Common Randomness in Information Theory and Cryptography

Part II: CR capacity

Rudolf Ahlswede
Universität Bielefeld
Fakultät für Mathematik
Postfach 100131
33501 Bielefeld
Germany

Imre Csiszár *
Mathematical Institute of the
Hungarian Academy of Sciences
P.O. Box 127
H-1364 Budapest
Hungary

Abstract

The CR capacity of a two-teminal model is defined as the maximum rate of common randomness that the terminals can generate using resources specified by the given model. We determine CR capacity for several models, including those whose statistics depend on unknown parameters. The CR capacity is shown to be achievable robustly, by common randomness of nearly uniform distribution no matter what the unknown parameters are. Our CR capacity results are relevant for the problem of identification capacity, and also yield a new result on the regular (transmission) capacity of arbitrarily varying channels with feedback.

Key words: common randomness, identification capacity, correlated sources, arbitrarily varying channel, feedback, randomization.

*I. Csiszár was partially supported by the Hungarian National Foundation for Scientific Research, Grant T16386.

1 Introduction

Suppose two terminals, called Terminal $\mathcal{X}$ and Terminal $\mathcal{Y}$, have resources such as access to side information and communication links that allow them to observe and (perhaps cooperatively) generate certain random variables (RV's). The permissible rules for this are specified by the particular model, but it is always assumed that the terminals have unrestricted computational power, thus the RV's that can be generated and observed at a terminal at a given time include, as a minimum, all functions of the RV's previously observed there. Common randomness (CR) of $\mathcal{X}$ and $\mathcal{Y}$ means, intuitively, a RV generated by them and observable to both, perhaps with a small probability of error.

A RV generated by a terminal is not necessarily observable there, e.g., when Terminal $\mathcal{X}$ inputs a RV X into a noisy channel to Terminal $\mathcal{Y}$, he thereby generates an output Y observable only at $\mathcal{Y}$. If Terminal $\mathcal{X}$ suitably encodes the RV X he wants to transmit, enabling $\mathcal{Y}$ to decode, then this X will represent CR. If noiseless feedback from $\mathcal{Y}$ to $\mathcal{X}$ is available then the output Y will always represent CR.

In Part 1 [7] we were interested in CR under an additional secrecy constraint, with the motivation that the generated CR will be used as an encryption key. In this paper we do not require secrecy, and just study the maximum amount of CR afforded by a given model, the amount measured by entropy. The most convenient form of CR is uniform common randomness (UCR), i.e., CR represented by a uniformly (or nearly uniformly) distributed RV. For the type of models we will consider, the maximum attainable amount of CR and UCR will be asymptotically the same.

As a very simple example, suppose that there is a DMC from Terminal $\mathcal{X}$ to Terminal $\mathcal{Y}$, Terminal $\mathcal{X}$ can randomize (= can generate RV's with arbitrary distributions), and can input into the DMC any random sequence X^n he has generated (of given "large" length n). Terminal $\mathcal{Y}$ can observe the output Y^n but the terminals have no other resources. It is intuitively clear that in this case $\mathcal{X}$ has to chose X^n to be uniformly distributed on the $\approx \exp(nC)$ codewords of an optimum code; then $\mathcal{Y}$ can decode, and the achieved $\approx nC$ amount of CR is best possible. If noiseless feedback is available, it is better for $\mathcal{X}$ to send independent repetitions of a RV X that produces maximum output entropy $H(Y)$. As now $\mathcal{X}$ can observe Y^n, in this way CR of amount $nH(Y)$ results, clearly the largest possible. Notice that here, too, the optimum could (almost) be attained by a nearly uniform RV, obtained by applying a compression code to Y^n.

As a combinatorial example, let $\mathcal{G}$ be a bipartite graph with vertex sets $\mathcal{X}$ and $\mathcal{Y}$ (we continue the practice of Part 1 that the symbols of the terminals also denote sets assigned to them). Nature selects an edge $(x, y) \in \mathcal{G}$ at random, Terminal $\mathcal{X}$ observes x, Terminal $\mathcal{Y}$ observes y. The terminals can communicate over a noiseless channel, but at most b binary digits may be transmitted, in any number of rounds. No other resources are available (above the minimum described in the first paragraph), in particular, neither terminal can randomize. Then, clearly, $\log |\mathcal{G}|$ is an upper bound to CR, which can be attained iff the communication complexity $C_\infty(\mathcal{H}, P_\mathcal{V}, P_\mathcal{E})$ does not exceed b (with the notation of [17]); here $\mathcal{H}$ denotes the hypergraph with vertex set $\mathcal{V} = \mathcal{X}$ and edge set $\mathcal{E}$ consisting of the sets $\{x : (x, y) \in \mathcal{G}\}$, $y \in \mathcal{Y}$. It may be an interesting study in communication complexity to determine the maximum amount of CR when $b < C_\infty(\mathcal{H}, P_\mathcal{E}, P_\mathcal{V})$, i.e., the maximum entropy of a function on $\mathcal{G}$ that may be computed at both terminals with communication of at most b bits.

One obvious motivation of our interest in CR is that if the two terminals have access to the outcome of the same random experiment, this knowledge may be used to implement correlated random protocols, perhaps leading to much faster algorithms than deterministic ones or those using independent randomization only. In information theory, in particular for arbitrarily varying channels (AVC's), correlated random codes may much outperform deterministic (or randomized) codes; indeed, they may be necessary to attain positive capacity [8]. E.g. for the additive Gaussian AVC with power constraints, (average error) capacity for deterministic codes equals random code capacity only if the sender's power exceeds the jammer's, otherwise the deterministic code capacity is 0 [11].

An even more striking application of CR appears in the theory of identification capacity [4]. It was shown in [5] that for any kind of channel, if sender and receiver can build up nR bits of UCR, this can be used to construct ID codes for $\approx 2^{2^{nR}}$ messages, with small probability of misidentification and misrejection, provided that the channel capacity is positive. The asymptotic optimality of this construction was also established in [5], for DMC's with no feedback and with complete feedback. Similar results for multiuser channels were obtained in [6].

One feature of this paper is that we also study "robust common randomness". This concept refers to models whose statistical properties are not completely specified but depend on certain parameters ("state") out of control of the two terminals and at least partially unknown to them. Then the distribution of the RV representing CR will depend on the actual state, and the minimum of its entropy (for all possible states) may be called the amount of robust CR. Most desirable is to have robust UCR, i.e., such RV representing CR whose distribution is nearly uniform, no matter what the actual state is. Again, for the type of models we will consider, the maximum attainable amount of robust CR and robust UCR will be asymptotically the same.

We will study robust UCR for AVC's and the results allow us to determine identification capacity for various AVC models. Quite remarkably, we also obtain a new result on regular (transmission) capacity, namely that the average error capacity of an AVC with complete feedback always equals the random code capacity of this AVC.

The problem of robust uniform randomness is of interest even if it is not required that distant terminals have access to it. Then the problem is that, if several probability distributions (PD's) are given on a set $\mathcal{V}$, how large can be the number of values of a function f on $\mathcal{V}$ whose distribution is nearly uniform under each of the given PD's. Here we state a simple combinatorial lemma, similar in spirit to the hypergraph coloring lemmas of [3]. It says that if the given PD's on $\mathcal{V}$ are umiform distributions on the edges $E \in \mathcal{E}$ of a hypergraph $(\mathcal{V}, \mathcal{E})$ then the maximum number of values of an f with the required property is not much smaller than the smallest edge size $|E|$. We believe that this lemma will help the reader to develop intuition, as it helped us to arrive at the results on robust VCR in Section 3.

Log's and exp's are to the base 2. Natural logarithms are denoted by ℓn.

Lemma 1.1. *(Balanced coloring) Let $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ be a hypergraph with $|\mathcal{E}| = N$ edges, each of size $|E| \geq d$. Then for any $0 < \varepsilon < \frac{1}{2}$ and $k < d\varepsilon^2/\ell n(2N)$ there exists an ε–balanced vertex coloring with k colors, i.e., a function $f : \mathcal{V} \to \{1, \ldots, k\}$, such that*

$$\left| \frac{|f^{-1}(i) \cap E|}{|E|} - \frac{1}{k} \right| < \frac{\varepsilon}{k} \quad \textit{for every} \;\; 1 \leq i \leq k \;\; \textit{and} \;\; E \in \mathcal{E}. \tag{1.1}$$

Proof: Let $\{Z(v), v \in \mathcal{V}\}$ be a family of i.i.d. RV's such that $Pr\{Z(v) = i\} = \frac{1}{k}$, $i = 1, \ldots, k$, and let $Z_i(v) = 1$ if $Z(v) = i$, and 0 otherwise. Then for the random coloring $f(v) = Z(v)$ we have $|f^{-1}(i) \cap E| = \sum_{v \in E} Z_i(v)$, and the standard large deviation bound for the binomial distribution gives, for every fixed $1 \leq i \leq k$ and $E \in \mathcal{E}$, that

$$Pr\left\{|f^{-1}(i) \cap E| < \frac{1-\varepsilon}{k}|E|\right\} \leq \exp\left\{-|E|D\left(\frac{1-\varepsilon}{k}\|\frac{1}{k}\right)\right\}$$

$$Pr\left\{|f^{-1}(i) \cap E| > \frac{1+\varepsilon}{k}|E|\right\} \leq \exp\left\{-|E|D\left(\frac{1+\varepsilon}{k}\|\frac{1}{k}\right)\right\},$$

where $D(p\|q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.

Calculus shows that $D\left(\frac{1+\varepsilon}{k}\|\frac{1}{k}\right) - \frac{\varepsilon^2}{k\ell n 2}$ is a convex function of ε in the interval $-\frac{1}{2} \leq \varepsilon \leq \frac{1}{2}$, with minimum equal to 0 attained at $\varepsilon = 0$. It follows that the probability that (1.1) does not hold for the random coloring $f(v) = Z(v)$ is upper bounded by $N \cdot 2 \exp(-d\varepsilon^2/k\ell n 2)$. Under the hypothesis of Lemma 1.1, this bound is less than 1, and the assertion follows.

2 Preliminaries

A key concept studied in this paper for various models is what we call CR capacity. In this section, we first formally describe one model to be considered, and define achievable CR rates and CR capacity for that model. Then we indicate the changes needed for other models, including those where the underlying statistics depend on unknown parameters. (For all models considered, *alternative* definitions of capacities — or capacity functions — lead to the same values.) As one of our reasons for studying CR capacity is its relationship to ID capacity, at the end of this section we sketch how the latter concept can be defined for the type of models we are interested in, as a straightforward extension of the definition of ID capacity of a DMC without or with feedback, cf. [4], [5]. A general definition of transmission capacity is also included.

In Section 3 we will establish some general results, including the achievability of CR capacity with UCR, i.e., with nearly uniformly distributed RV's. For models where the statistics depend on unknown parameters, this UCR result holds in a robust sense. Then the result of Ahlswede and Dueck [5] refered to in the Introduction affords the conclusion that for the type of models considered in this paper, CR capacity is always a lower bound to ID capacity, whenever the transmission capacity is positive.

Our results on the CR capacity of particular models will be stated and proved in Sections 4 and 5.

As in Part 1, we use the terminology of the book [16], and refer to it for notation not defined here.

One of the stimuli for this investigation came from [18], where first basic observations are made and first results are established for the binary symmetric case of the model we now describe.

Model (i): Two–source with one–way communication.

Given a discrete memoryless multiple source (DMMS) with two components, with alphabets $\mathcal{X}, \mathcal{Y}$ and generic variables X, Y, the n–length source outputs are observable at Terminals $\mathcal{X}$ and $\mathcal{Y}$, respectively. Moreover, $\mathcal{X}$ can send information to $\mathcal{Y}$ over a noiseless channel of capacity R, namely, he can noiselessly transmit any function $f(X^n)$ of X^n to $\mathcal{Y}$, subject to the rate constraint

$$\frac{1}{n} \log \|f\| \leq R. \tag{2.1}$$

Other resources are not available to the terminals. We will say that a pair of RV's (K, L) is permissible if K and L are functions of the data available at $\mathcal{X}$ resp. $\mathcal{Y}$, i.e.,

$$K = K(X^n), \ L = L\big(Y^n, f(X^n)\big). \tag{2.2}$$

A permissible pair (K, L) represents ε–common randomness if

$$Pr\{K \neq L\} < \varepsilon. \tag{2.3}$$

As K and L represent the same CR, intuition requires that the entropy rates $\frac{1}{n}H(K)$ and $\frac{1}{n}H(L)$ be arbitrary close if ε is small, independently of n. In order to ensure this, via Fano's inequality, we impose the technical condition that K and L take values in the same set $\mathcal{K}$ whose cardinality satisfies

$$|\mathcal{K}| \leq \exp(cn) \tag{2.4}$$

for some c not depending on n.

For Model (i), we adopt the following definition that, with suitable interpretation, will be apropriate also for other models.

Definition 2.1: A number H is an achievable CR rate if for some constant c and every $\varepsilon > 0$, $\delta > 0$, for all sufficiently large n there exists a permissible pair of RV's (K, L) satisfying (2.3), (2.4), such that

$$\frac{1}{n}H(K) > H - \delta. \tag{2.5}$$

The largest achievable CR rate is the CR capacity.

Remark: In Part 1 [7] we considered the related concept of key capacity, where also a secrecy requirement was imposed on the CR. There, nearly uniform distribution was also required in the sense that the entropy rate $\frac{1}{n}H(K)$ be close to $\frac{1}{n}\log|\mathcal{K}|$. As stated before, CR of nearly uniform distribution or UCR is desirable also in the present context. It turns out, however, that the CR capacity in the sense of Definition 2.1 can always be attained with nearly uniformly distributed RV's, even in the stronger sense of variation distance, i.e., with K satisfying

$$\sum_{k\in\mathcal{K}} \left| Pr\{K=k\} - \frac{1}{|\mathcal{K}|} \right| < \varepsilon.$$

Actually, it will be seen in Theorem 3.2 that a still stronger kind of near uniformity can be attained, with the variation distance above going to 0 exponentially as $n\to\infty$. Of course, then also $H(K)$ will be exponentially close to $\log|\mathcal{K}|$.

For orientation notice that the CR capacity for Model (i) never exceeds $H(X)$. If $H(X|Y) < R$ then an f satisfying (2.1) can be chosen to let $\mathcal{Y}$ recover X^n from $f(X^n)$ and Y^n with small probability of error (Slepian and Wolf, [15]). Thus in this case the CR capacity equals $H(X)$.

The question of how large CR rate can be attained in the extreme case $R=0$ of Model (i), when no communication is permitted between $\mathcal{X}$ and $\mathcal{Y}$, was asked by the second author in 1970. It was answered by Gács and Körner [12] who showed that it was equal to the largest entropy of a common function of X and Y, hence always 0 if X and Y had indecomposable joint distribution. This paper was one of the starting points of multiuser information theory, at least for the Hungarian research group. It will turn out that the (by now) standard "multiuser" techniques permit to determine the CR capacity both for Model (i) and its extensions considered in Section 4.

In the model described above, randomization was not permitted. As in Part 1, we will always regard randomization (at either terminal) as generating a RV at the very start, and let further actions depend on this RV, but already in a deterministic way. Thus, Model (i) with randomization at $\mathcal{X}$ means that a RV $M=M_{\mathcal{X}}$ (of arbitrary distribution, but independent of X^n, Y^n) may be generated at $\mathcal{X}$; then the information sent to $\mathcal{Y}$ may be $f(X^n, M)$ (still subject to (2.1)), and Definition 2.1 applies with the understanding of permissible pairs as $K=K(X^n,M)$, $L=L\big(Y^n, f(X^n,M)\big)$. Randomization at $\mathcal{Y}$ might also be permitted, then $\mathcal{Y}$ could generate a RV $M_{\mathcal{Y}}$ (independent of $X^n, Y^n, M_{\mathcal{X}}$), and let L be a function of $M_{\mathcal{Y}}$, too. Notice that whereas randomization at $\mathcal{X}$ may increase the CR capacity of Model (i), randomization at $\mathcal{Y}$ can not.

A variant of Model (i) is when the given channel from $\mathcal{X}$ to $\mathcal{Y}$ is not noiseless but a DMC, say with the same word–length n as the observed source output. The input is selected by Terminal $\mathcal{X}$ as a function of X^n (or of X^n and M if randomization is permitted) and Terminal $\mathcal{Y}$ observes the output, say Z^n. Then the change required in the definition of permissible pairs (K,L) is that now $L=L(Y^n,Z^n)$.

A somewhat different model is

Model (ii): DMC with active feedback.

Given a DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$, Terminal $\mathcal{X}$ selects the inputs, Terminal $\mathcal{Y}$ observes the outputs, and $\mathcal{Y}$ can send back information to $\mathcal{X}$ over a noiseless channel of capacity R. We assume that $\mathcal{X}$ is permitted to randomize but $\mathcal{Y}$ is not. Formally, the terminals' permissible actions are as follows. Initially, $\mathcal{X}$ generates a randomization RV $M_{\mathcal{X}} = M$, then he inputs $X_1 = f_1(M)$ into the DMC. The output Y_1 is observed by $\mathcal{Y}$ who then noiselessly sends $\mathcal{X}$ a message $g_1(Y_1)$. Then $\mathcal{X}$ sends $X_2 = f_2\big(M, g_1(Y_1)\big)$ over the DMC, $\mathcal{Y}$ observes the output Y_2 and sends back $g_2(Y_1, Y_2)$. Next $\mathcal{X}$ sends $X_3 = f_3\big(M, g_1(Y_1), g_2(Y_1, Y_2)\big)$ and $\mathcal{Y}$ sends back $g_3(Y_1, Y_2, Y_3)$, etc. through n rounds.

The individual feedback messages may be arbitrary, but $g = (g_1, \ldots, g_n)$ is supposed to satisfy the global rate constraint

$$\frac{1}{n} \log \|g\| \leq R. \tag{2.6}$$

E.g., $g_1, \ldots, g_n$ may be binary words of variable length with prefix property, then (2.6) will mean that their total length is $\leq nR$.

In this model, the permissible pairs (K, L) (which will represent ε–common randomness if they satisfy (2.3)) are of the form

$$K = K(M, g_1, \ldots, g_n), \;\; L = L(Y^n). \tag{2.7}$$

With this understanding, Definition 2.1 of the achievable CR rates and CR capacity applies to the present model. Notice that some of the messages g_i may be empty, indeed it is permissible that Terminal $\mathcal{Y}$ sends only one message to $\mathcal{X}$ after having received the whole Y^n (of course, then the input X^n must be a function of M alone). We will show that the CR capacity for Model (ii) is always attainable that way.

In another version of Model (ii) also Terminal $\mathcal{Y}$ is permitted to randomize, which formally means that he, too, generates a randomization RV $M_{\mathcal{Y}}$ at the start (independent of $M_{\mathcal{X}}$), and then $g_1, \ldots, g_n$ as well as L may depend also on $M_{\mathcal{Y}}$. Still another version would be when neither terminal is allowed to randomize, but that will not be considered here.

Just as Model (i) could be modified replacing the noiseless channel from $\mathcal{Y}$ to $\mathcal{X}$ by a DMC, the same is possible also for Model (ii). Actually, several such variants of Model (ii) could be considered, one of them is when the i'th input of the backward channel is a function $g_i(Y_1, \ldots, Y_i)$ of the first i outputs of the forward channel, and Terminal $\mathcal{X}$ observes the corresponding output Z_i before selecting the input X_{i+1} to the forward channel. Then the permissible pairs (K, L) are defined by letting $K = K(M, Z^n)$, while $L = L(Y^n)$ as before.

Remark: Our terminology "active feedback" refers to the freedom of Terminal $\mathcal{Y}$ to select the inputs of the backward channel. It differs from the terminology of the book [16] where "active feedback" means that $\mathcal{Y}$ is allowed to randomize. By "passive feedback" we mean that the inputs of the backward channel are equal to the outputs of the forward channel. In particular, noiseless passive feedback (also called complete feedback) means that the outputs of the DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$ are observable not only to Terminal $\mathcal{Y}$ but also to Terminal $\mathcal{X}$. The variant of Model (ii) with complete feedback has been hinted at in the Introduction as a simple example for which the problem of CR–capacity is trivial. The variant with noisy passive feedback deserves interest, but will not be considered in this paper.

The paper [21] came recently to our attention. It gives a detailed treatment of that version of our Model ii, where the backward channel is a DMC, including the case when both terminals have available a limited amount of private randomness.

A more complex version of the two–source model is

Model (iii): Two–source with two–way noiseless communication.

Given a DMMS as in Model (i), suppose that after Terminal $\mathcal{Y}$ received the message sent by $\mathcal{X}$ over a noiseless channel of capacity R_1, he in turn can send $\mathcal{X}$ a message over a noiseless channel of capacity R_2. This can be any function g of Y^n and the received $f(X^n)$ (or $f(X^n, M_{\mathcal{X}})$), subject to the rate constraint

$$\frac{1}{n} \log \|g\| \leq R_2. \tag{2.8}$$

If $\mathcal{Y}$ is permitted to randomize, g may also depend on $\mathcal{Y}$'s randomization RV $M_{\mathcal{Y}}$, chosen at the start, independently of $(M_{\mathcal{X}}, X^n, Y^n)$.

Now (K, L) is a permissible pair of RV's if $K = K(X^n, g)$ or $K = K(X^n, M_{\mathcal{X}}, g)$ and $L = L(Y^n, f)$ or $L = L(Y^n, M_{\mathcal{Y}}, f)$, according as randomization is permitted or not. With this understanding of the permissible pairs, Definition 2.1 applies as before.

It is obvious how to extend the model to permit several rounds of communication between $\mathcal{X}$ and $\mathcal{Y}$, each transmission subject to a rate constraint. Alternatively, the transmissions may not be constrained individually only their total rate is. The CR capacity can always be defined as in Definition 2.1, letting the permissible (K, L) pairs be functions of the data that become available at the corresponding terminals after having executed a protocol allowable by the particular model.

Models with robust CR.

The simplest model of this kind is that when both terminals can observe the output of an arbitrary varying source (AVS) but have no other resources whatsoever. An AVS with alphabet $\mathcal{X}$ and state set $\mathcal{S}$ (both finite) is determined by a family $\{P(\cdot|s), s \in \mathcal{S}\}$ of PD's on $\mathcal{X}$. The distribution of the n–length source output X^n depends on the state sequence $\mathbf{s} \in \mathcal{S}^n$, and equals

$$P(\cdot|\mathbf{s}) = P(\cdot|s_1) \times \cdots \times P(\cdot|s_n) \quad \text{if} \quad \mathbf{s} = (s_1, \ldots, s_n). \tag{2.9}$$

In this model, any function $K = K(X^n)$ represents CR, thus the largest CR, for any fixed block–length n, is represented by $K = X^n$. In the definition of achievable CR rates, the condition (2.5) is now required to hold independently of the underlying statistics, i.e., for all $\mathbf{s} \in \mathcal{S}^n$. Thus the CR capacity for this model equals $H_{\min} = \min_{s \in \mathcal{S}} H\big(P(\cdot|s)\big)$. It is non–trivial but will be shown in Theorem 3.1 that this CR capacity can be attained with robust UCR, i.e., that $K = K(X^n)$ satisfying (2.6) for all possible $\mathbf{s} \in \mathcal{S}^n$ can be given, such that

$$\frac{1}{n} \log |\mathcal{K}| > H_{\min} - \delta.$$

We will consider various AVC models in this paper. An AVC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and state set $\mathcal{S}$, each finite, is determined by a family $\mathcal{W} = \{W(\cdot|\cdot, s), s \in \mathcal{S}\}$ of channels $W(\cdot|\cdot, s) : \mathcal{X} \to \mathcal{Y}$. Terminal $\mathcal{X}$ selects the inputs, Terminal $\mathcal{Y}$ observes the outputs, and the state sequence $\mathbf{s} \in \mathcal{S}^n$ governing the n–length transmission may be arbitrary. Several different models are possible according to the availability of information to Terminal $\mathcal{X}$ about the states and the previous outputs when selecting the input X_i, and whether or not randomization is allowed.

We now formally describe two models, both with randomization permitted at $\mathcal{X}$, thus Terminal $\mathcal{X}$ first generates a randomization RV $M_{\mathcal{X}} = M$. In the "no feedback" model, Terminal $\mathcal{X}$ selects the input sequence X^n as a function of M. In the "complete feedback" model, the inputs $X_1, \dots, X_n$ are selected successively as $X_i = f_i(M, Y_1, \dots, Y_{i-1})$, where $Y_1, \dots, Y_{i-1}$ are the previous outputs ("seen" by Terminal $\mathcal{X}$ through a noiseless feedback channel from $\mathcal{Y}$ to $\mathcal{X}$). In both models, the joint distribution of M and the output sequence Y^n, when the state sequence is $\mathbf{s} = (s_1, \dots, s_n)$, is given by

$$Pr\{M = m, Y^n = \mathbf{y}\} = Pr\{M = m\} \prod_{i=1}^{n} W(y_i|x_i, s_i). \tag{2.10}$$

Here x_i denotes the i'th input symbol when $M = m$ (in the no feedback model) or when $M = m$, $Y_1 = y_1, \dots, Y_{i-1} = y_{i-1}$ (in the complete feedback model). For both models, the CR capacity is defined as in Definition 2.1, requiring (2.3) and (2.5) to hold robustly, i.e., for every $\mathbf{s} \in \mathcal{S}^n$. The permissible pairs K, L are of form $K = K(M)$, $L = L(Y^n)$ in the no feedback case, and formally, K should be replaced by $K = K(M, Y^n)$ in the complete feedback case; for the latter model, however, $K = L = L(Y^n)$ may be taken, without restricting generality.

Both the "no feedback" and "complete feedback" AVC models can be modified by letting Terminal $\mathcal{X}$ know the state sequence $\mathbf{s}$. Then the inputs $X_1, \dots, X_n$ and the RV K may depend also on $\mathbf{s}$. Also the AVC analogue of Model (ii), i.e., AVC with active feedback could be considered.

We will not attempt to give a general formal definition of the class of models we are interested in, but all our models involve the block–length n of observable source RV's or permissible channel transmissions (or both). For such models, Definition 2.1 always makes sense if we specify, for every n, the class of permissible pairs of RV's that may be generated by the terminals as functions of the data available to them. We now sketch how ID codes and ID capacity can be defined for arbitrary models of this kind, as a straightforward extension of the corresponding definitions for channels without or with feedback, [4], [5].

Suppose one of N contingencies $k \in \{1, \dots, N\}$ takes place, Terminal $\mathcal{X}$ knows this k, and the goal is to let Terminal $\mathcal{Y}$ reliably decide, for any $1 \le j \le N$ he may choose, whether or not $k = j$. To this end, the terminals perform a protocol permissible by the given model, for some block–length n, with the understanding that the actions of Terminal $\mathcal{X}$, but not those of Terminal $\mathcal{Y}$, may explicitly depend on k. E.g., for Model (ii), the functions $f_i\big(M, g_1(Y_1), \ldots, g_{i-1}(Y_1, \ldots, Y_{i-1})\big)$ specifying the channel inputs X_i, will depend on k, whereas for the feedback messages $g_i(Y_1, \ldots, Y_i)$ no such dependence is allowed, except for implicit dependence through the Y_i's. Let U denote all information available at Terminal $\mathcal{Y}$ after having performed the protocol, e.g., for Model (ii), $U = Y^n$. Then, if $\mathcal{Y}$ wants to decide whether or not $k = j$, he decides "yes" if $U \in D_j$ and "no" if $U \notin D_j$, where D_j, $1 \le j \le N$ are certain subsets of $\mathcal{U}$, the range of U.

Definition 2.2: A protocol as above together with a family $\{D_j, 1 \le j \le N\}$ of subsets of $\mathcal{U}$ is called an (N, n, ε) ID code for the given model if for each distinct k, j in $\{1, \ldots, N\}$

$$P_j(D_j^c) \le \varepsilon; \quad P_k(D_j) \le \varepsilon. \tag{2.11}$$

Here P_k denotes the distribution of U when contingency k has taken place. The ID capacity of the given model is the supremum of the numbers R such that for every $\varepsilon > 0$ and sufficiently large n there exists an (N, n, ε) ID code with $N \ge \exp\exp(nR)$.

For models whose statistics depend on unknown parameters ("state"), Definition 2.2 applies with the obvious modification. Namely, as then the distributions P_k also depend on the state, we require (2.11) to hold robustly, i.e., for all possible states. In particular, for an AVC without feedback (with $\mathcal{X}$ permitted to randomize) an (N, n, ε) ID code is defined by a family $\{Q_j, \ \ 1 \le j \le N\}$ of PD's on $\mathcal{X}^n$, Q_j representing the distribution of the input sequence when contingency j takes place, together with a family $\{D_j, \ \ 1 \le j \le N\}$ of subsets of $\mathcal{Y}^n$, such that for each distinct k, j in $\{1, \ldots, N\}$ and all $\mathbf{s} \in \mathcal{S}^n$

$$\sum_{\mathbf{x} \in \mathcal{X}^n} Q_j(x) W^n(D_j^c | \mathbf{x}, \mathbf{s}) \le \varepsilon, \quad \sum_{\mathbf{x} \in \mathcal{X}^n} Q_k(\mathbf{x}) W^n(D_j | \mathbf{x}, \mathbf{s}) \le \varepsilon. \tag{2.12}$$

It is important to emphasize that the sets D_j in Definition 2.2 need not be disjoint. If they were, Terminal $\mathcal{Y}$ could infer k (as that j for which $U \in D_j$) with probability of error less than ε, thus the ID code would become a transmission code. Whereas for ID codes N can grow doubly exponentially with the block–length n, for transmission codes only exponential growth is possible.

As a straightforward generalization of the concept of channel capacity, we can define the transmission capacity of a general model as the supremum of numbers R such that for every $\varepsilon > 0$ and sufficiently large n there exists an (N, n, ε) transmission code. Notice that for transmission codes, i.e., when the sets D_j, $1 \le j \le N$, are disjoint, it suffices to impose the first inequality in (2.11). More exactly, the transmission capacity defined in this way is that for the "maximum error" criterion, whereas transmission capacity for the "average error" criterion is obtained if the transmission codes are required to satisfy only

$$\frac{1}{N} \sum_{j=1}^{N} P_j(D_j^c) \le \varepsilon, \tag{2.13}$$

a weaker condition than (2.11). Just as for standard channel capacity, these two concepts of transmission capacity coincide for models with uniquely determined statistics, but transmission capacity for average error can be larger than that for maximum error when the statistics depend on unknown states.

Remark: For models with randomization allowed at Terminal $\mathcal{X}$, transmission capacity (for average error) is always a lower bound to CR capacity. Indeed, a trivial way of generating CR is that Terminal $\mathcal{X}$ generates a RV uniformly distributed on $\{1, \ldots, N\}$ and then transmits it to Terminal $\mathcal{Y}$ with probability of error less than ε. From the point of view of CR capacity, the interesting models are those for which this trivial scheme is not optimal.

3 Some general results

Lemma 3.1. *Let $\mathcal{P}$ be any family of N PD's $P = \{p(v), v \in V\}$ on a finite set V, let $0 < \varepsilon \leq \frac{1}{9}$ and let $d > 0$ be such that for every $P \in \mathcal{P}$ the set*

$$E(P, d) = \{v : p(v) \leq \frac{1}{d}\} \tag{3.1}$$

has P–probability

$$P\big(E(P, d)\big) \geq 1 - \varepsilon. \tag{3.2}$$

Then for $k \leq \frac{\varepsilon^2}{3\log(2N)} d$, there exists $f : V \to \{1, \ldots, k\}$ such that for every $1 \leq i \leq k$ and $P \in \mathcal{P}$ the conditional P–probability of $f(v) = i$ on the condition $v \in E(P, d)$ differs from $\frac{1}{k}$ by less than $\frac{\varepsilon}{k}$, i.e.,

$$\left| \frac{P\big(f^{-1}(i) \cap E(P, d)\big)}{P\big(E(P, d)\big)} - \frac{1}{k} \right| < \frac{\varepsilon}{k}, \quad 1 \leq i \leq k, P \in \mathcal{P}. \tag{3.3}$$

In particular, the variation distance of the distribution of f from the uniform distribution on $\{1, \ldots, k\}$ is less than 3ε, i.e.,

$$\sum_{i=1}^{k} \left| P\big(f^{-1}(i)\big) - \frac{1}{k} \right| < 3\varepsilon, \tag{3.4}$$

for each of the PD's $P \in \mathcal{P}$.

Proof: Similar to that of Lemma 1.1 but requires a little more calculation. See Appendix.

Consider now the problem of robust uniform randomness obtainable by encoding the n–length output X^n of an AVS, where the distribution of X^n depending on the state sequence $\mathbf{s} \in \mathcal{S}^n$ is given by (2.9). We are interested in mappings $f : \mathcal{X}^n \to \mathcal{M}$ of possibly large rate $\frac{1}{n}\log|\mathcal{M}|$ for which $f(X^n)$ represents robust ε–uniform randomness, i.e., the variation distance of the distribution of $f(X^n)$ from the uniform distribution on $\mathcal{M}$ is less than ε, no matter what is the state sequence $\mathbf{s}\in\mathcal{S}^n$.

Theorem 3.1. *Let us be given an AVS by a set of PD's $\{P(\cdot|s), s\in\mathcal{S}\}$ on $\mathcal{X}$, such that $H_{\min}=\min_{s\in\mathcal{S}} H\big(P(\cdot|s)\big)>0$. Then for every $0<\varepsilon<\frac{1}{3}$ and every n there exists a mapping $f:\mathcal{X}^n\to\mathcal{M}$ of rate*

$$\frac{1}{n}\log|\mathcal{M}| > H_{\min}-\delta(\varepsilon,n) \tag{3.5}$$

such that $f(X^n)$ represents robust ε–uniform randomness, where

$$\delta(\varepsilon,n)=\sqrt{\frac{2\ell n 3/\varepsilon}{n}}\log|\mathcal{X}|+\frac{2\log 1/\varepsilon}{n}+\frac{\log\log(2|\mathcal{S}|)}{n}+0\left(\frac{\log n}{n}\right) \tag{3.6}$$

if $|\mathcal{X}|\ge 3$, and $|\mathcal{X}|$ should be replaced by 3 if $|\mathcal{X}|=2$; the $0\left(\frac{\log n}{n}\right)$ term in (3.6) does not depend on ε and the AVS, not even on $\mathcal{X}$ and $\mathcal{S}$.

Remarks: One feature of Theorem 3.1 that will be used in Theorem 3.2 below is that it brings out explicitly the dependence of $\delta(\varepsilon,n)$ on $\mathcal{X}$ and $\mathcal{S}$. For a fixed AVS, Theorem 3.1 shows that robust ε–uniform randomness for (arbitrarily small but) constant ε can be attained by mappings of rate approaching $H_{\min}$ with speed $0(n^{-\frac{1}{2}})$, and the rate will approach $H_{\min}$ even if $\varepsilon=\varepsilon_n\to 0$, providing it goes to 0 slower than exponentially. Moreover, robust ε–uniform randomness with rate $\frac{1}{n}\log|\mathcal{M}|>H_{\min}-\delta$ with an arbitrarily small but constant $\delta>0$ is attainalbe even with ε going to 0 exponentially.

Proof: Apply Lemma 3.1 to the family of PD's $P(\cdot|\mathbf{s})$, $\mathbf{s}\in\mathcal{S}^n$, on $V=\mathcal{X}^n$, with ε replaced by $\varepsilon/3$ (in order to get ε–uniform rather than 3ε–uniform randomness, cf. (3.4)). Then $N=|\mathcal{S}|^n$, and we will choose the number d in (3.1) as

$$d=\exp\big[n(H_{\min}-\xi)\big], \tag{3.7}$$

with $\xi>0$ such that (3.2) (with ε replaced by $\varepsilon/3$) is fulfilled for each $P=P(\cdot|\mathbf{s})$. As shown in the Appendix,

$$\xi=\sqrt{\frac{2\ell n 3/\varepsilon}{n}}\log|\mathcal{X}| \tag{3.8}$$

is an adequate choice, with the understanding (as also in the rest of the proof) that $|\mathcal{X}|$ should be replaced by 3 if $|\mathcal{X}|=2$. Then Lemma 3.1 gives that for

$$|\mathcal{M}|\le\frac{(\varepsilon/3)^2}{3\log(2|\mathcal{S}|^n)}\exp\left[n\left(H_{\min}-\sqrt{\frac{2\ell n 3/\varepsilon}{n}}\log|\mathcal{X}|\right)\right] \tag{3.9}$$

there exists $f:\mathcal{X}^n\to\mathcal{M}$ such that $f(X^n)$ represents ε–uniform randomness, for each $\mathbf{s}\in\mathcal{S}^n$. Comparison of (3.5) and (3.9) shows that both can be satisfied with $\delta(\varepsilon,n)$ as in (3.5).

Having available Theorem 3.1, we now prove that for the type of models treated in this paper, CR capacity can be attained with uniform CR. Although we did not give a formal definition of this class of models, we recall from Section 2 that all our models involve the specification of permissible pairs of RV's (K, L), for each block–length n. The following definition postulates a property common to all models we are interested in.

Definition 3.1: A model permits independent concatenations if for any pairs of RV's (K_1', L_1') and (K_2', L_2') permissible for block–lengths n_1 and n_2, there exists a pair (K, L) permissible for block–length n_1+n_2 such that $K = (K_1, K_2)$, $L = (L_1, L_2)$, where (K_1, L_1) and (K_2, L_2) are independent and have the same distribution as (K_1', L_1') and (K_2', L_2'). When the underlying statistics are not uniquely determined but depend on some parameters ("state"), the last condition means that under any permissible statistics for block–length $n_1 + n_2$, (K_1, L_1) and (K_2, L_2) are independent, with distributions equal to those of (K_1', L_1') and (K_2', L_2') under one of permissible statistics for block–length n_1 resp. n_2.

For models with statistics depending on "states", let $\mathcal{S}(n)$ denote the set of possible states for block–length n. We will assume that this set does not grow faster than doubly exponentially, more exactly, that $\frac{1}{n}\log\log|\mathcal{S}(n)|$ is bounded by a constant. This holds for all models we are aware of, e.g. for the standard AVS and AVC models $|\mathcal{S}(n)| = |\mathcal{S}|^n$ grows only exponentially. Even for the variant of the AVC where the state sequence $\mathbf{s}$ may depend on the input sequence $\mathbf{x}$, in which case $\mathcal{S}(n)$ is the set of all mappings of $\mathcal{X}^n$ into $\mathcal{S}^n$, the growth rate of $|\mathcal{S}(n)|$ is "only" doubly exponential.

Theorem 3.2. *Let us be given a model permitting independent concatenations. If the statistics are not uniquely determined, we assume that $\frac{1}{n}\log\log|\mathcal{S}(n)|$ is bounded. Then for any fixed $\varepsilon > 0$, every H less than CR capacity, and sufficiently large n, there exists a permissible pair of RV's (K, L), both distributed on a set $\mathcal{M}$ satisfying $\frac{1}{n}\log|\mathcal{M}| \geq H$, such that*

$$Pr\{K \neq L\} < \varepsilon, \ \sum_{k\in\mathcal{M}} \left| Pr\{K = k\} - \frac{1}{|\mathcal{M}|} \right| < \varepsilon, \tag{3.10}$$

for every possible choice of the underlying statistics.

Remark: It will be clear from the proof that the near uniformity of K can be attained also in a stronger sense, namely in the second inequality in (3.10), instead of a fixed $\varepsilon > 0$ one could take a sequence ε_n going to 0 exponentially as $n \to \infty$ (with a sufficiently small exponent). A similar improvement of the first inequality in (3.10) is possible providing in the definition of CR capacity, the fixed $\varepsilon > 0$ in (2.3) can be replaced by ε_n going to 0 exponentially; this holds for all the models treated in this paper.

Proof: As H is less than CR capacity, there exists $H' > H$ which is still an achievable CR rate. Applying Definition 2.1 to H' in the role of H, with $\delta' = \frac{H'-H}{2}$, and $\varepsilon' > 0$ specified later, it follows that for sufficiently large m there exists a pair (K', L') permissible for block–length m such that their common range $\mathcal{K}$ satisfies

$$|\mathcal{K}| \leq \exp(cm), \tag{3.11}$$

and

$$Pr\{K' \neq L'\} < \varepsilon', \tag{3.12}$$

$$\frac{1}{m} H(K') > H' - \delta' = H + \delta', \tag{3.13}$$

for every choice of the underlying statistics. Clearly, the case of uniquely determined statistics $\big(|\mathcal{S}(m)| = 1\big)$ need not be considered separately.

As the model permits independent concatenations, for every r there exists a pair (K^r, L^r) permissible for block–length $n = rm$, with $K^r = K_1 \dots K_r$, $L^r = L_1 \dots L_r$, such that for every possible statistics for block–length n the pairs (K_i, L_i), $i = 1, \dots, r$ are independent, with distributions equal to that of (K', L') for some possible statistics for block–length m (possibly different for each i). In particular, K^r may be regarded as the r–length output of an AVS with alphabet $\mathcal{K}$ and state set $\mathcal{S}(m)$. For this AVS, $H_{\min} > m(H + \delta')$ by (3.13). Thus by Theorem 3.1 there exists a mapping $f : \mathcal{K}^r \to \mathcal{M}$ with

$$\frac{1}{r} \log |\mathcal{M}| > m(H + \delta') - \delta(\varepsilon, r) \tag{3.14}$$

such that the distribution of $f(K^r)$ is robustly ε–close to the uniform distribution on $\mathcal{M}$, where

$$\delta(\varepsilon, r) = \sqrt{\frac{2\ell n 1/\varepsilon}{r}} \log |\mathcal{K}| + \frac{2 \log 1/\varepsilon}{r} + \frac{\log \log |2\mathcal{S}(m)|}{r} + 0 \left(\frac{\log r}{r} \right). \tag{3.15}$$

Using (3.11) and the assumption on the growth rate of $\mathcal{S}(m)$, it follows from (3.14) and (3.15) that $\frac{1}{rm} \log |\mathcal{M}| > H$ if r is sufficiently large, dependeing on ε but not on m.

With such an r we set $K = f(K^r)$, $L = f(L^r)$ for block–length $n = rm$. Then K and L are distributed on $\mathcal{M}$ satisfying $\frac{1}{n} \log |\mathcal{M}| > H$, and the second inequality in (3.10) holds for every possible choice of the underlying statistics. Finally, the first inequality in (3.10) follows from (3.12), if we choose $\varepsilon' = \varepsilon / r$. This completes the proof, because it clearly suffices to restrict attention to block–lengths n which are multiples of a constant r.

Theorem 3.3. *For all models as in Theorem 3.2, the CR capacity is a lower bound to ID capacity, provided the transmission capacity (for the maximum error criterion) is positive.*

Proof: Immediate from Theorem 3.2 and the result of Ahlswede and Dueck [5].

4 Common randomness in models (i), (ii), and (iii)

Theorem 4.1. *For Model (i) described in Section 2, the CR capacity equals*

$$C_1(R) = \max_U \big[I(U \wedge X) | I(U \wedge X) - I(U \wedge Y) \leq R \big] \tag{4.1}$$

if no randomization is permitted, and

$$\tilde{C}_1(R) = \begin{cases} C_1(R) & \text{if } \; R \leq H(X|Y) \\ R + I(X \wedge Y) & \text{if } \; R \geq H(X|Y) \end{cases} \tag{4.2}$$

if Terminal $\mathcal{X}$ is allowed to randomize. Here the maximum is for all RV's U that satisfy the Markov condition $U \multimap X \multimap Y$, and the range constraint $|\mathcal{U}| \leq |\mathcal{X}|$, and R is the capacity of the noiseless channel in the model. Moreover, the CR capacity of the variant of Model (i) where the noiseless channel is replaced by a DMC, is still given by (4.1) resp. (4.2), with R replaced by the capacity of that DMC.

Remark: If $\mathcal{X}$ is permitted to randomize, a trivial way to create CR is that $\mathcal{X}$ generates nR random bits and transmits them to $\mathcal{Y}$, disregarding the DMMS. Theorem 4.1 shows that this is suboptimal, and for $R \geq H(X|Y)$ the CR capacity exceeds R (attained by the trivial scheme) by exactly $I(X \wedge Y)$. This means that although mutual information does not represent a "common information" (as shown in [12]), it does represent a kind of hidden common randomness that can be recovered if sufficient transmission capacity is available. It is interesting to compare this interpretation of mutual information with that obtained in Part 1 in a context involving secrecy.

Proof: A short proof is available using standard results of multiuser information theory, cf. the proof of Theorem 4.2 below. Here we prefer an independent proof, which later will be extended to the case of two–way communication.

We state, also for later reference, an identity also used in Part 1 (Lemma 4.1; cf. also [16], p. 409): For arbitrary RV's S, T and sequences of RV's X^n, Y^n

$$\begin{aligned} & I(S \wedge X^n|T) - I(S \wedge Y^n|T) \\ &= \sum_{i=1}^{n} \big[I(S \wedge X_i|X_1, \ldots, X_{i-1}, Y_{i+1} \ldots Y_n) - I(S \wedge Y_i|X_1 \ldots X_{i-1} Y_{i+1} \ldots Y_n) \big] \\ &= n \big[I(S \wedge X_J|V) - I(S \wedge Y_J|V) \big] \end{aligned} \tag{4.3}$$

where J is a RV independent of all the previous ones, uniformly distributed on $\{1, \ldots, n\}$, and

$$V = X_1 \ldots X_{J-1} Y_{J+1} \ldots Y_n T J. \tag{4.4}$$

(a) Converse part.

Consider first the "no randomization" case. Suppose (K, L) satisfy (2.1) – (2.4). Write

$$H(K|Y^n) = I\big(K \wedge f(X^n)|Y^n\big) + H\big(K|Y^n, f(X^n)\big). \tag{4.5}$$

Here the first term is $\leq nR$, by (2.1), and the second term is $\leq H(K|L) \leq \varepsilon cn + 1$ by (2.2) and Fano's inequality, using (2.3), (2.4). Thus we have

$$H(K) - I(K \wedge Y^n) = H(K|Y^n) \leq nR + \varepsilon cn + 1. \tag{4.6}$$

Apply (4.3) to the present X^n, Y^n with $S = K, T = \varnothing$. Then V in (4.4) is independent of (X_J, Y_J), hence the last line in (4.3) can also be written as

$$I(U \wedge X_J) - I(U \wedge Y_J), \quad \text{with} \quad U = KV.$$

Thus

$$H(K) - I(K \wedge Y^n) = I(K \wedge X^n) - I(K \wedge Y^n) = n\big[I(U \wedge X_J) - I(U \wedge Y_J)\big] \tag{4.7}$$

where $U = KX_1 \dots X_{J-1}Y_{J+1} \dots Y_n$ satisfies the Markov condition $U \multimap X_J \multimap Y_J$. Notice also that

$$\begin{aligned} H(K) = I(K \wedge X^n) &= \sum_{i=1}^{n} I(K \wedge X_i | X_1 \dots X_{i-1}) \\ &= nI(K \wedge X_J | X_1, \dots, X_{J-1}) \le nI(U \wedge X_J). \end{aligned} \tag{4.8}$$

As X_J, Y_J may be identified with the generic variables X, Y of our DMMS, (4.6), (4.7) and (4.8) show that $\frac{1}{n}H(K)$ is upper bounded by the maximum of $I(U \wedge X)$ subject to $I(U \wedge X) - I(U \wedge Y) \le R + \varepsilon c + \frac{1}{n}$, for RV's U satisfying the Markov condition $U \multimap X \multimap Y$. It is routine to show that there exists U attaining the maximum that satisfies the range constraint $|\mathcal{U}| \le |\mathcal{X}|$ (direct application of the Support Lemma of [16], p. 310 gives only $|\mathcal{U}| \le |\mathcal{X}| + 1$, but for a U yielding an extremal value, this bound can be improved by 1, cf. [19]).

This completes the proof for the "no randomization" case. Notice that in (4.1) necessarily $I(U \wedge Y) \le I(X \wedge Y)$ hence

$$C_1(R) \le R + I(X \wedge Y), \quad \text{equality holds if} \quad R = H(X|Y). \tag{4.9}$$

When $\mathcal{X}$ may randomize, we will conveniently regard his randomization RV $M_{\mathcal{X}}$ as an i.i.d. sequence M^n (of course, independent of X^n, Y^n). This reduces the present case to the previous one, replacing X by XM, where M is independent of X, Y. Thus we need to maximize $I(U \wedge XM)$ subject to

$$I(U \wedge XM) - I(U \wedge Y) \le R, \; U \multimap XM \multimap Y. \tag{4.10}$$

It follows similarly to (4.9) that (4.10) implies $I(U \wedge XM) \le R + I(X \wedge Y)$, thus for the case $R \ge H(X|Y)$ we are done. For $R < H(X|Y)$, notice that since $I(U \wedge XM) = I(U \wedge X) + I(U \wedge M|X)$, and the Markov condition in (4.10) implies $U \multimap X \multimap Y$, it follows from (4.9) that

$$I(U \wedge XM) \le C_1\big(R - I(U \wedge M|X)\big) + I(U \wedge M|X). \tag{4.11}$$

It is easy to check that the function defined by (4.1) is concave, hence, by (4.9), its slope is ≥ 1 if $R \le H(X|Y)$. Thus the right hand side of (4.11) is $\le C_1(R)$. This completes the proof for the randomized case.

Finally, if the channel from $\mathcal{X}$ to $\mathcal{Y}$ is not noiseless but a DMC, the only modification needed in the above proof is to replace $f(X^n)$ in eq. (4.5) by the output of that DMC. Denote the input of this DMC by T^n and the output by Z^n. Whether or not Terminal $\mathcal{X}$ randomizes, the Markov condition $Y^n \multimap X^n K \multimap T^n \multimap Z^n$ must hold. Thus the first term in (4.5) with $f(X^n)$ replaced by Z^n can be bounded as

$$I(K \wedge Z^n|Y^n) \le I(X^n K \wedge Z^n|Y^n) \le I(T^n \wedge Z^n|Y^n) \le nC,$$

establishing our claim.

(b) Direct part.

It suffices to consider the case $R \le H(X|Y)$, with no randomization. By continuity, it suffices to show that $C_1(R')$ is an achievable CR rate for every $R' < R$. We are going to show this by exhibiting for arbitrary U satisfying

$$U \multimap X \multimap Y, \quad I(U \wedge X) - I(U \wedge Y) < R \tag{4.12}$$

and for any $\varepsilon > 0$, $\delta > 0$ and sufficiently large n, a permissible pair K, L as defined by (2.1), (2.2), such that K, L satisfy (2.3), (2.4), and (2.5) with $H = I(U \wedge X)$.

Assuming without any loss of generality that the distribution of U is a possible type for block–length n, select at random $\exp\{n(I(U \wedge X) + \delta)\}$ sequences $\mathbf{u} \in \mathcal{U}^n$ of type P_U, denoted as $\mathbf{u}_{ij}$, $1 \le i \le N_1$, $1 \le j \le N_2$, with

$$N_1 = \exp\{n(I(U \wedge X) - I(U \wedge Y) + 3\delta)\}, N_2 = \exp\{n(I(U \wedge Y) - 2\delta)\}. \tag{4.13}$$

Then for every X–typical $\mathbf{x} \in \mathcal{X}^n$ the probability that neither $\mathbf{u}_{ij}$ is jointly UX–typical with $\mathbf{x}$ is doubly exponentially small. Hence with probability close to 1, every typical $\mathbf{x}$ is jointly typical with some $\mathbf{u}_{ij}$.

Let $K(\mathbf{x})$ be equal to an $\mathbf{u}_{ij}$ jointly typical with $\mathbf{x}$ (either one if there are several), and let $f(\mathbf{x}) = i$ if $K(\mathbf{x}) = \mathbf{u}_{ij}$; both functions are set constant when $\mathbf{x}$ is not typical. Further let $L\big(\mathbf{y}, f(\mathbf{x})\big) = \mathbf{u}_{ij}$ if $f(\mathbf{x}) = i$ and $\mathbf{u}_{ij}, \mathbf{y}$ are jointly UY–typical. If there is no such $\mathbf{u}_{ij}$ or there are several, L is set equal to a constant. Then, by (4.12), (4.13), the rate constraint (2.1) on f is satisfied if δ is sufficiently small, and $K = K(X^n)$, $L = L\big(Y^n, f(X^n)\big)$ obviously satisfy (2.4). Also (2.5) is satisfied since

$$\begin{aligned} Pr\{K = \mathbf{u}_{ij}\} &\le P_X^n\big(\{\mathbf{x} : (\mathbf{u}_{ij}, \mathbf{x}) \ \text{ jointly typical}\}\big) \\ &= \exp\big(-nI(U \wedge X) + o(n)\big) \end{aligned} \tag{4.14}$$

implies that $H(K) \ge nI(U \wedge X) + o(n)$.

It remains to check (2.3), i.e., $Pr\{K \ne L\} \le \varepsilon$. Notice that for any jointly UX–typical pair $(\mathbf{u}, \mathbf{x})$, the set of $\mathbf{y}$'s jointly typical with $(\mathbf{u}, \mathbf{x})$ has conditional probability arbitrarily close to 1 on the condition $U^n = \mathbf{u}$, $X^n = \mathbf{x}$, and hence by markovity, also on the condition $X^n = \mathbf{x}$. It follows that the set A of those pairs $(\mathbf{x}, \mathbf{y})$ for which $\big(K(\mathbf{x}), \mathbf{x}, \mathbf{y}\big)$ are jointly UXY–typical has P_{XY}^n arbitrarily close to 1. Let us denote by B the set of those pairs $(\mathbf{x}, \mathbf{y}) \in A$ for which in addition to $\mathbf{u}_{ij} = K(\mathbf{x})$, some other $\mathbf{u}_{ij}$ (with the same first index

i) is also jointly typical with $\mathbf{y}$. To complete the proof, it suffices to show that $P^n_{XY}(B)$ with be arbitrarily small, with large probability with respect to the random choice of $\{\mathbf{u}_{ij}\}$.

Now, for fixed $(\mathbf{x}, \mathbf{y})$, the probability that B determined by the random $\{\mathbf{u}_{ij}\}$ contains $(\mathbf{x}, \mathbf{y})$, is upper bounded by

$$\begin{aligned}
&\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{\substack{\ell=1 \\ \ell \neq j}}^{N_2} Pr\big\{(\mathbf{u}_{ij}, \mathbf{x}) \text{ jointly typical, } (\mathbf{u}_{ij}, \mathbf{y}) \text{ jointly typical } \big\} \\
&= N_1 N_2^2 \exp\big[-nI(U \wedge X) + o(n)\big] \exp\big[-nI(U \wedge Y) + o(n)\big] \\
&= \exp\big[-n\delta + o(n)\big]; \qquad (4.15)
\end{aligned}$$

here we used (4.13) and that the $\mathbf{u}_{ij}$ are independent, chosen with uniform distribution from the sequences of type P_U. Hence the expectation of $P^n_{XY}(A)$, as a RV depending on $\{\mathbf{u}_{ij}\}$, is also upper bounded by $\exp\big[-n\delta + o(n)\big]$. This completes the proof.

Consider now the following generalization of Model (i) to generating CR at $r+1$ (rather than 2) terminals. Given a DMMS with $r+1$ components, with generic variables $X, Y_1, \ldots, Y_r$, Terminal $\mathcal{X}$ can observe X^n and send messages $f_i(X^n)$ to Terminals $\mathcal{Y}_i$, subject to rate constraints

$$\frac{1}{n} \log \|f_i\| \leq R_i, \ i = 1, \ldots, r. \qquad (4.16)$$

Terminal $\mathcal{Y}_i$ can observe Y_i^n, and the message $f_i(X^n)$ sent him by Terminal $\mathcal{X}$. Achievable CR rates and CR capacity are defined by the natural extension of Definition 2.1, namely the role of permissible pairs (K, L) is now played by permissible $(r+1)$–tuples $(K, L_1, \ldots, L_r)$ defined in analogy to (2.2), and the role of condition (2.3) is played by r similar conditions $Pr\{K \neq L_i\} < \varepsilon$, $i = 1, \ldots, r$.

Theorem 4.2. *For the above model, with no randomization permitted, the CR capacity equals the maximum of $I(U \wedge X)$ subject to the constraints*

$$I(U \wedge X) - I(U \wedge Y_i) \leq R_i, \ i = 1, \ldots, k, \ U \multimap X \multimap Y_1, \ldots, Y_k, \qquad (4.17)$$

where U may be supposed to satisfy the range constraint $|\mathcal{U}| \leq |\mathcal{X}| + r - 1$. If Terminal $\mathcal{X}$ is permitted to randomize, the CR capacity is still the same if $R_i < H(X|Y_i)$ for some i, and it equals $\min_{1 \leq i \leq k} \big[R_i + I(X \wedge Y_i)\big]$ if $R_i \geq H(X|Y_i)$, $i = 1, \ldots, k$

Proof: If H is an achievable CR rate and $\delta > 0$, take (for large n) $K = K(X^n)$ that can be ε–reproduced at each terminal by $L_i = L_i\big(Y_i^n, f_i(X^n)\big)$, $i = 1, \ldots, r$, and such that

$$\left| \frac{1}{n} H(K) - H \right| < \delta. \qquad (4.18)$$

Although the definition of achievable CR rates postulates $\frac{1}{n}H(K) > H - \delta$ only, it clearly does not restrict generality to require $\frac{1}{n}H(K) < H + \delta$, as well. In order that K could be ε–reproduced at $\mathcal{Y}_i$ it is necessary that

$$\frac{1}{n}H(K|Y_i^n) < R_i + \delta \tag{4.19}$$

(formally, $H(K|Y_i^n)$ can be written as a sum of two terms as in (4.5), and bounded as there).

On the other hand, if to a number H for all $\delta > 0$ and sufficiently large n there exists a function $K = K(X^n)$ that satisfies (4.18), (4.19), then H is an achievable entropy rate. Indeed, from Y_i^{nr} and a suitable code $f_i(K^r)$ of rate $\frac{1}{nr}\log\|f_i\| \le R_i + \delta$ of the r–fold repetition of K, Terminal $\mathcal{Y}_i$ can reproduce K^r with arbitrarily small probability of error, by Slepian–Wolf. Although the permissible rate is only R_i, this can be remedied by taking block–length $N = n'r$ with n' slightly larger than n, satisfying $n'R_i \ge n(R_i + \delta)$, $i = 1, \dots, r$, and disregarding the last $N - nr$ source outputs. Thus for block–length N, the terminals can produce ε–common randomness of rate $\frac{1}{N}rH(K) = \frac{1}{n'}H(K)$, arbitrarily close to H.

Thus we have obtained a "product space characterization" of achievable CR rates, namely that H is achievable iff for every $\delta > 0$ and sufficiently large n there exists a function $K = K(X^n)$ satisfying (4.18), (4.19). This can be easily single–letterized, using results available in the literature. To this, notice that in the above product space characterization, (4.18), (4.19) may be replaced by

$$\left|\frac{1}{n}H(X^n|K) - \big(H(X) - H\big)\right| < \delta, \frac{1}{n}H(Y_i^n|K) \le R_i - H + H(Y_i) + 2\delta. \tag{4.20}$$

Now, by [16], p. 352, an $(r+1)$–tuple $\tilde{R}_0, \tilde{R}_1, \dots, \tilde{R}_r$ has the property that for every $\delta > 0$ and sufficiently large n there exists a function $f(X^n)$ satisfying

$$\left|\frac{1}{n}H\big(x^n|f(X^n)\big) - \tilde{R}_0\right| < \delta, \quad \frac{1}{n}H\big(Y_i^n|f(X^n)\big) \le \tilde{R}_i + \delta$$

iff there exists a RV U with $U \multimap X \multimap Y_1, \dots, Y_r$ such that

$$H(X|U) = \tilde{R}_0, \quad H(Y_i|U) \le \tilde{R}_i.$$

Substituting here $\tilde{R}_0 = H(X) - H$, $\tilde{R}_i = R_i - H + H(Y_i)$, we get

$$I(U \wedge X) = H, \ I(U \wedge X) - I(U \wedge Y_i) \le R_i,$$

and this completes the proof for the no randomization case (up to the routine range constraint).

If randomization is permitted, we replace X by XM and proceed as in the proof of Theorem 4.1.

Theorem 4.3. *For Model (ii) described in Section 2, the CR capacity equals*

$$C_2(R) = \max_X \big[I(X \wedge Y) + \min\big(R, H(Y|X)\big)\big] \tag{4.21}$$

if randomization at $\mathcal{Y}$ is not permitted, and

$$\tilde{C}_2(R) = C(W) + R \tag{4.22}$$

if Terminal $\mathcal{Y}$ is allowed to randomize. In (4.21), the maximum is taken for random inputs X to the DMC $\{W\}$ given in the model, Y denoting the corresponding output. R is the capacity of the backward noiseless channel in the model, and $C(W)$ is the capacity of $\{W\}$. Moreover, the CR capacity of the variant of Model (ii) where the backward channel is replaced by a DMC, is still given by (4.21) resp. (4.22), with R replaced by the capacity of that DMC.

Remarks: (i) Comparing (4.21) and (4.22) shows that if R is not larger than $H(Y|X)$ for a capacity achieving X then $\tilde{C}_2(R) = C_2(R)$. Thus, similarly to Model (i), randomization helps only when R is "large". For many DMC's, the maximum of $H(Y)$ is attained for a capacity achieving X. In those cases $C_2(R) = H(Y)$ for all R "small" in the above sense.

(ii) One possible strategy of Terminal $\mathcal{X}$ in Model (ii) is to use an i.i.d. sequence X^n as channel input, which leads to the situation of Model (i), with the roles of $\mathcal{X}$ and $\mathcal{Y}$ reversed. Comparing Theorems 4.1 and 4.3 shows that this reduction to Model (i) suffices to achieve CR capacity for Model (ii) when the max in (4.21) is attained for some X with $H(Y|X) \leq R$, but not otherwise. (Namely, in that case, iq.'s (4.21) and (4.1), the lattes with X and Y reversed, give $C_2(R) = C_1(R) = H(Y)$.)

Proof: (a) Direct part.

If Terminal $\mathcal{Y}$ can randomize (recall that Terminal $\mathcal{X}$ always can in this model), a CR rate as in (4.22) can be attained in a trivial way: For large block–length n, Terminal $\mathcal{X}$ generates and transmits to $\mathcal{Y}$ a RV uniformly distributed on a set of size $\exp\big[(C(W) - \delta)\big]$; $\mathcal{Y}$ can decode it with small probability of error. Terminal $\mathcal{Y}$, in turn, generates a RV uniformly distributed on a set of size $\exp nR$, and transmits it to $\mathcal{X}$.

If $\mathcal{Y}$ can not randomize, a CR rate as in (4.21) can be attained as follows. For large n, take (X, Y) almost attaining the maximum in (4.21) such that P_X is a possible type for block–length n. Terminal $\mathcal{X}$ generates a RV M uniformly distributed on a set of size $\exp\big[n\big(I(X \wedge Y) - \delta\big)\big]$ and transmits it to $\mathcal{Y}$ using a code of fixed composition P_X. $\mathcal{Y}$ sends nothing back until he has received all n outputs. Then $\mathcal{Y}$ decodes M. Terminal $\mathcal{Y}$ can decode with small probability of error, and he gets access to additional randomness from the channel output. Namely, he can numerate the words in each of his decoding sets from 1 to $\exp\big[nH(Y|X) + o(n)\big]$, then the RV Z equal to the number assigned to the observed output sequence will be almost independent of M and have entropy $\big[nH(Y|X) + o(n)\big]$. If $R > H(Y|X)$, this Z can be transmitted back to $\mathcal{X}$, and if $R < H(Y|X)$ then a suitable function of Z of entropy $nR + o(n)$ can be transmitted back.

(b) Converse part.

Let (K,L) be a permissible pair for block–length n, thus $K=K(M,g_1,\dots,g_n)$, $L=L(Y^n)$, with $g=(g_1,\dots,g_n)$ satisfying (2.6). Supposing that (K,L) satisfies the condition in Definition 2.1 we decompose $H(L|M)$ in analogy to (4.5), replacing $f(X^n)$ there by $g=(g_1,\dots,g_n)$. Bounding as there we obtain

$$H(L|M)=I(L\wedge g|M)+H(L|M,g)\le nR+\varepsilon nc+1. \tag{4.23}$$

Further, (2.7) and the memoryless character of the DMC $\{W\}$ imply

$$\begin{aligned} I(L\wedge M) &\le I(M\wedge Y^n)=\sum_{i=1}^n I(M\wedge Y_i|Y^{i-1})\\ &\le \sum_{i=1}^n I(X_i\wedge Y_i|Y^{i-1})\le \sum_{i=1}^n I(X_i\wedge Y_i)\le nI(X_J\wedge Y_J) \end{aligned} \tag{4.24}$$

where J is an auxiliary RV uniformly distributed on $\{1,\dots,n\}$, independent of (M,Y^n). On the other hand,

$$H(L)\le H(Y^n)\le \sum_{i=1}^n H(Y_i)\le H(Y_J).$$

Combining this with (4.23) – (4.24) we get that

$$\begin{aligned} \frac{1}{n}H(L) &\le \min\bigl[I(X_J\wedge Y_J)+R+\varepsilon c+\frac{1}{n},H(Y_J)\bigr]\\ &\le I(X_J\wedge Y_J)+\min\bigl[R,H(Y_J|X_J)\bigr]+\varepsilon c+\frac{1}{n}. \end{aligned}$$

As Y_J is the channel output for input X_J, this completes the converse proof also for the no randomization case. When the backward channel is not noiseless but a DMC, then denoting its input and output by T^n and Z^n, the only difference will be that g in (4.23) has to be replaced by Z^n.

Then the first term will be bounded as

$$I(L\wedge Z^n|M)\le I(T^n\wedge Z^n|M)\le nC,$$

where C is the capacity of the backward channel.

Theorem 4.4. *For Model (iii) described in Section 2, the CR capacity without randomization is equal to*

$$\begin{aligned} C_3(R_1,R_2)=\max_{U,V}\bigl[I(U\wedge X)+I(V\wedge Y|U)\mid I(U\wedge X)-I(U\wedge Y)\le R_1,\\ I(V\wedge Y|U)-I(V\wedge X|U)\le R_2\bigr] \end{aligned} \tag{4.25}$$

where the maximization is for RV's U and V satisfying the Markov conditions

$$U \multimap X \multimap Y, \qquad X\multimap YU\multimap V. \tag{4.26}$$

Moreover, the range sizes of U and V can be bounded by $|\mathcal{X}|+2$ and $|\mathcal{Y}|$, resp.

Remark: It is reasuring to check that (4.25) reduces to the expected simple results when either $R_1 \geq H(X|Y)$ or $R_2 \geq H(Y|X)$. In the first case $U = X$ is a permissible choice, then the Markov condition for V becomes void, and it follows that

$$C_3(R_1, R_2) = H(X) + \min\bigl(R_2, H(Y|X)\bigr) \quad \text{if} \quad R_1 \geq H(X|Y). \tag{4.27}$$

In the second case $V = Y$ is a permissible choice, which leads to

$$I(U \wedge X) + I(V \wedge Y|U) = I(U \wedge X) + H(Y|U) = I(U \wedge X) - I(U \wedge Y) + H(Y).$$

Hence

$$C_3(R_1, R_2) = H(Y) + \min\bigl(R_1, H(X|Y)\bigr) \quad \text{if} \quad R_2 \geq H(Y|X). \tag{4.28}$$

Proof: (a) Converse part.

Let (K, L) be a permissible pair for Model (iii) without randomization, i.e., $K = K(X^n, g)$, $L = L(Y^n, f)$, $f = f(X^n)$, $g = g(Y^n, f)$, where f and g satisfy the rate constraints (2.1), (2.8). Suppose that (K, L) satisfy the conditions (2.3), (2.4) of Definition 2.1.

Our key tool is the identity (4.3), which will be applied twice. First we get

$$nR_1 \geq H(f) = I(f \wedge X^n) \geq I(f \wedge X^n) - I(f \wedge Y^n) = n\bigl(I(U \wedge X) - I(U \wedge Y)\bigr), \tag{4.29}$$

with

$$X = X_J, Y = Y_J,\ U = fX_1 \ldots X_{J-1}Y_{J+1} \ldots Y_n J \tag{4.30}$$

(where we proceed as in the derivation of eq. (4.7), the role of K there now played by f).

Notice now that just as $H(K|Y^n)$ was bounded in eq. (4.5), we have the bound

$$H(L|X^n) \leq nR_2 + \varepsilon cn + 1. \tag{4.31}$$

Applying the identity (4.3) again, we get

$$\begin{aligned} -H(L|X^n) &= -H(L|X^n, f) \\ &= I(L \wedge X^n|f) - H(L|f) \\ &= I(L \wedge X^n|f) - I(L \wedge Y^n|f) \\ &= n\bigl(I(L \wedge X|U) - I(L \wedge Y|U)\bigr) \end{aligned} \tag{4.32}$$

where X, Y, U are (luckily) the same as in (4.30).

By (4.31) (with sufficiently small ε) and (4.32) we have for any fixed $\delta > 0$

$$R_2 \geq I(L \wedge Y|U) - I(L \wedge X|U) - \delta. \tag{4.33}$$

Finally we can write

$$
\begin{aligned}
I(L \wedge X^n) &= \sum_{i=1}^{n} I(L \wedge X_i | X_1 \dots X_{i-1}) \\
&= \sum_{i=1}^{n} I(LX_1 \dots X_{i-1} \wedge X_i) \le nI(LU \wedge X).
\end{aligned} \tag{4.34}
$$

Combining (4.32) and (4.34) gives

$$
\begin{aligned}
H(L) &= I(L \wedge X^n) + H(L|X^n) \\
&\le n\big[I(LU \wedge X) + I(L \wedge Y|U) - I(L \wedge X|U)\big] \\
&= n\big[I(U \wedge X) + I(L \wedge Y|U)\big].
\end{aligned} \tag{4.35}
$$

Replacing L with V, we have thus proved that achievable CR–rates are bounded above by an expression as in (4.25) (the Markov conditions (4.26) are easily verified), perhaps with R_2 replaced by $R_2 + \delta$; the latter is inconsequential, by continuity.

(b) Direct part.

As in the proof of Theorem 4.1, it suffices to prove that $I(X \wedge U) + I(V \wedge Y|U)$ is an achievable CR rate whenever U and V satisfy (in addition to (4.26)) the inequalities in (4.25) with strict inequality. The form of eq. (4.25) suggests that in the first round CR of rate $I(U \wedge X)$ ought to be generated, and in the second round additional CR of rate $I(V \wedge Y|U)$.

We use the same construction as in the proof of Theorem 4.1. First we generate $\{\mathbf{u}_{ij}, 1 \le i \le N_1, 1 \le j \le N_2\}$ and associate with them functions $K_1(\mathbf{x})$ and $f(\mathbf{x})$ and $L_1(\mathbf{y}, i)$ as there (we write K_1 and L_1 rather than K and L, for now these functions will represent only the first part of the CR).

Then, by the proof of Theorem 4.1, for every pair $(\mathbf{x}, \mathbf{y})$ not in the exceptional set $A^c \cup B$ of arbitrarily small P^n_{XY}–probability, $K_1(\mathbf{x}) = L_1\big(\mathbf{y}, f(\mathbf{x})\big)$.

Next, to each $\mathbf{u}_{ij}$ as above, we generate at random $\exp\big[n\big(I(V \wedge Y|U) + \delta\big)\big]$ sequences $\mathbf{v} \in V^n$ of joint type with $\mathbf{u}_{ij}$ equal to P_{UV}, denoted as $\mathbf{v}^{(ij)}_{k\ell}$, $1 \le k \le M_1$, $1 \le \ell \le M_2$, where

$$
M_1 = \exp\big[n\big(I(V \wedge Y|U) - I(V \wedge X|U) + 3\delta\big)\big], M_2 = \exp\big[n\big(I(V \wedge X|U) - 2\delta\big)\big]. \tag{4.36}
$$

Then for every $\mathbf{y} \in \mathcal{Y}^n$ jointly UY–typical with $\mathbf{u}_{ij}$, the probability that neither $\mathbf{v}^{(ij)}_{k\ell}$ is jointly UYV–typical with $(\mathbf{u}_{ij}, \mathbf{y})$ is doubly exponentially small. Hence with probability close to 1, to every jointly typical pair $(\mathbf{u}_{ij}, \mathbf{y})$ there is a $\mathbf{v}^{(ij)}_{k\ell}$ such that $(\mathbf{u}_{ij}, \mathbf{y}, \mathbf{v}^{ij}_{k\ell})$ is jointly typical; we denote by $L(\mathbf{y}, \mathbf{u}_{ij})$ such a $\mathbf{v}^{(ij)}_{k\ell}$ (either one if there are several). Then for each $\mathbf{y}$ and $1 \le i \le N_1$ we take for $\mathbf{u}_{ij} = L_1(\mathbf{y}, i)$ the unique $\mathbf{u}_{ij}$ with the given first index i which is jointly typical with $\mathbf{y}$, or a constant if no or several such $\mathbf{u}_{ij}$ exist, and define $L_2(\mathbf{y}, i)$ as the $\mathbf{v}_{k\ell}^{(ij)}$ selected to this $\mathbf{u}_{ij}$ and $\mathbf{y}$:

$$L_2(\mathbf{y}, i) = L\big(\mathbf{y}, L_1(\mathbf{y}, i)\big) = \mathbf{v}_{k\ell}^{(ij)}. \tag{4.37}$$

Moreover, we define $g(\mathbf{y}, i)$ to equal the first index k of $\mathbf{v}_{k\ell}^{(ij)}$ in (4.37). Finally, for $\mathbf{x} \in \mathcal{X}^n$ and $1 \le k \le M_1$ we define $K_2(\mathbf{x}, k)$ as the unique $\mathbf{v}_{k\ell}^{(ij)}$ jointly typical with $(\mathbf{u}_{ij}, \mathbf{x})$, where $\mathbf{u}_{ij} = K_1(\mathbf{x})$, or set $K_2(\mathcal{X}, k) = \text{const}$ if no or several such $\mathbf{v}$ exist.

Then, by (4.36), g satisfies the rate constraint (2.8) if δ is sufficiently small. It is also clear that

$$\begin{aligned} K &= \big(K_1(X^n), K_2\big(X^n, g(Y^n, f(X^n))\big)\big) \\ L &= \big(L_1\big(Y^n, f(X^n)\big), L_2\big(Y^n, f(X^n)\big)\big) \end{aligned}$$

represent a permissible pair for Model (iii), satisfying (2.4), and one shows as in the proof of Theorem 4.1 that

$$\frac{1}{n} H(L) \ge I(U \wedge X) + I(V \wedge Y|U) - \delta.$$

It remains only to show that the condition (2.3), i.e., $Pr\{K = L\} > 1 - \varepsilon$ is also satisfied, at least with large probability with respect to the random selections. $Pr\{K_1 = L_1\} > 1 - \varepsilon$ has already been demonstrated in the proof of Theorem 3.1. The remaining part $Pr\{K_2 = L_2\} > 1 - \varepsilon$ can be proved similarly, though with a little more work.

5 Common randomness, identification, and transmission for arbitrarily varying channels

Recall the definition of an AVC in Section 2 by a class $\mathcal{W} = \{W(\cdot|\cdot, s), s \in \mathcal{S}\}$ of channels $W(\cdot|\cdot, s) \colon \mathcal{X} \to \mathcal{Y}$. There also CR capacities, ID capacities, and transmission capacities have been defined for various models involving an AVC. We present now our results.

Model (a): AVC without feedback and any other side information.

First we recall some well-known results for transmission capacities, cf. [16].

A random code $(\mathcal{C}_1, \dots, \mathcal{C}_M, Q)$ is defined by deterministic codes $\mathcal{C}_1, \dots, \mathcal{C}_M$ of the same block-length n and a PD Q on $\{1, \dots, M\}$, with the understanding that $\mathcal{C}_i$ will be used with probability $Q(i)$. The error criterion is that the maximum or the average (for k) of $\sum_{i=1}^{M} Q(i) e_k(i, \mathbf{s})$ be small for every $\mathbf{s} \in \mathcal{S}^n$, where $e_k(i, \mathbf{s})$ denotes the probability of not decoding correctly the message k when the code $\mathcal{C}_i$ is used and the state sequence is $\mathbf{s}$. Both criteria lead to the same random code capacity C_R. Notice that random codes can be used for transmission only if sender and receiver have acces to CR, the outcome of a random experiment with distribution Q.

It was shown in [8] that

$$C_R = \max_P \min_{W \in \overline{\mathcal{W}}} I(P,W) = \min_{W \in \overline{\mathcal{W}}} C(W). \tag{5.1}$$

Here $I(P,W)$ denotes the mutual information of input and output RV's with joint distribution $P(x)W(y|x)$, $C(W) = \max_P I(P,W)$ is the Shannon capacity of the channel W, and $\overline{\mathcal{W}}$ denotes the convex hull of $\mathcal{W}$.

By an elimination technique – based on an idea called now "derandomization" in computer science – it was shown in [1] that C_R can be attained by random codes $(\mathcal{C}_1, \ldots, \mathcal{C}_M, Q)$ with M not larger than the square of the block-length n and with uniform Q. As a consequence, the capacity for deterministic codes and the average probability of error criterion, denoted by $\overline{C}$, satisfies

$$\overline{C} = C_R \qquad \text{if} \qquad \overline{C} > 0. \tag{5.2}$$

Random codes should be distinguished from codes with randomized encoding, which do not need CR, the decoding being deterministic. It was also shown in [1] that with randomized encoding, both the maximum and average error criteria lead to the same capacity, and

$$\text{capacity under randomized encoding} = \overline{C}. \tag{5.3}$$

We note for later reference that (5.2) and (5.3) remain valid also for AVC's with noiseless feedback, if $\overline{C}$ is replaced by $\overline{C}_f$, the average error capacity for deterministic codes with feedback.

A necessary and sufficient condition for $\overline{C} > 0$, given in [1], is that for some n there exist PD's Q_1, Q_2 on $\mathcal{X}^n$ and disjoint subsets D_1, D_2 of $\mathcal{Y}^n$ such that

$$\min_{\mathbf{s} \in \mathcal{S}^n} \sum_{\mathbf{x} \in \mathcal{X}^n} Q_i(\mathbf{x}) W^n(D_i|\mathbf{x}, \mathbf{s}) > \frac{1}{2}, \quad i = 1, 2. \tag{5.4}$$

A single-letter necessary and sufficient condition for $\overline{C} > 0$ was given in [10]: $\overline{C} > 0$ iff $\mathcal{W}$ is not symmetrizable, where symmetrizability of $\mathcal{W}$ means the existence of a channel $U \colon \mathcal{X} \to \mathcal{S}$ such that

$$\sum_{s \in \mathcal{S}} U(s|x')W(y|x,s) = \sum_{s \in \mathcal{S}} U(s|x)W(y|x',s) \tag{5.5}$$

for every x, x' in $\mathcal{X}$ and y in $\mathcal{Y}$.

With these results and Theorems 3.2 and 3.3 the following theorem is readily obtained.

Theorem 5.1. *For an AVC without feedback, both ID capacity and CR capacity with sender permitted to randomize are equal to average error transmission capacity for deterministic codes:*

$$C_{ID} = C_{CR} = \overline{C}. \tag{5.6}$$

Their common value equals C_R given by (5.1) if $\mathcal{W}$ is not symmetrizable, and 0 otherwise.

Proof: (i) $C_{CR} = \overline{C}$: the non–trivial part $C_{CR} \le \overline{C}$ follows from Theorem 3.2 and (5.3). Indeed, a permissible pair (K, L) that satisfies (3.10) for every choice of $\mathbf{s} \in \mathcal{S}^n$ gives rise to a code with randomized encoder of rate $\frac{1}{n} \log |\mathcal{M}|$ and average probability of error $< 2\varepsilon$.

(ii) $C_{ID} \ge C_{CR} = \overline{C}$: In the non–trivial case $\overline{C} > 0$, this is a consequence of Theorem 3.3 and of the fact that $\overline{C}$ equals the maximum error capacity for randomized encoding.

(iii) $C_{ID} \le \overline{C}$: Notice that $C_{ID} > 0$ implies $\overline{C} > 0$, because Q_1 and Q_2 as in (2.12) with $D_1' = D_1 \setminus D_2$, $D_2' = D_2 \setminus D_1$ satisfy (5.4) if $\varepsilon < 1/4$ in (2.12). Thus, on account of (5.2), it suffices to show that $C_{ID} \le C_R$. It follows from (2.12) that an (N, n, ε) ID code for the AVC is, for each $W \in \overline{\mathcal{W}}$, also an (N, n, ε) code for the DMC $\{W\}$. Since the ID capacity of a DMC equals its transmission capacity, this and (5.1) imply the claimed inequality.

Model (b): AVC with noiseless (passive) feedback.

Let C_{CRF} and C_{CRf} denote the CR capacity and C_{IDF} and C_{IDf} the identification capacity of the AVC with noiseless (passive) feedback, according as Terminal $\mathcal{X}$ is permitted to randomize or not. As now $\mathcal{X}$ knows everything that $\mathcal{Y}$ does, C_{CRF} equals the limit as $n \to \infty$ of the maximum, for all protocols as described in the passage containing eq. (2.10), of

$$\frac{1}{n} \min_{\mathbf{s} \in \mathcal{S}^n} H(Y^n). \tag{5.7}$$

C_{CRf} is obtained similarly, with the maximum taken for the deterministic protocols (formally, with M =const in (2.10)).

Theorem 5.2. *For an AVC with noiseless feedback,*

$$C_{CRF} = \max_P \min_{W \in \mathcal{W}} H(PW), \tag{5.8}$$

$$C_{CRf} = \max_P \min_{W \in \mathcal{W}} H(W|P) \quad \textit{if } C_{CRf} > 0, \tag{5.9}$$

$$C_{CRf} > 0 \quad \textit{iff} \ \min_{W \in \mathcal{W}} H(W(\cdot|x)) > 0 \quad \textit{for some } x \in \mathcal{X}. \tag{5.10}$$

Here $H(PW)$ and $H(W|P)$ denote the entropy $H(Y)$ and conditional entropy $H(Y|X)$ for RV's X, Y with joint distribution $P(x)W(y|x)$.

Remark: These single-letter characterizations have been obtained independently also by Ning Cai [9].

Proof: (i) For a protocol that disregards the feedback information and selects i.i.d. inputs $X_1, \dots, X_n$ with distribution P, the quantity (5.7) becomes

$\min_{W\in\mathcal{W}} H(PW)$. This proves that the right hand side of (5.8) is an achievable CR rate. For the converse, we prove by induction that for any given protocol,

$$\min_{\mathbf{s}\in\mathcal{S}^k} H(Y^k) \le k \max_P \min_{W\in\mathcal{W}} H(PW), \tag{5.11}$$

for $k = 1,\ldots,n$. Indeed, (5.11) clearly holds for $k=1$. Now, if $\min_{\mathbf{s}\in\mathcal{S}^k} H(Y^k)$ is attained for $\tilde{\mathbf{s}} = \tilde{s}_1 \ldots \tilde{s}_k$, let $\widetilde{P}$ denote the distribution of X_{k+1} when (the given protocol is used and) $s_i = \tilde{s}_i$, $i = 1,\ldots,k$. Then

$$\begin{aligned}\min_{\mathbf{s}\in\mathcal{S}^{k+1}} H(Y^{k+1}) &\le \min_{\mathbf{s}\in\mathcal{S}^{k+1}} (H(Y^k) + H(Y_{k+1})) \\ &\le \min_{\mathbf{s}\in\mathcal{S}^k} H(Y^k) + \min_{W\in\mathcal{W}} H(\widetilde{P}W).\end{aligned} \tag{5.12}$$

Hence (5.11) holds for $(k+1)$ if it does for k.

(ii) For a deterministic protocol, when X_i is a function of Y^{i-1}, we have

$$H(Y^n) = \sum_{i=1}^n H(Y_i|Y^{i-1}) = \sum_{i=1}^n H(Y_i|Y^{i-1}X_i) = \sum_{i=1}^n H(Y_i|X_i). \tag{5.13}$$

Using (5.13), an induction as above shows that the right hand side of (5.9) is an upper bound to (5.7) for any deterministic protocol.

Now, let P^* be the PD achieving the maximum in (5.9). Supposing $C_{CRf} > 0$, it follows from Theorem 3.2 that to any $\varepsilon > 0$ there exists $k = k(\varepsilon)$, a protocol of block-length k, and a mapping f of $\mathcal{Y}^k$ into $\mathcal{X}$, such that the distribution of $f(Y^k)$ differs by less than ε from P^*, in variation distance, no matter what is the state sequence $\mathbf{s}\in\mathcal{S}^k$. We extend this protocol to block-length n, by letting $X_i = f(Y^k)$ for $i = k+1,\ldots,n$. Then, by (5.13), the limit of (5.7) as $n\to\infty$ will be arbitrarily close to the right hand side of (5.9), if $\varepsilon > 0$ is sufficiently small.

(iii) Obviously, the condition in (5.10) is sufficient for $C_{CRf} > 0$. To prove its necessity, suppose indirectly that to each $x\in\mathcal{X}$ there is an $s = s(x)$ such that $W(\cdot|x,s)$ is the point mass at some $y = y(x)$. Given any deterministic protocol, consider $\mathbf{x}\in\mathcal{X}^n$, $\mathbf{s}\in\mathcal{S}^n$, $\mathbf{y}\in\mathcal{Y}^n$ defined recursively such that $s_i = s(x_i)$, $y = y(x_i)$, and x_{i+1} is the input symbol that the given protocol specifies when the past output sequence is $y_1\ldots y_i$. For this particular state sequence $\mathbf{s}$, the given protocol leads to a unique output sequence $\mathbf{y}$, proving that quantity (5.7) is equal to 0 for every deterministic protocol, hence $C_{CRf} = 0$.

Our result on the CR capacity leads to a noticable conclusion for the classical transmission problem.

Theorem 5.3. *The average error capacity $\overline{C}_f$ of an AVC with noiseless feedback, for deterministic coding, is always equal to C_R given by (5.1). Further,*

$$C_{IDF} = C_{CRF}, C_{IDf} = C_{CRf} \quad \textit{if} \quad C_R > 0 \tag{5.14}$$

$$C_{IDF} = C_{IDf} = 0 \quad \textit{if} \quad C_R = 0. \tag{5.15}$$

Proof: (i) The random code $(\mathcal{C}_1, \ldots, \mathcal{C}_M, Q)$ in the paragraph containing eq. (5.2) can be used for transmission if $\mathcal{X}$ and $\mathcal{Y}$ have access to $2\log n$ bits of robust UCR, i.e., to RV's K, L satisfying (3.10) for every $\mathbf{s} \in \mathcal{S}^n$ with $|\mathcal{M}| = n^2$. Since $C_R > 0$ implies $C_{CRF} > 0$, cf. (5.1), (5.8), such UCR may be generated using a protocol of block-length $n' = c\log n$, by Theorem 3.2. This proves that C_R is an achievable transmission rate, at least if randomization is permitted (randomization may be needed in the CR-generating protocol of negligible block-length $n' = c\log n$, whose outcome will identify the $\mathcal{C}_i$ actually used). The proof is completed by reference to the feedback versions of (5.2) and (5.6).

(ii) If $\overline{C}_f = C_R > 0$, the inequalities $C_{IDF} \geq C_{CRF}$, $C_{IDf} \geq C_{CRf}$ are proved analogously to the proof of Theorem 5.1, part (ii). The reversed inequalities follow by the method of [5], where the ID capacity of a DMC with feedback has been determined. If $C_R = 0$ then $C(W) = 0$ for some $W \in \overline{\mathcal{W}}$. Then the feedback ID capacity of the DMC $\{W\}$ is 0 by [5], and (5.15) follows.

Model (c): Strongly arbitrarily varying channel (SAVC).

It is assumed here that the jammer can make his choice of $\mathbf{s} \in \mathcal{S}^n$ depend on the sent $\mathbf{x} \in \mathcal{X}^n$. Formally, the parameter determining the statistics is now an arbitrary mapping from $\mathcal{X}^n$ to $\mathcal{S}^n$.

Since the number of such mappings is doubly exponential in n, the hypothesis of Theorem 3.2 is still satisfied. The criterion (2.11) for an (N, n, ε) ID code becomes

$$\sum_{\mathbf{x}\in\mathcal{X}^n} Q_j(\mathbf{x}) \max_{\mathbf{s}} W^n(D_j^c|\mathbf{x},\mathbf{s}) \leq \varepsilon, \quad \sum_{\mathbf{x}\in\mathcal{X}^n} Q_k(\mathbf{x}) \max_{\mathbf{s}} W^n(D_j|\mathbf{x},\mathbf{s}) \leq \varepsilon. \tag{5.16}$$

The first inequalities here (with disjoint sets D_j) represent the maximum probability of error criterion for transmission codes with randomized encoding.

Any (N, n, ε) transmission code with randomized encoding gives rise to a deterministic (N, n, ε) code, with codewords $\mathbf{x}_j = \arg\min_{\mathbf{x}}(\max_{\mathbf{s}} W^n(D_j^c|\mathbf{x},\mathbf{s}))$. Hence the maximum error capacity of a SAVC for deterministic and randomized encoding is the same. It is also (well known and) easy to see that this capacity coincides with the average error capacity for deterministic codes, and it equals the maximum error capacity for deterministic codes of the AVC defined by the same $\mathcal{W}$. We shall denote this capacity by $\overline{\overline{C}}$. As shown in [14], $\overline{\overline{C}} > 0$ iff there exists x and x' in $\mathcal{X}$ with $\mathcal{T}(x) \cap \mathcal{T}(x') = \emptyset$ where $\mathcal{T}(x)$ denotes the convex hull of the set of PD's $W(\cdot|x, s)$, $s \in \mathcal{S}$.

The row-convex hull $\overline{\overline{\mathcal{W}}}$ of $\mathcal{W}$ is the set of all channels $W \colon \mathcal{X} \to \mathcal{Y}$ such that $W(\cdot|x) \in \mathcal{T}(x)$, $x \in \mathcal{X}$. Write

$$D = \min_{W\in\overline{\overline{\mathcal{W}}}} C(W). \tag{5.17}$$

Theorem 5.4. *For a SAVC, the CR capacity C_{CR}^s and ID capacity C_{ID}^s (with $\mathcal{X}$ permitted* *to randomize) satisfy*

$$\bar{\bar{C}} \le C^s_{CR} \le C^s_{ID} \le D \tag{5.18}$$

$$C^s_{ID} > 0 \quad \textit{iff} \quad \bar{\bar{C}} > 0. \tag{5.19}$$

Remark: Under not too restrictive hypotheses, $\bar{\bar{C}} = D$, cf. [2] for $\mathcal{W}$ satisfying $\mathcal{T}(x) \cap \mathcal{T}(x') = \emptyset$ whenever $x \neq x'$, and [13] under a weaker hypothesis; there are, however, examples of $0 < \bar{\bar{C}} < D$. For $\mathcal{W}$ with $\bar{\bar{C}} = D$, Theorem 5.4 gives a conclusive result, but we do not know whether $C^s_{ID} = C^s_{CR}$ and/or $C^s_{CR} = \bar{\bar{C}}$ hold for every SAVC. C^s_{CR} always equals the average error capacity for randomized encoding, but it appears unknown whether the latter can ever be larger than $\bar{\bar{C}}$.

Proof: The first inequality of (5.18) is obvious, and if $\bar{\bar{C}} > 0$, the second inequality follows from Theorem 3.3. It remains to prove that $C^s_{ID} \le D$ and that $\bar{\bar{C}} = 0$ implies $C^s_{CR} = C^s_{ID} = 0$.

Consider an auxiliary model where at each instant i the state s_i may depend on x_i but not on the other x'_js. Formally, this is an AVC model, with state set $\mathcal{S}^*$ consisting of all mappings $s^* \colon \mathcal{X} \to \mathcal{S}$, defined by the set of channels

$$\mathcal{W}^* = \{W^*(\cdot|\cdot, s^*), s^* \in \mathcal{S}^*\}, \quad W^*(\cdot|x, s^*) = W(\cdot|x, s^*(x)). \tag{5.20}$$

Clearly, the CR and ID capacities of this AVC are upper bounds to C^s_{CR} and C^s_{ID}. Thus, on account of Theorem 5.1, it suffices to show that (i) the random code capacity of the AVC defined by (5.20) equals D and (ii) $\mathcal{W}^*$ is symmetrizable if $\bar{\bar{C}} = 0$.

(i) is obvious from (5.1) and (5.17) since $\overline{\mathcal{W}^*} \stackrel{=}{=}\!\!\!\!\to \mathcal{W}$.

To prove (ii), use either the Strong Separation Lemma of [1] or, alternatively, recall that $\bar{\bar{C}} = 0$ iff $\mathcal{T}(x) \cap \mathcal{T}(x')$ is never empty, i.e., for suitable PD's $U(\cdot|x, x')$ on $\mathcal{S}$,

$$\sum_{s \in \mathcal{S}} U(s|x, x')W(y|x, s) = \sum_{s \in \mathcal{S}} U(s|x', x)W(y|x', s) \tag{5.21}$$

for every x, x' and y. (5.21) means that $\mathcal{W}^*$ satisfies (5.5), with $U^* \colon \mathcal{X} \to \mathcal{S}^*$ defined by

$$U^*(s^*|x) = \prod_{\tilde{x} \in \mathcal{X}} U(s^*(\tilde{x})|\tilde{x}, x). \tag{5.22}$$

Remark: Work relevant for problems concerning feedback with *noise* can be found in [20].

6 Appendix

Proof of Lemma 3.1: Choosing f at random as in the proof of Lemma 1.1, with $Z_i(v)$ as there, we have

$$P\big(f^{-1}(i) \cap E(P,d)\big) = \sum_{v \in E(P,d)} p(v) Z_i(v). \tag{A.1}$$

Chernoff bounding gives that for any $A \subset \mathcal{V}$

$$\begin{aligned}
&Pr\left\{\sum_{v\in A} p(v)Z_i(v) > \frac{1+\varepsilon}{k}P(A)\right\} \\
&= Pr\left\{\exp\left[\beta\sum_{v\in A} p(v)Z_i(v)\right] > \exp\left(\beta\frac{1+\varepsilon}{k}P(A)\right)\right\} \\
&\le E\left(\exp\left[\beta\sum_{v\in A} p(v)Z_i(v)\right]\right)\exp\left(-\beta\frac{1+\varepsilon}{k}P(A)\right) \\
&= \exp\left(-\beta\frac{1+\varepsilon}{k}P(A)\right)\prod_{v\in A}\left[1+\frac{1}{k}\big(\exp(\beta p(v))-1\big)\right]
\end{aligned} \tag{A.2}$$

where $\beta > 0$ is arbitrary, and similarly

$$\begin{aligned}
&Pr\left\{\sum_{v\in A} p(v)Z_i(v) < \frac{1-\varepsilon}{k}P(A)\right\} \\
&\le \exp\left(\beta\frac{1-\varepsilon}{k}P(A)\right)\prod_{v\in A}\left[1+\frac{1}{k}\big(\exp(-\beta p(v))-1\big)\right].
\end{aligned} \tag{A.3}$$

Apply (A.2) to $A = E(P,d)$ with $\beta = \varepsilon d$. Then for $v \in A = E(P,d)$ we have $\beta p(v) \le \varepsilon$, by (3.1), and therefore

$$\begin{aligned}
\exp\big(\beta p(v)\big) - 1 &= \sum_{j=1}^{\infty}\frac{\big(\beta p(v)\ell n2\big)^j}{j!} \\
&< \beta p(v)\left[1+\frac{1}{2}\sum_{j=1}^{\infty}(\varepsilon\ell n2)^j\right]\ell n2 = \beta p(v)(1+\varepsilon^*)\ell n2,
\end{aligned}$$

where $\varepsilon^* = \frac{\varepsilon\ell n2}{2(1-\varepsilon\ell n2)}$.

Using the inequality $1 + t\ell n2 \le \exp t$, it follows that the last product in (A.2) is upper bounded by

$$\exp\left[\sum_{v\in E(P,d)}\frac{1}{k}\beta p(v)(1+\varepsilon^*)\right] = \exp\left[\frac{\beta}{k}(1+\varepsilon^*)P\big(E(P,d)\big)\right].$$

Thus (A.2) gives, using the assumption (3.2) and recalling that $\beta = \varepsilon d$,

$$\begin{aligned} &Pr\left\{\sum_{v\in E(P,d)} p(v)Z_i(v) > \frac{1+\varepsilon}{k}P\big(E(P,d)\big)\right\} \\ &< \exp\left[-\frac{\beta}{k}(\varepsilon-\varepsilon^*)P\big(E(P,d)\big)\right] < \exp\left(-\frac{\varepsilon d(\varepsilon-\varepsilon^*)(1-\varepsilon)}{k}\right) \\ &< \exp\left(-\frac{\varepsilon^2}{3k}d\right); \end{aligned} \tag{A.4}$$

here, in the last step, we used that $(\varepsilon-\varepsilon^*)(1-\varepsilon) = \varepsilon\left(1-\frac{\ell n2}{2(1-\ell\varepsilon n2)}\right)(1-\varepsilon) > \frac{\varepsilon}{3}$, if $\varepsilon < 3-2\log e$, and that condition does hold by the assumption $\varepsilon \le \frac{1}{9}$. It follows from (A.3) in a similar but even simpler way (as $\exp\big(-\beta p(v)-1\big)$ can be bounded by $\beta p(v)\left(-1+\frac{1}{2}\varepsilon\ell n2\right)\ell n2)$ that the left hand side of (A.3) is also bounded by $\exp\left(-\frac{\varepsilon^2}{3k}d\right)$.

Recalling (A.1), we have thereby shown that the probability that (3.3) does not hold for a randomly chosen f is $< 2N\exp\left(-\frac{\varepsilon^2}{3k}d\right)$. Hence this probability is less than 1 if $k \le \frac{\varepsilon^2}{3\log(2N)}d$. This completes the proof of Lemma 3.1, because (3.4) is an immediate consequence of (3.3).

Completion of the proof of Theorem 3.1:

We have to show that the P–probability of the complementary set of (3.1) with $P = P(\cdot|\mathbf{s})$ defined by (2.9) is $\le \varepsilon$ if d is as in (3.7), with ξ given by (3.8). This probability can be written as

$$Pr\big\{P(X^n|\mathbf{s}) > \exp\big[-n(H_{\min}-\xi)\big]\big\} \tag{A.5}$$

where Pr denotes probability under $P(\cdot|\mathbf{s})$. Now, for every $t > 0$

$$\begin{aligned} &Pr\big\{P(X^n|\mathbf{s}) > \exp\big[-n(H_{\min}-\xi)\big]\big\} = Pr\big\{P^t(X^n|\mathbf{s}) > \exp\big[-nt(H_{\min}-\xi)\big]\big\} \\ &< \exp\big[nt(H_{\min}-\xi)\big]E\big(P^t(X^n|\mathbf{s})\big) \\ &= \exp\big[nt(H_{\min}-\xi)\big]\prod_{i=1}^{n}\sum_{x\in\mathcal{X}} P^{1+t}(x|s_i), \end{aligned} \tag{A.6}$$

where E denotes expectation under $P(\cdot|\mathbf{s})$. To bound the last product in (A.6), notice that for any PD $P = \big\{p(x)\big\}$ on $\mathcal{X}$

$$\begin{aligned} \sum_{x\in\mathcal{X}} p^{1+t}(x) &= \sum_{x\in\mathcal{X}} p(x)\sum_{j=0}^{\infty}\frac{\big(t\ell n\ p(x)\big)^j}{j!} \\ &\le 1 + t\sum_{x\in\mathcal{X}} p(x)\ell n\ p(x) + \frac{t^2}{2}\sum_{x\in\mathcal{X}} p(x)\big[\ell n\ p(x)\big]^2 \\ &= 1 - tH(P)\ell n\ 2 + \frac{t^2}{2}\sum_{x\in\mathcal{X}} p(x)\big[\ell n\ p(x)\big]^2. \end{aligned} \tag{A.7}$$

Calculus shows that the last sum in (A.7) is maximum when P is the uniform distribution on $\mathcal{X}$, providing $|\mathcal{X}| \geq 3$. Hence

$$\sum_{x\in\mathcal{X}} p^{1+t}(x) \leq 1 - tH(P)\ell n\ 2 + \frac{t^2}{2}(\ell n|\mathcal{X}|)^2 \leq \exp\big[-tH(P) + \frac{t^2}{2}(\log|\mathcal{X}|)^2\ell n\ 2\big], \quad \text{(A.8)}$$

with the understanding (as also in the rest of the proof) that $|\mathcal{X}|$ should be replaced by 3 if $|\mathcal{X}| = 2$.

As $H\big(P(\cdot|s_i)\big) \geq H_{\min}$ by definition, (A.6) and (A.8) give that the probability (A.5) is upper bounded by $\exp\big[-nt\xi + n\frac{t^2}{2}(\log|\mathcal{X}|)^2\ell n\ 2\big]$, for each $t > 0$. Setting $t = \xi/(\log|\mathcal{X}|)^2\ell n\ 2$, we get

$$Pr\big\{P(X^n|\mathbf{s}) > \exp\big[-n(H_{\min} - \xi)\big]\big\} < \exp\left[-n\frac{\xi^2}{2(\log|\mathcal{X}|)^2\ell n\ 2}\right]. \quad \text{(A.9)}$$

For ξ given by (3.8), the right hand side of (A.9) is equal to $\frac{\varepsilon}{3}$, establishing our claim.

Completion of the proof of Theorem 4.4:

Here we show that to any RV's U, V satisfying the Markov conditions (4.26) there exist $\tilde{U}, \tilde{V}$ satisfying the same conditions, with range sizes $|\tilde{\mathcal{U}}| \leq |\mathcal{X}| + 2$, $|\tilde{\mathcal{V}}| \leq |\mathcal{Y}|$ such that

$$I(\tilde{U} \wedge X) - I(\tilde{U} \wedge Y) = I(U \wedge X) - I(U \wedge Y) \quad \text{(A.10)}$$

$$I(\tilde{V} \wedge Y|\tilde{U}) - I(\tilde{V} \wedge X|\tilde{U}) \leq I(V \wedge Y|U) - I(V \wedge X|U) \quad \text{(A.11)}$$

$$I(\tilde{U} \wedge X) + I(\tilde{V} \wedge Y|\tilde{U}) \geq I(U \wedge X) + I(V \wedge Y|U). \quad \text{(A.12)}$$

(i) Given U, V satisfying (4.26), introduce an equivalence relation on $\mathcal{U}$ by letting $u_1 \sim u_2$ iff $P_{X|U=u_1} = P_{X|U=u_2}$. Our first claim is that U, V can be replaced by U', V' without changing the relevant mutual informations, such that no distinct elements of the range of U' are equivalent in the above sense.

Let $f(u)$ denote the equivalence class of u. Then clearly

$$P_{XY|U=u} = P_{XY|f(U)=f(u)} \quad \text{(A.13)}$$

hence

$$I(U \wedge X) = I\big(f(U) \wedge X\big),\ I(U \wedge Y) = I\big(f(U) \wedge Y\big).$$

This, in turn, implies that

$$I(V \wedge X|U) = I(UV \wedge X) - I(U \wedge X) = I\big(UVf(U) \wedge X\big) - I\big(f(U) \wedge X\big) = I\big(UV \wedge X|f(U)\big),$$

and similarly

$$I(V \wedge Y|U) = I\big(UV \wedge Y|f(U)\big).$$

Thus $U' = f(U)$, $V' = UV$ satisfy our claim, since $U' \multimap X \multimap Y$ is obvious from $U \multimap X \multimap Y$ and (A.13), and $X \multimap YU' \multimap V'$ follows as

$$P_{X|Y=y,f(U)=f(u),U=u,V=v} = P_{X|Y=y,U=u,V=v} = P_{X|Y=y,U=u} = P_{X|Y=y,f(U)=f(u)},$$

where the second equality holds by $X \multimap YU \multimap V$ and the third by (A.13).

(ii) By (i), it suffices to consider U, V (satisfying (4.26)) such that the PD's $P_u = P_{X|U=u}, u \in \mathcal{U}$ are all distinct. This will enable us to use the Support Lemma (see [19] or [16], p. 310) to reduce the range size of U. To this end, define the stochastic matrix valued function $F(P)$ for $P \in \{P_u, u \in \mathcal{U}\}$ by letting $\Pr\{V = v|Y = y, U = u\}$ be the (y, v) entropy of $F(P_u)$. Then extend $F(P)$ continuously but otherwise arbitrarily to the set $\mathcal{P}(\mathcal{X})$ of all PD's on $\mathcal{X}$. Now apply the Support Lemma to the following continuous functions on $\mathcal{P}(\mathcal{X})$:

$$
\begin{aligned}
&f_1(P) = H(X) - H(Y) - H(P) + H(PW), \quad \text{where } \ W = P_{Y|X} \\
&f_2(P) = H(X) - H(P) + I\big(PW, F(P)\big) \\
&f_3(P) = I\big(PW, F(P)\big) - I\big(P, WF(P)\big) \\
&f_i(P) = P(x_{j-3}), \ 4 \le j \le |\mathcal{X}| + 2, \quad \text{where } \ \mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}.
\end{aligned}
$$

It follows that there exist PD's $P_i \in \mathcal{P}(\mathcal{X})$, $i = 1, \ldots, |\mathcal{X}|+2$, and a PD $\{\alpha_1, \ldots, \alpha_{|\mathcal{X}|+2}\}$ on $\{1, \ldots, |\mathcal{X}| + 2\}$ such that

$$
\sum_{u \in \mathcal{U}} \Pr\{U = u\} f_j(P_u) = \sum_{i=1}^{|\mathcal{X}|+2} a_i f_j(P_i), \ j = 1, \ldots, |\mathcal{X}| + 2. \tag{A.14}
$$

The last $|\mathcal{X}| - 1$ identities in (A.14) mean that a RV $\tilde{U}$ with range $\tilde{\mathcal{U}} = \{1, \ldots, |\mathcal{X}| + 2\}$ and distribution $\{\alpha_1, \ldots, \alpha_{|\mathcal{X}|+2}\}$ exists such that $P_{X|\tilde{U}=i} = P_i$, $i = 1, \ldots, |\mathcal{X}|+2$. Letting this $\tilde{U}$ satisfy $\tilde{U} \multimap X \multimap Y$, the first identity in (A.14) gives (A.10). Further letting $\tilde{V}$ be such that $X \multimap Y\tilde{U} \multimap \tilde{V}$, $P_{\tilde{V}|Y,\tilde{U}=i} = F(P_i)$, the second and third identities in (A.14) mean that (A.11) and (A.12) hold with equality.

(iii) Finally, it remains to show that $\tilde{V}$ in (ii) can be replaced by some $\tilde{V}'$ with range size $\le |\mathcal{Y}|$ and $X \multimap Y\tilde{U} \multimap \tilde{V}'$ such that $\tilde{V}'$ still satisfies (A.11), (A.12). Now, by the range constraint result of Theorem 4.1, applied to RV's with joint distribution $P_{YX|\tilde{U}=u}$ in the role of X, Y, for each fixed $\tilde{u} \in \tilde{\mathcal{U}}$ there exists $V_{\tilde{u}}$ distributed on a set of size $\le |\mathcal{Y}|$ and conditionally independent of X on the conditions $Y = y$, $\tilde{U} = \tilde{u}$, for every $y \in \mathcal{Y}$, such that

$$
\begin{aligned}
&I(V_{\tilde{u}} \wedge Y|\tilde{U} = \tilde{u}) \ge I(\tilde{V} \wedge Y|\tilde{U} = \tilde{u}) \\
&I(V_{\tilde{u}} \wedge Y|\tilde{U} = \tilde{u}) - I(V_{\tilde{u}} \wedge X|\tilde{U} = \tilde{u}) \le I(\tilde{V} \wedge Y|\tilde{U} = \tilde{u}) - I(\tilde{V} \wedge X|\tilde{U} = \tilde{u}).
\end{aligned}
$$

But then we can define a RV $\tilde{V}'$ with $X \multimap YU \multimap \tilde{V}'$ such that $P_{\tilde{V}|Y=y,\tilde{\mathcal{U}}=\tilde{u}} = P_{V_{\tilde{u}}|Y=y,\tilde{\mathcal{U}}=\tilde{u}}$ for every $y \in \mathcal{Y}$, $\tilde{u} \in \tilde{\mathcal{U}}$. This $\tilde{V}'$ of range size $\le |\mathcal{Y}|$ will satisfy the last inequalities for every $\tilde{u} \in \tilde{\mathcal{U}}$, and hence also (A.11), (A.12), as required.

7 References

[1] Ahlswede, R., "Elimination of correlation in random codes for arbitrarily varying channels", Z. Wahrscheinlichkeitsth. Verw. Gebiete 33 (1978), 159-175.

[2] Ahlswede, R., "A method of coding and an application to arbitrarily varying channels", J. Combinatorics, Inform. and Systems Sci. 5 (1980), 1035.

[3] Ahlswede, R., "Coloring hypergraphs: a new approach to multiuser source coding", J. Comb. Inf. Syst. Sci. 1 (1979), 76-115 and 2 (1980), 220-268.

[4] Ahlswede, R. and Dueck, G., "Identification via channels", IEEE Trans. Inform. Theory 35 (1989), 15-29.

[5] Ahlswede, R. and Dueck, G., "Identification in the presence of feedback – a discovery of new capacity formulas", IEEE Trans. Inform. Theory 35 (1989), 30-36.

[6] Ahlswede, R. and Verboven, B., "On identification via multi-way channels with feedback", IEEE Trans. Inform. Theory 37 (1991), 1519-1526.

[7] Ahlswede, R. and Csiszár, I., "Common randomness in information theory and cryptography. Part 1: Secret sharing", IEEE Trans. Inform. Theory 39 (1993), 1121-1132.

[8] Blackwell, D., Breiman, L. and Thomasian, A.J., "The capacities of certain channel classes under random coding", Ann. Math. Statist. 31 (1960), 558-567.

[9] Cai, N., "Personal communication", (1995).

[10] Csiszár, I. and Narayan, P., "The capacity of the arbitrarily varying channel revisited: positivity, constraints", IEEE Trans. Inform. Theory 34 (1988), 181-193.

[11] Csiszár, I. and Narayan, P., "Capacity of the Gaussian arbitrarily varying channel", IEEE Trans. Inform. Theory 37 (1991), 18-26.

[12] Gács, P. and Körner, J., "Common information is far less than mutual information", Probl. Control and Inform. Theory 21 (1973), 149-162.

[13] Csiszár, I. and Körner, J., "On the capacity of the arbitrarily varying channels for maximum probability of error", Z. Wahrscheinlichkeitsth. Verw. Gebiete 57 (1981), 87-101.

[14] Kiefer, J. and Wolfowitz, J., "Channels with arbitrarily varying channel probability functions", Information and Control 5 (1962), 44-54.

[15] Slepian, D. and Wolf, J.K., "Noiseless coding of correlated information sources", IEEE Trans. Inform. Theory 19 (1973), 471-480.

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

[17] Ahlswede, R., Cai, N., and Zhang, Z., "On interactive communication", Preprint 93–066, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, IEEE Trans. on Inform. Theory, Vol. 43, No. 1 (1997), 22–37.

[18] Ahlswede, R. and Balakirsky, V.B., "Identification under random processes", Preprint 95–098, SFB 343, "Diskrete Strukturen in der Mathematik", Universität Bielefeld, Problemy peredachii informatsii (special issue devoted to M.S. Pinsker), Vol. 32, No. 1 (1996), 144–160.

[19] Salehi, M., "Cardinality bounds on auxiliary variables in multiple–user theory via the method of Ahlswede and Körner", Stanford Technical Report (1978).

[20] Ahlswede, R. and Zhang, Z., "New directions in the theory of identification via channels", Preprint 94–010, SFB 343 "Diskrete Strukturen in der Mathematik, Universität Bielefeld, IEEE Trans. on Inform. Theory, Vol. 41, No. 4 (1995), 1040–1050.

[21] Venkatesan, S. and Anantharam, V., "The common randomness capacity of independent discrete memoryless channels", Memorandum No. UCB/ERL M95/85 (Sep. 1995).