{1\Delta}^n$ which is at an average distortion less than or equal to $D + g(\Delta)$ (**note the change** $D + g(\Delta)$ **instead of** D) from the output sequence, declare that x_1^n was transmitted, else declare error. Mathematically, if $\exists ! x_1^n \in \mathcal{C}_R$ such that $\frac{1}{n}\sum_{t=1}^n d(x_t, y_t) \leq D + g(\Delta)$, declare that x_1^n was transmitted, else declare error.

This decoding rule has reduced the problem to Case 1(finite $\mathcal{X}$ and $\mathcal{Y}$), and we can use results from there. Thus, we can transmit at rates $R < R_{X_\Delta}(D + g(\Delta))$ using this decoding rule. It can be shown using the appropriate continuity arguments that $\lim_{\Delta \rightarrow 0} R_{X_\Delta}(D + g(\Delta)) = R_X(D)$. This proves that we can transmit at all rates $< R_X(D)$.

The sequence in which n, Δ, ϵ need to be chosen depending on the desired rate $R < R_X(D)$ and the error probability p_e is:

1) Choose Δ small enough so that $R < R_{X_\Delta}(D + g(\Delta))$.
2) Choose ϵ small enough so that $R < \inf_{X_\Delta \in p_{X_\Delta} \pm \epsilon} R_{X_\Delta}(D + g(\Delta))$.
3) Choose n large enough so that the sum of error probabilities of events E_1, E_2, E_3 is less than p_e.

Case 4, where $\mathcal{X}, \mathcal{Y}$ are non-finite while the "coverstory" $\mathcal{V}$ is finite, is proved in exactly the same way — by quantizing X, Y finely enough.

Next we consider Case 5, that is, $\mathcal{X}, \mathcal{Y}$ are finite and $\mathcal{V}$ is non-finite. We assume that $\mathcal{V}$ is a bounded subset of $\mathcal{R}^\eta$ for some positive integer η.

We introduce some notation regarding $\mathcal{V}$.

- $\mathcal{V} \rightarrow$ bounded subset of $\mathcal{R}^\eta$.
- $s \rightarrow$ generic element of $\mathcal{V}$.
- $\mathcal{V}_{\Delta'} \rightarrow \Delta'$ hypercube quantization of $\mathcal{V}$. The boundary of the hypercube can be put in any of the adjoining sets. Quantization is taken as the center of the hypercube. We use Δ' instead of Δ because we use Δ for quantizing $\mathcal{X}$ and $\mathcal{Y}$.
- $s_{\Delta'} \rightarrow$ Generic point of $\mathcal{V}_{\Delta'}$. $s_{\Delta'}$ in $\mathcal{V}_{\Delta'}$ is got by quantizing s in $\mathcal{V}$.
- $\mathcal{S}_{\Delta'} \rightarrow$ quantization region (hypercube) of $\mathcal{V}$ containing the point $s_{\Delta'} \in \mathcal{V}_{\Delta'}$.
- $p_{V_{\Delta'}} \rightarrow$ probability distribution on $\mathcal{V}_{\Delta'}$ got from p_V on $\mathcal{V}$ in the obvious way.

What is not obvious, though, is how to define $p_{X|V_{\Delta'}}$. We need to make definitions in such a way that we can do probability of error calculations for the event E_3 (the other two events, E_1 and E_2 will be trivial as usual).

$$p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}} = \sup_{s \in \mathcal{S}_{\Delta'}} p_{X|V}(i|s), \ i \in \mathcal{X} \tag{31}$$

$$p^{\text{inf}}_{X|V_{\Delta'}=s_{\Delta'}} = \inf_{s \in \mathcal{S}_{\Delta'}} p_{X|V}(i|s), \ i \in \mathcal{X} \tag{32}$$

$p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}$ is not, in general, a probability measure. It is a measure with mass ≥ 1 and denotes a measure which "dominates" all probability measures $p_{X|V=s}$ over the quantization region of $\mathcal{V}$ which contains $s_{\Delta'}$.

$p^{\text{inf}}_{X|V_{\Delta'}=s_{\Delta'}}$ is not, in general, a probability measure. It is a measure with mass ≤ 1. It denotes a measure which "is dominated by" all probability measures $p_{X|V=s}$ over the quantization region of $\mathcal{V}$ which contains $s_{\Delta'}$.

Intuitively, if we make some continuity assumptions on $p_{X|V=s}$ as $s \in \mathcal{V}$ varies, then $p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}$ and $p^{\text{inf}}_{X|V_{\Delta'}=s_{\Delta'}}$ will be close to each other. For small enough Δ', all $s \in \mathcal{S}_{\Delta'}$ are almost the same in the distribution induced on $\mathcal{X}$.

Another reason for defining $p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}$ is that it helps us to do error probability calculations. This is demonstrated by the following lemma:

Lemma 1: Let p_X be a probability distribution on $\mathcal{X}$. Let μ_X be a measure on $\mathcal{X}$ such that $\mu_X(i) > p_X(i)$ for all $i \in \mathcal{X}$ (that is, μ_X dominates p_X). Let q_X be another probability distribution on $\mathcal{X}$.

Then, probability that an n length sequence generated iid p_X has type q_X

$$p^n_X(T(q_X)) \leq 2^{-nD(q_X||\mu_X)} \tag{33}$$

where $D(q_X||\mu_X)$ is defined in the obvious way, $D(q_X||\mu_X) \triangleq \sum_{i \in \mathcal{X}} q_X(i) log \frac{q_X(i)}{\mu_X(i)}$
Proof:

$$\begin{aligned} & p^n_X(T(q_X)) \\ \leq \ & 2^{-nD(q_X||p_X)} \text{ (by method of types)} \\ \leq \ & 2^{-nD(q_X||\mu_X)} \text{ (trivial by definition of } D(q_X||\mu_X)) \end{aligned}$$

This lemma gives us a way of upper bounding the error probability of a type class when we do not know the generating distribution, but have an upper bound on the same, and this is precisely the situation we are in.

We define $p^{\text{avg}}_{X|V_{\Delta'}=s_{\Delta'}}$ as the probability measure obtained by normalizing $p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}$.

If we have some continuity conditions (which we will make rigorous later) on $p_{X|V}$, as measures, $p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}$, $p^{\text{inf}}_{X|V_{\Delta'}=s_{\Delta'}}$, $p^{\text{avg}}_{X|V_{\Delta'}=s_{\Delta'}}$, $\{p_{X|V=s}, s \in \mathcal{S}_{\Delta'}\}$ will be quite close to each other.

Also, the distributions $p_{V_{\Delta'}}$ and $p^{\text{avg}}_{X|V_{\Delta'}=s_{\Delta'}}$ result in a probability distribution on $(V_{\Delta'}, X)$ which we denote by $p^{\text{avg}}_{V_{\Delta'},X}$.

Next, we state the codebook formation and decoding rule:

Codebook Construction: Generate 2^{nR} codewords iid $p_{X|V}$. This is the codebook $\mathcal{C}$.

Decoding: Fix $\epsilon > 0$. Restrict attention to those codewords x^n_1 that $(x^n_1, v^n_{1\Delta'})$ have an empirical type $q_{X,V_{\Delta'}}$ that is $p^{\text{avg}}_{V_{\Delta'},X}$ typical. Denote this restricted set of codewords by $\mathcal{C}_R$.

Let y^n_1 denote the output of the attacker. If there is a unique x^n_1 in the restricted codebook $\mathcal{C}_R$ that is at an average distortion less than or equal to D from the output sequence, declare that x^n_1 was transmitted, else declare error.

We impose the following technical condition[10] on $p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}$ and $p^{\text{inf}}_{X|V_{\Delta'}=s_{\Delta'}}$, which captures mathematically, the closeness of $p_{X|V=s_1}$ and $p_{X|V=s_2}$ for s_1 and s_2 close.

Technical Condition: $\forall i \in \mathcal{X}$

$$\lim_{\Delta' \to 0} \max_{s_{\Delta'} \in \mathcal{V}_{\Delta'}} \left| p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}(i|s_{\Delta'}) - p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}(i|s_{\Delta'}) \right| = 0 \tag{34}$$

This condition says that $p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}(i|s_{\Delta'}) - p^{\text{sup}}_{X|V_{\Delta'}=s_{\Delta'}}(i|s_{\Delta'}) \to 0$ as$\Delta' \to 0$ **uniformly** over all partitions of $\mathcal{V}$.

We now do the probability of error calculations.

It is easy to check that with the above decoding rule, the probabilities of error event $E_2 \to 0$ as $n \to \infty$. For E_1, all that is required is for n to be large enough while ϵ is also large enough relative to Δ' so that $[p^{\text{inf}}_{V_{\Delta'},X} - \frac{\epsilon}{2}, p^{\text{sup}}_{V_{\Delta'},X} + \frac{\epsilon}{2}] \in p^{\text{avg}}_{V_{\Delta'},X} \pm \epsilon$. At that point, the weak law of large numbers is enough to guarantee what is desired.

For $\Pr(E_3)$, we follow the steps in the proof of Case 2 ($\mathcal{X}$ finite $\mathcal{Y}$ finite, $\mathcal{V}$ finite) (Theorem 3) and use Lemma 1 to replace $p_{X|V}$ with $p^{\text{sup}}_{X|V_{\Delta'}}$. It follows that we can transmit at rates

$$R < \lim_{\epsilon \to 0} \inf_{\substack{q_{V_{\Delta'},X} \in p^{\text{avg}}_{V_{\Delta'},X} \pm \epsilon \\ Ed_{q_{V_{\Delta'},XY}}(X,Y) \leq D}} D\left(q_{XY|V_{\Delta'}}||p^{\text{sup}}_{X|V_{\Delta'}}q_{Y|V_{\Delta'}}|q_{V_{\Delta'}}\right) \tag{35}$$

First thing that we need to take care of $p^{\text{sup}}_{X|V_{\Delta'}}$ appearing above - we want to somehow replace it by $p^{\text{avg}}_{X|V_{\Delta'}}$. Using the technical condition (34), it is easy to see that there is a function h such that we can transmit $R <$

$$\lim_{\epsilon \to 0} \inf_{\substack{q_{V_{\Delta'},X} \in p^{\text{avg}}_{V_{\Delta'},X} \pm \epsilon \\ Ed_{q_{V_{\Delta'},XY}}(X,Y) \leq D}} D\left(q_{XY|V_{\Delta'}}||p^{\text{avg}}_{X|V_{\Delta'}}q_{Y|V_{\Delta'}}|q_{V_{\Delta'}}\right) - h(\Delta') \tag{36}$$

where $h(\Delta') \to 0$ as $\Delta' \to 0$. The first term above is the same as that appearing in the proof of Case 2, the conditional case with $\mathcal{X}, \mathcal{Y}, \mathcal{V}$ finite, Equation 28. It follows that we can transmit at all rates

$$R < \inf_{\substack{(V_{\Delta'},X) \sim p^{\text{avg}}_{V_{\Delta'},X} \\ Ed(X,Y) \leq D}} I(X;Y|V_{\Delta'}) - h(\Delta') = R_{X|V_{\Delta'}}(D) - h(\Delta') \tag{37}$$

[10] It can be shown to be satisfied for any joint distribution for X, V that satisfies weak convergence in that $p(X|V = s_n) \to p(X|V = s)$ whenever $s_n \to s$.

Now, $\lim_{\Delta' \to 0} R_{X|V_{\Delta'}}(D) - h(\Delta') = R_{X|V}(D)$ (we need to use the technical condition (34) for proving this), and it follows that we can transmit at all rates less than $R_{X|V}(D)$.

The sequence in which we choose n, ϵ, Δ' depending on the rate R and the probability of error p_e is

1) Choose Δ' small enough so that $R < R_{X|V_{\Delta'}}(D) - h(\Delta')$
2) Choose ϵ small enough such that $R < \inf_{(V_{\Delta'}, X) \in p^{\text{avg}}_{V_{\Delta'}, X} \pm \epsilon} R_{X|V_{\Delta'}}(D) - h(\Delta')$
3) Choose n large enough so that sum of error probabilities of events $E_1, E_2, E_3 < p_e$.

Finally, we consider Case 6, that of $\mathcal{X}, \mathcal{Y}, \mathcal{V}$ non-finite. This is just a mixture of decoding rules for Case 4 ($\mathcal{X}$ non-finite, $\mathcal{Y}$ non-finite, $\mathcal{V}$ finite) and the previous case, Case 5 ($\mathcal{X}$ finite, $\mathcal{Y}$ finite , $\mathcal{V}$ non-finite).

First quantize $\mathcal{X}, \mathcal{Y}$ to size Δ. This way, we get $p_{X_\Delta|V}$. This reduces the problem to previous case where $\mathcal{X}$ and $\mathcal{Y}$ are finite and by combining the decoding rules of Case 4 and Case 5, it is easy to see that we can transmit at all rates $R < R_{X_\Delta|V}(D + g(\Delta))$ where $g(\Delta)$ is defined analogous to that in Case 3.

Taking $\Delta \to 0$, it follows that we can transmit at all rates $R < R_{X|V}(D)$.

Clearly, the technical condition in place of (34) in this case of $\mathcal{X}, \mathcal{Y}$ non-finite, but bounded support, is: $\forall x_\Delta \in \mathcal{X}$.

$$\lim_{\Delta' \to 0} \max_{s_{\Delta'} \in \mathcal{V}_{\Delta'}} \left| p^{\sup}_{X_\Delta|V_{\Delta'}=s_{\Delta'}}(x_\Delta|s_{\Delta'}) - p^{\sup}_{X_\Delta|V_{\Delta'}=s_{\Delta'}}(x_\Delta|s_{\Delta'}) \right| = 0 \tag{38}$$

This is just saying that the technical condition of the finite $\mathcal{X}$ case should hold for all partitions of $\mathcal{X}$ in this non-finite case.

The sequence in which we choose $\epsilon, n, \Delta, \Delta'$ to achieve a rate R and probability of error $< p_e$ is

1) Choose Δ small enough so that $R < R_{X_\Delta|V}(D+g(\Delta))$.
2) Choose Δ' small enough so that $R < R_{X_\Delta|V_{\Delta'}}(D + g(\Delta)) - h(\Delta')$.
3) Choose ϵ small enough so that $R < \inf_{(V_{\Delta'}, X_\Delta) \in p^{\text{avg}}_{V_{\Delta'}, X_\Delta} \pm \epsilon} R_{X_\Delta|V_{\Delta'}}(D + g(\Delta)) - h(\Delta')$
4) Choose n large enough so that sum of error probabilities caused by events E_1, E_2, E_3 add up to less than p_e.

Next, we state (without proof) sufficient conditions for the technical conditions, Equations (34) and (38) to hold.

1) Case 5, that is, $\mathcal{X}, \mathcal{Y}$ finite, $\mathcal{V}$ non-finite: The following weak convergence condition is sufficient for the technical condition (34) to hold:
$$s_\alpha \to s \implies p_{X|V=s_\alpha} \xrightarrow{w} p_{X|V=s} \tag{39}$$
2) Case 6, that is, $\mathcal{X}, \mathcal{Y}, \mathcal{V}$ non-finite: what we want is that after discretizing $\mathcal{X}$ and $\mathcal{Y}$, the same technical condition should hold. Assuming that $p_{X|V=s}$ **have densities**, the above condition,
$$s_\alpha \to s \implies p_{X|V=s_\alpha} \xrightarrow{w} p_{X|V=s} \tag{40}$$
is sufficient for the technical condition (38) to hold.

B. Unbounded support

The compact support condition is what allowed us to use quantization to reduce everything to the finite-alphabet case where the method of types could work since the number of possible types grew only polynomially in the block-length n. Dealing with this requires an appropriate truncation argument. For space reasons, we merely sketch the essential ideas here:

1) Pick a small $\delta > 0$.
2) Pick a sufficiently large compact region $\mathcal{X}_c \times \mathcal{V}_c$ (with the obvious modifications if there is no coverstory) so that it satisfies the following properties:
   - $P(\mathcal{X}_c \times \mathcal{V}_c) \geq 1 - \delta$
   - $P(\mathcal{X}_c | V = s) \geq 1 - \delta$ for all $s \in \mathcal{V}_c$
   - Let X_c, V_c be the random variables X, V conditioned on their values lying within the compact region $\mathcal{X}_c \times \mathcal{V}_c$. Then $R_{X_c|V_c}(D) \geq (1-\delta) R_{X|V}(D)$.

   Given this, the distribution for $P(X|V = s)$ can be written as a convex combination $(1 - \delta) P_{X_c|V_c=s} + \delta P'_{X|V_c=s}$ for some other distribution $P'_{X|V_c=s}$.
3) Employ a two-part strategy for generating the random codebook. First, we classify positions in the codebook as “clean” or “dirty” or “bad”:
   - Mark as “dirty” all positions where V_t is not in $\mathcal{V}_c$.
   - Flip a commonly random iid biased coin with δ probability of coming up heads for each position. Mark as “bad” all positions where the coin turns up heads.
   - All remaining positions are “clean.”

   Next, we generate the 2^{nR} random codewords iid using $P_{X_c|V_c}$ in the clean positions. For dirty positions, we draw from $P_{X|V}$ while bad positions are drawn from $P'_{X|V_c}$. The resulting codewords look as though they are drawn from $P_{X|V}$.
4) For decoding, look at only the clean positions. If their number is less than $(1 - 4\delta)n$, declare error. Beyond that, we treat it as in the previous cases dealing with compact support, using the appropriate quantization and nearest typical neighbor decoding.

In terms of the probability of error, there is now a new error event E_0 which corresponds to there being more than $4\delta n$ bad or dirty positions. By the weak law of large numbers (since bad and dirty positions arrive no faster than a Bernoulli processes with expected rate 2δ), this cannot happen very often and so $P(E_0) \to 0$ as $n \to \infty$.

The other terms in the probability of error can be bounded by pretending that the attacker knows not only the dirty positions, but also the bad ones. Assume it also knows that our decoding rule is going to ignore all the dirty and bad positions. With this knowledge, the worst thing it can do is choose to allocate no distortion to those positions and spend that distortion over the clean positions. However, this only increases average distortion by a factor $\frac{1+4\delta}{1-4\delta}$ over the clean positions that figure in the decoding process. By choosing δ sufficiently small, we can be sure that $R < R(D(\frac{1+4\delta}{1-4\delta}))$. Everything else proceeds as before.

XI. Stationary-Ergodic Sources

So far, the information-embedding arguments seemed to depend strongly on the assumption of memorylessness. This is what allowed the method-of-types to be used. To deal with more general sources with memory, we can just apply a trick similar to the truncation argument in Section X-B. Once again, in the interest of space, we simply sketch the key ideas in the context of finite-alphabet rate-distortion problems.

Suppose that the source process $\{X_t\}$ is stationary[11] and ergodic. In such cases, the rate-distortion and conditional rate-distortion functions are defined in terms of limits of longer and longer finite-horizon problems X_1^t. So, for any t sufficiently long, then $R < R_X(D)$ implies also that $tR < R_{X_1^t}(tD)$. But before we simply pick a t long enough, we need to impose a technical condition that requires the process to "mix" appropriately uniformly fast towards its stationary distribution.

Assume that for every $\lambda > 0, \beta > 0$, there exists a uniform delay τ so that for all $t > 0$, all possible values[12] x_1^t, all $k > 0$, and all measurable subsets A of $\mathcal{X}^k$:

$$P(X_{t+d}^{t+d+k-1} \in A | X_1^t = x_1^t) = (1-\lambda) P_{\text{stat}}^{\beta}(X_1^k \in A) + \lambda P'(A) \tag{41}$$

where P' is a probability measure that can depend explicitly on t, d, x_1^t while P_{stat}^{β} is a measure that does not have any such dependence and is within $\pm\beta$ of the stationary probability distribution for the original process.

Essentially, (41) just captures the idea that the process has fading memory and that if we wait long enough, the process will return to its stationary distribution regardless of what values the process might have taken in the past. It is easy to verify that (41) holds for all finite-state stationary ergodic Markov chains[13] as well as hidden Markov models with an underlying finite-state stationary ergodic Markov chain.

With this condition, the codebook construction proceeds in the following sequence:

1) Pick small enough λ, β
2) Based on the technical condition, calculate the required delay d to make the process "forget" its past.
3) Pick a t sufficiently long so that $\frac{t}{t+d}$ is close to 1, and the finite horizon rate-distortion function is close to its infinite-horizon limit.
4) Segment time regularly with t time units of potentially embedded data followed by d time units of dead-time.
5) Use common-randomness to generate Bernoulli(λ) random variables used to mark t-long slots as being bad. This is done for the entire codebook, not on a codeword by codeword basis.
6) For the codewords, independently generate the t-long slots that are not bad by drawing from the stationary distribution for X_1^t. Draw bad slots using P' from (41) and the prefix of the codeword[14] so far.
7) Generate the d-length dead-time slots in between by sampling from the appropriate conditional distribution once the following t-long slot has been chosen.

It is clear that every codeword is thus a simulation of the original process with memory. Conditioned on knowing where the good slots of length t are, the process is iid from both the encoder and decoder's point of view and so reverts to the previous case. The decoder can focus entirely on the good slots viewed as an iid process. Once again, the probability of having fewer than a $(1 - 2\lambda)$ proportion of good slots goes to zero. Decoding error can be bounded by supposing that the attacker knew which slots were good and what time-segments were "dead-time." Thus, the attacker can choose to concentrate all its distortion on the good slots. This increases the average distortion by a factor of at most $\frac{t+d}{t}(\frac{1+2\lambda}{1-2\lambda})$ — which is as close to 1 as we want.

References

[1] C. E. Shannon, Coding theorems for a discrete source with a fidelity criterion, *IRE National Convention Record*, vol. 7, no. 4, pp 142-163.

[2] A. Sahai, S. K. Mitter, The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link. Part I: scalar systems, *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp 3369-3395.

[3] M. Agarwal, A. Sahai, S. K. Mitter, A direct equivalence perspective on the separation theorem, *IEEE Transactions on Information Theory*, in preparation.

[4] D. Blackwell, L. Breiman, A. J. Thomasian, The capacity of a class of channels, *Ann. Math. Statistics*, vol. 30, no. 4, December 1959

[5] D. Blackwell, L. Breiman, A. J. Thomasian, The capacity of certain channel classes under random coding, *Ann. Math. Statistics*, vol. 31, no. 4, September 1960

[6] I. G. Stiglitz, A coding theorem for a class of unknown channels, *IEEE Tran. Info. Th.*, Vol. 13, Issue 2, Apr 1967, pp 217-220

[7] I. Csiszar, P. Narayan, Channel capacity for a given decoding metric, *IEEE Tran. Info. Th.*, Vol. 41, Issue 1, Jan. 1995, pp 35-43

[8] R. W. Hamming, Error detecting and error correcting codes, *Bell System Technical Journal*, Vol. 29, Apr. 1950, pp 147-160

[9] P. Moulin, J. A. O'Sullivan, Information-theoretic analysis of information hiding, *IEEE Tran. Info. Th.*, Vol. 49, Issue 3, March 2003, pp 563-593

[10] A. Somekh-Baruch, N. Merhav, On the error exponent and capacity games of private watermarking systems, *IEEE Tran. Info. Th*, Vol. 49, Issue 3, March 2003, pp 537-562

[11] I. Csiszar, J. Korner, *Information Theory: Coding theorems for discrete memoryless systems*, Akademiai Kiado, 1981

[12] V. Guruswami, *List decoding of error correcting codes*, PhD Thesis, MIT, Aug. 2001

[11] Since time for us starts at 1, assume that it has been initialized into its stationary distribution.

[12] All the arguments here immediately generalize to the conditional rate-distortion case if the technical condition holds uniformly over all possible realizations for the cover-story sequence V_1^∞. Essentially, we want to capture the idea that the cover-story should not be able to force the $\{X_t\}$ process to strongly remember what it did in its distant past. This condition can be relaxed so that it is only required to hold for most realizations of the cover-story process.

[13] Because they must mix exponentially fast based on the second largest eigenvalue of the transition matrix.

[14] If the block code is intended to be used over and over again, then in general the dead-times must be interpolated in a way that takes into account what was transmitted in the distant past. This is not a problem for Markov or Hidden Markov processes.