IDENTIFICATION PLUS TRANSMISSION OVER CHANNELS WITH PERFECT FEEDBACK

by

S. Venkatesan and V. Anantharam

Memorandum No. UCB/ERL M96/34

7 May 1996

IDENTIFICATION PLUS TRANSMISSION OVER CHANNELS WITH PERFECT FEEDBACK

by

S. Venkatesan and V. Anantharam

Memorandum No. UCB/ERL M96/34

7 May 1996

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Identification plus Transmission over Channels with Perfect Feedback *

S. Venkatesan† V. Anantharam‡§

Abstract

We study the following problem: a transmitter is connected to N different receivers by a single discrete memoryless channel W, with positive Shannon capacity C, whose output is available to all the receivers *and* the transmitter, i.e., there is perfect and instantaneous feedback. In n channel uses, the transmitter wishes to send one of M messages to *one* of the N receivers, whose identity is a priori unknown to the receivers. The requirements are that the intended recipient should decode the transmitted message correctly with high probability, and any other receiver should recognize with high probability that the message is not intended for it (these probabilities are averaged over the M messages). In this situation, we prove that M and N can *simultaneously* grow as $\exp[nR_1 - o(n)]$ and $\exp\{\exp[nR_2 - o(n)]\}$, respectively, for all rate-pairs (R_1, R_2) satisfying $R_1 \leq C$ and $R_2 \leq \max_{P:I(P;W)\geq R_1} H(PW)$, provided the transmitter can use randomized encoding. If the encoding must be deterministic, we prove that all rate-pairs (R_1, R_2) satisfying $R_1 \leq C$ and $R_2 \leq R_1 + \max_{P:I(P;W)\geq R_1} H(W|P)$ are achievable. In both cases, we establish these rate regions as optimal by proving "strong" converses.

Keywords: Identification theory, identification plus transmission, point-to-multipoint communication.

---

*Research supported by NSF IRI 9005849, IRI 9310670, NCR 9422513, and the AT&T Foundation.

†Cornell University and U.C. Berkeley.

‡Univ. of California, Berkeley.

§Address all correspondence to the first author: 341, Cory Hall, Dept. of EECS, U.C. Berkeley, Berkeley, CA 94720.

1 Introduction

Consider the following point-to-multipoint communication problem: a transmitter is connected to N different receivers by a single discrete memoryless channel (DMC) W, whose output can be observed by all the receivers. The transmitter wishes to send one of M messages to *one* of the N receivers, but the receivers do not know beforehand who the intended recipient is. The transmitter must therefore encode both the *address* and the *message* in a codeword, and send it across the channel (the address is just the index of the intended recipient). Each receiver must then observe the channel output and decide whether the message is intended for it, and, if so, decode it. The requirements are that (a) the intended recipient should decode the transmitted message correctly with high probability, and (b) any other receiver should recognize with high probability that the message is not intended for it (these probabilities are assumed to be averaged over the M possible messages). Receivers other than the intended one do not need to know either the message or the actual recipient. The question of interest, then, is how fast M and N can simultaneously grow with the number of channel uses permitted to the transmitter.

If there is only one receiver, i.e. $N = 1$, then this is just the classical *data transmission* problem of information theory. In n uses of the channel, the transmitter can then reliably send one of $M = \exp[nR - o(n)]$ messages to the receiver, provided $R \leq C$, where $C = \max_P I(P; W)$ is the Shannon capacity of the channel W (the maximum, over all input probability distributions P, of the corresponding mutual information between the channel input and output). It is also well known that allowing randomization in the encoding process does not help in any way.

On the other hand, if $M = 1$, we have the *identification* problem introduced by Ahlswede and Dueck [2]. Here, we could think of the transmitter as simply sending an alarm or "identifying" signal to one of the N receivers (there is no message to be transmitted). In [2], it was shown that one of $N = \exp\{\exp[nR - o(n)]\}$ receivers (*doubly* exponential in n) could be reliably identified in n channel uses, *provided* randomization was allowed in the encoding process. The identification capacity of the channel, defined as the maximum achievable (second-order) rate R in this situation, was shown to equal its Shannon capacity C! Here, randomization in the encoding is crucial in the sense that it is impossible to achieve any positive second-order rate deterministically, i.e., the deterministic identification capacity is zero.

The general problem of simultaneous *identification and transmission* was studied by Han and Verdu [5], as a variant of the identification problem with potential applications in multiuser communication. The straightforward solution to this problem would be to use classical transmission codes to encode the address in a header of αn symbols and the message in the remaining $(1 - \alpha)n$ symbols (n is

the number of channel uses, and $0 < \alpha < 1$ is a timesharing parameter), and this would permit $N = \exp[n\alpha C - o(n)]$ and $M = \exp[n(1-\alpha)C - o(n)]$. However, as pointed out in [5], this scheme is far from optimal because it needlessly conveys the address and message information to *all* the receivers.

We could therefore improve on the above scheme by using an identification code (instead of a transmission code) to convey the address in αn symbols, and encoding the message in $(1-\alpha)n$ symbols as before. This scheme would permit $N = \exp\{\exp[n\alpha C - o(n)]\}$ and $M = \exp[n(1-\alpha)C - o(n)]$, provided randomization was allowed in the address encoding. While this is certainly better than the straightforward solution, it is still suboptimal. In [5], it was shown that in fact $N = \exp\{\exp[nC - o(n)]\}$ and $M = \exp[nC - o(n)]$ are *simultaneously* achievable, if the address and message are *jointly* encoded using an "identification plus transmission" code (or IT code), instead of separately as above; moreover, the joint encoding does *not* require randomization.

In this paper, we study an analogous identification plus transmission problem when the transmitter has *perfect and instantaneous feedback* from the output of the DMC connecting it to the receivers. It is well known that feedback does not increase the transmission capacity of a DMC. In marked contrast, feedback can have a dramatic effect on the identification capacity of a DMC, as demonstrated by Ahlswede and Dueck in [1] (the sequel to [2]). There, it was shown that feedback allows positive second-order identification rates even when the encoding has to be deterministic. In fact, the deterministic feedback identification capacity of a DMC W with positive Shannon capacity was shown to be $\max_x H[W(\cdot|x)]$ (the maximum, over all input symbols x, of the conditional output entropy when x is transmitted). Allowing randomized encoding was shown to increase the feedback identification capacity to $\max_P H(PW)$ (the maximum, over all input probability distributions P, of the corresponding unconditional output entropy).

In this paper, we determine the region of all rate-pairs (R_1, R_2) such that it is possible for the transmitter to reliably send one of $M = \exp[nR_1 - o(n)]$ messages to one of $N = \exp\{\exp[nR_2 - o(n)]\}$ receivers across the DMC W, with positive Shannon capacity C, when it is equipped with perfect and instantaneous feedback (the reliability requirements are as described in the first paragraph). As in [1], we consider both the case where randomized encoding is allowed, and the case where the encoding is restricted to be deterministic. The "identification plus transmission" rate region turns out to be defined by the conditions

$$R_1 \leq C \quad \text{and} \quad R_2 \leq \max_{P:I(P;W)\geq R_1} H(PW)$$

in the general case, and by the conditions

$$R_1 \leq C \quad \text{and} \quad R_2 \leq R_1 + \max_{P:I(P;W)\geq R_1} H(W|P)$$

in the deterministic case. The identification theorems of [1] can be viewed as special cases of the above results, obtained by setting the transmission rate requirement R_1 to zero. As in [1], the converses proved here are "strong." As a by-product of these converses, we also have a new proof of the strong converse to Shannon's coding theorem for DMCs with feedback, a result first proved by Kemperman [6].

We will formulate the problem more precisely in Section 2 and then state our main result in Theorem 2.1.

2 Statement of problem and results

The following conventions will be in effect in the rest of the paper. All logarithms and exponentials will be to the base e. If J is an integer, then $[J]$ will denote the set $\{1, 2, \ldots, J\}$. Finally, the notation for all standard information-theoretic quantities will be that of [4].

The discrete memoryless channel (DMC) connecting the transmitter and the receivers is assumed to have finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, and transition probability function $W = \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$. In the presence of feedback, the transmitter can choose its channel input at each step of communication based on the outputs from all previous steps, according to some *feedback strategy*. As in [1], we will consider strategies that can be randomized, as well as strategies that must be deterministic.

We first define an *n-step feedback function* for the channel W as a vector $f = (f_1, \ldots, f_n)$, where $f_k : \mathcal{Y}^{k-1} \to \mathcal{X}$. When the transmitter uses this feedback function for communication, it sends the channel input symbol $X_k = f_k(Y^{k-1})$ in step k, $1 \leq k \leq n$, where Y^{k-1} is the sequence of channel outputs in the previous $k-1$ steps. For each $y^n \in \mathcal{Y}^n$, we will denote by $f(y^n)$ the $x^n \in \mathcal{X}^n$ given by $x_k = f_k(y^{k-1})$, $1 \leq k \leq n$.

Let $W_f(x^n, y^n)$ be the probability that $(X^n, Y^n) = (x^n, y^n)$ when the transmitter uses the feedback function f, and let $Q_f(y^n)$ be the corresponding marginal probability that $Y^n = y^n$. Clearly,

$$\begin{aligned} Q_f(y^n) &= \prod_{k=1}^{n} W(y_k \,|\, f_k(y^{k-1})) \\ &= W^n(y^n \,|\, f(y^n)), \end{aligned}$$

and

$$W_f(x^n, y^n) = \begin{cases} Q_f(y^n) & \text{if } x^n = f(y^n); \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{F}_n$ denote the set of all n-step feedback functions (note that $\mathcal{F}_n$ is finite). An *n-step feedback strategy* F for the channel W is defined as a probability

distribution (p.d.) on $\mathcal{F}_n$. To communicate according to the strategy F, the transmitter randomly chooses a feedback function $f \in \mathcal{F}_n$ with distribution F, and then uses this f, as described earlier, to decide its channel input at each step. We will denote the probability that $(X^n, Y^n) = (x^n, y^n)$ under the strategy F by $W_F(x^n, y^n)$, and the corresponding marginal probability that $Y^n = y^n$ by $Q_F(y^n)$. Then,

$$\begin{aligned} W_F(x^n, y^n) &= \sum_{f \in \mathcal{F}_n} F(f)\, W_f(x^n, y^n), \\ Q_F(y^n) &= \sum_{f \in \mathcal{F}_n} F(f)\, Q_f(y^n). \end{aligned}$$

A general strategy, as defined above, is allowed to use randomization. The strategy F is called *deterministic* if $F(f) = 1$ for some $f \in \mathcal{F}_n$. Clearly, such a strategy does not require any randomization. Moreover, there is an obvious 1-1 correspondence between $\mathcal{F}_n$ and the set of n-step deterministic strategies.

We will now define the identification plus transmission codes, or IT codes, to be studied here.

Definition 2.1 *An (n, N, M, λ, μ) identification plus transmission (IT) code is a collection $\{(F_{a,m}, \mathcal{D}_{a,m}) : (a.m) \in [N] \times [M]\}$. where $F_{a,m}$ is an n-step strategy, $\mathcal{D}_{a,m} \subseteq \mathcal{Y}^n$, and, for each $a \in [N]$,*

1. *$\mathcal{D}_{a,m} \cap \mathcal{D}_{a,m'}$ is empty if $m \neq m'$.*
2. *$(1/M) \sum_{m=1}^{M} Q_{F_{a,m}}(\mathcal{D}_{a,m}) > 1 - \lambda$.*
3. *$(1/M) \sum_{m=1}^{M} Q_{F_{a,m}}(\mathcal{D}_{a'}) < \mu$ for all $a' \neq a$, where $\mathcal{D}_{a'} = \bigcup_{m=1}^{M} \mathcal{D}_{a',m}$.*

The code is called deterministic if all the strategies $F_{a,m}$ are deterministic.

The interpretation of the above code is as follows: if the transmitter wishes to send message m to receiver a, it communicates according to the strategy $F_{a,m}$. After n steps of communication, receiver a decides that it is indeed the intended recipient if the received sequence Y^n falls in $\mathcal{D}_a = \bigcup_m \mathcal{D}_{a,m}$; in this case, because of Condition (1), there is a unique $m \in [M]$ such that $Y^n \in \mathcal{D}_{a,m}$, and it takes this m as the transmitted message. Otherwise, i.e. if $Y^n \notin \mathcal{D}_a$, receiver a decides that the message is not intended for it. Condition (2) in the above definition guarantees that the intended recipient decodes the transmitted message correctly with probability greater than $1-\lambda$, while Condition (3) guarantees that any other receiver wrongly decides it is the recipient with probability less than μ. Note that these probabilities are averaged over the M possible messages.

Definition 2.2 *The region of (λ,μ)-achievable rate-pairs $\mathcal{R}(\lambda,\mu)$ is the set of (R_1,R_2) for which there exists a sequence of (n,N_n,M_n,λ,μ) IT codes satisfying*

$$\liminf_{n\to\infty} n^{-1}\log M_n \;=\; R_1 \quad \textit{and} \quad \liminf_{n\to\infty} n^{-1}\log\log N_n \;=\; R_2. \tag{1}$$

The region of deterministically (λ,μ)-achievable rate-pairs $\mathcal{R}_D(\lambda,\mu)$ is the set of (R_1,R_2) for which there exists a sequence of (n,N_n,M_n,λ,μ) deterministic IT codes satisfying (1).

The main result in this paper is the determination of the capacity regions $\mathcal{R}(\lambda,\mu)$ and $\mathcal{R}_D(\lambda,\mu)$ when $\lambda>0$, $\mu>0$, and $\lambda+\mu<1$.

Theorem 2.1 *Let $\lambda>0$, $\mu>0$, and $\lambda+\mu<1$. If the channel W has positive Shannon capacity $C=\max_P I(P;W)$, then*

$$\mathcal{R}(\lambda,\mu)=\left\{(R_1,R_2):\; R_1\le C \;\textit{and}\; R_2 \le \max_{P:I(P;W)\ge R_1} H(PW)\right\}, \tag{2}$$

and

$$\mathcal{R}_D(\lambda,\mu)=\left\{(R_1,R_2):\; R_1\le C \;\textit{and}\; R_2 \le R_1+\max_{P:I(P;W)\ge R_1} H(W|P)\right\}. \tag{3}$$

We will prove the achievability parts of Theorem 2.1 in Section 3, and the converse parts in Section 4.

Remarks:

1) The assumptions $\lambda>0$ and $\mu>0$ are of course reasonable. We also need the assumption $\lambda+\mu<1$ in order to get meaningful results. For if $\lambda+\mu>1$ then arbitrarily high identification rates R_2 are achievable. To see this, note that if $1-\lambda<(1/M)\sum_m Q_{F_m}(\mathcal{D}_m)<\mu$ for some n-step strategies $F_1,\ldots,F_M$ and pairwise disjoint subsets $\mathcal{D}_1,\ldots,\mathcal{D}_M$ of $\mathcal{Y}^n$, then, by setting $F_{a,m}=F_m$ and $\mathcal{D}_{a,m}=\mathcal{D}_m$ for all (a,m), we have an (n,N,M,λ,μ) IT code with arbitrarily large N. (If λ itself is greater than 1, then we can also achieve arbitrarily large transmission rates R_1.)

2) The assumption that $C>0$ prevents trivialities at the other end of the spectrum. It can be verified that if $C=0$, then there does not exist any (n,N,M,λ,μ) IT code with $N>1$ or $M>(1-\lambda)^{-1}$, assuming $\lambda>0$, $\mu>0$, and $\lambda+\mu<1$.

3) We will actually prove the achievability results with Condition (2) in Definition 2.1 replaced by the stronger condition "$Q_{F_{a,m}}(\mathcal{D}_{a,m})>1-\lambda$ for all m." However, it is *not* possible to similarly replace Condition (3) by "$Q_{F_{a,m}}(\mathcal{D}_{a'})<\mu$ for all m and all $a'\neq a$" without affecting the results. Averaging over messages is essential in controlling the probability of a receiver wrongly deciding that it is the intended recipient (the reason will become clear from the achievability proof).

3 Proofs of the achievability parts

We will prove the achievability result for general IT codes in Section 3.1. With very minor changes, the same proof will work in the deterministic case as well (see Section 3.2). But first we state a couple of lemmas needed in the proofs.

Lemma 3.1 *Let P be an n-type on $\mathcal{X}$ for some $n \geq 1$ (i.e., P is a p.d. on $\mathcal{X}$ such that $nP(x)$ is an integer for all $x \in \mathcal{X}$).*

1. *If $R' \geq 0$ and $J \leq \exp(nR')$, then there exist sequences $\mathbf{c}_1, \ldots, \mathbf{c}_J$ in $\mathcal{X}^n$, all of type P, and a partition of $\mathcal{Y}^n$ into subsets $\mathcal{C}_1, \ldots, \mathcal{C}_J$, such that*

$$1 - W^n(\mathcal{C}_j|\mathbf{c}_j) \leq (n+1)^{4d} \exp\left[-nE'(R', P)\right] \qquad \text{for all } j \in [J].$$

   Here, $d = |\mathcal{X}||\mathcal{Y}|$, and

$$E'(R', P) = \min_V \left\{ D(V\|W|P) + [I(P;V) - R']^+ \right\}.$$

2. *If $R'' \geq 0$ and $K \leq \exp(nR'')$, then for any $\mathbf{c} \in \mathcal{X}^n$ of type P, and any $\mathcal{C} \subseteq \mathcal{Y}^n$, there exists a partition of $\mathcal{C}$ into subsets $\mathcal{C}_e, \mathcal{C}_1, \ldots, \mathcal{C}_K$, such that*

$$\begin{aligned} W^n(\mathcal{C}_k|\mathbf{c}) &= (1/K)\left[W^n(\mathcal{C}|\mathbf{c}) - W^n(\mathcal{C}_e|\mathbf{c})\right] \qquad \text{for all } k \in [K], \\ W^n(\mathcal{C}_e|\mathbf{c}) &\leq (n+1)^d \exp\left[-nE''(R'', P)\right]. \end{aligned}$$

   Here,

$$E''(R'', P) = \min_V \left\{ D(V\|W|P) + [H(V|P) - R'']^+ \right\}.$$

The minima in the definitions of $E'(\cdot,\cdot)$ and $E''(\cdot,\cdot)$ are both over the set of all DMCs V with alphabets $\mathcal{X}$ and $\mathcal{Y}$. Both $E'(R', \cdot)$ and $E''(R'', \cdot)$ are continuous functions. $E'(R', P)$ is positive if $R' < I(P;W)$ and zero otherwise. Similarly, $E''(R'', P)$ is positive if $R'' < H(W|P)$ and zero otherwise.

Proof: Appendix. □

Lemma 3.2 *Let $0 < \epsilon < 1$, and let $J \geq 2/\epsilon$, $S \geq 1$, and $N < \exp(\epsilon^2 S/2)$ be integers. Then, there exists an $N \times S$ array (N rows and S columns) with entries from $[J]$, any two rows of which are at a Hamming distance greater than $(1-\epsilon)S$ from each other (i.e., they differ in more than $(1-\epsilon)S$ positions).*

Proof: Appendix. □

The first part of Lemma 3.1 implies the existence of an "equitype" transmission code of blocklength n with about $\exp[nI(P;W)]$ codewords (P is the

common type of the codewords), whose maximal error probability over W decays exponentially with n (note that the code is for the DMC W *without* feedback). This is of course a standard result in channel coding [4] that can be used to prove Shannon's coding theorem for DMCs. We will use this result in the following way: given $R_1 < C$, the transmitter and receivers choose a type P such that $I(P;W) > R_1$, and agree on an equitype code as above. In the general case, the transmitter uses this code to reliably convey one of $M = \exp(nR_1)$ messages *and* an independently generated random variable of entropy about $n[I(P;W) - R_1]$, to the receivers. In the deterministic case, the transmitter simply conveys the message alone using the code.

By the second part of Lemma 3.1, the decoding set corresponding to the transmitted codeword can be partitioned into about $\exp[nH(W|P)]$ sets — all of which have *exactly* the same probability — and a remaining "error" set whose probability decays exponentially with n. Thus, by observing which of these subsets the channel output falls in, the transmitter and receivers can agree on a random variable of entropy about $nH(W|P)$ (the channel output is known to the transmitter because of feedback). Since the codeword itself carries an entropy of about $nI(P;W)$ (in the general case) or nR_1 (in the deterministic case), the transmitter and receivers have a total *common randomness* of about $nI(P;W) + nH(W|P) = nH(PW)$ (general case) or $nR_1 + nH(W|P)$ (deterministic case).

Now suppose $R_2 < H(PW)$ (general case) or $R_2 < R_1 + H(W|P)$ (deterministic case), and $N = \exp[\exp(nR_2)]$. Then, by Lemma 3.2, we can find an $N \times S$ array whose rows are far apart in Hamming distance, with S roughly equal to $\exp[nH(PW)]$ or $\exp\{n[R_1 + H(W|P)]\}$ (in the respective cases). We can index the rows of the array by the N receivers. The common randomness gained by the codeword transmission can then be used by the transmitter and receivers to agree on a random column of this array (in either case). If the transmitter now sends the array element in that column and the row corresponding to the intended recipient (this can be done reliably with a constant number of channel uses), then the Hamming distance properties of the array enable the receivers to decide reliably whether or not the message is intended for them.

The idea of using channel noise to generate randomness for identification coding in the presence of feedback is one of the key principles in the results of [1]. The details of the idea there are somewhat different, and the method outlined here (based on the second part of Lemma 3.1) seems to allow a somewhat simpler analysis.

Lemma 3.2 is based on the arguments in Section III of [1], though it is not stated there in this form. It is the essence of the "$\sqrt{n}$ trick" of [1] — so called because the equivalent of the array element encoding was done there with about $\sqrt{n}$ channel uses — and can in fact be used to prove all known achievability

results in identification theory. In its present form, the name "array trick" may be more appropriate.

3.1 The general case

Given $\lambda > 0$ and $\mu > 0$, let $\epsilon = (1/4)\min(\lambda, \mu)$. We will prove that any rate-pair (R_1, R_2) satisfying

$$0 \leq R_1 < C - \delta \quad \text{and} \quad 0 \leq R_2 < \max_{P:I(P;W)\geq R_1+\delta} H(PW) - \delta \tag{4}$$

for some $\delta > 0$ is in $\mathcal{R}(4\epsilon, 4\epsilon)$, hence in $\mathcal{R}(\lambda, \mu)$. Since $\mathcal{R}(\lambda, \mu)$ is a closed subset of $\mathbf{R}^2$ (because of the way achievable rate-pairs are defined), it will follow that any point in the closure of the set of all (R_1, R_2) satisfying (4) for some $\delta > 0$ is also in $\mathcal{R}(\lambda, \mu)$. By the continuity of $\max_{I(P;W)\geq R} H(PW)$ in R, this closure is precisely the RHS of (2), so that the desired achievability result will be proved.

From now on, fix a $\delta > 0$ and any (R_1, R_2) satisfying (4). Let $M_n = \lfloor \exp(nR_1) \rfloor$ and $N_n = \lfloor \exp[\exp(nR_2)] \rfloor$. To prove that $(R_1, R_2) \in \mathcal{R}(4\epsilon, 4\epsilon)$, we will show that for all sufficiently large n there exists an $(n + t, N_n, M_n, 4\epsilon, 4\epsilon)$ IT code. Here, t is a constant (i.e., it does not depend on n).

3.1.1 Preliminaries

Let P^* be any p.d. on $\mathcal{X}$ that maximizes $H(PW)$ subject to the constraint $I(P;W) \geq R_1 + \delta$. Let $R' = I(P^*;W) - \delta/2$ and $R'' = H(W|P^*) - \delta/2$. Then, $E'(R', P^*)$ and $E''(R'', P^*)$ are both positive. Pick any sequence $\{P_n\}$, with P_n an n-type on $\mathcal{X}$, such that $P_n \to P^*$ as $n \to \infty$. By continuity, $E'(R', P_n) \to E'(R', P^*) > 0$ and $E''(R'', P_n) \to E''(R'', P^*) > 0$, so that there must exist a γ satisfying

$$0 < \gamma < \min\{E'(R', P_n), E''(R'', P_n)\} \quad \text{for all large } n. \tag{5}$$

Let $L_n = \lfloor \exp[n(R' - R_1)] \rfloor$. Then, $M_n L_n \leq \exp(nR')$, and the first part of Lemma 3.1 guarantees the existence of sequences $\mathbf{c}_{ml} \in \mathcal{X}^n$ of type P_n, and sets $\mathcal{C}_{ml}$ partitioning $\mathcal{Y}^n$, such that $1 - W^n(\mathcal{C}_{ml}|\mathbf{c}_{ml}) \leq \alpha_n = \exp[-nE'(R', P_n) + o(n)]$, for all $(m, l) \in [M_n] \times [L_n]$. By (5), $\alpha_n < \epsilon$ for all large n.

Let $K_n = \lfloor \exp(nR'') \rfloor$. Then, by the second part of Lemma 3.1, $\mathcal{C}_{ml}$ can be partitioned into subsets $\mathcal{C}_{mlk}$, $k \in \{e\} \cup [K_n]$, such that $W^n(\mathcal{C}_{mlk}|\mathbf{c}_{ml})$ is the same for all $k \in [K_n]$, and $W^n(\mathcal{C}_{mle}|\mathbf{c}_{ml}) \leq \beta_n = \exp[-nE''(R'', P_n) + o(n)]$, for all $(m, l) \in [M_n] \times [L_n]$. By (5), $\beta_n < \epsilon$ for all large n.

Note that $\lim n^{-1} \log(M_n L_n K_n) = H(P^*W) - \delta > R_2$, by assumption. Hence, for all large n, $\exp(nR_2) < (\epsilon^2/2) M_n L_n K_n$ and $N_n < \exp[(\epsilon^2/2) M_n L_n K_n]$. By Lemma 3.2, then, there exists an $N_n \times (M_n L_n K_n)$ array with entries from $[J]$, any

two rows of which are at a Hamming distance greater than $(1-\epsilon)M_nL_nK_n$, if n is large. Here, we may take $J = \lceil 2/\epsilon \rceil$. We will denote this array by $\mathcal{A}$, and take its columns to be indexed by triples $(m,l,k) \in [M_n] \times [L_n] \times [K_n]$. $\mathcal{A}(a;m,l,k)$ will denote the array element in row a and column (m,l,k).

Finally, pick an integer t large enough that there exist sequences $\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_J$ in $\mathcal{X}^t$ and a partition $\tilde{\mathcal{C}}_1, \ldots, \tilde{\mathcal{C}}_J$ of $\mathcal{Y}^t$, satisfying $1 - W^t(\tilde{\mathcal{C}}_j|\tilde{\mathbf{c}}_j) < \epsilon$ for all $j \in [J]$. This is possible because the channel has positive Shannon capacity.

3.1.2 The strategies $F_{a,m}$ and decoding sets $\mathcal{D}_{a,m}$

We will now describe how the transmitter encodes the address-message pair $(a,m) \in [N_n] \times [M_n]$. The encoding is in two stages. In the first stage, it picks a random $l \in [L_n]$ with a uniform distribution, and sends the sequence $\mathbf{c}_{ml}$ across the channel. There is then a unique triple $(\hat{m},\hat{l},k) \in [M_n] \times [L_n] \times (\{e\} \cup [K_n])$ such that the corresponding channel output sequence lies in $\mathcal{C}_{\hat{m}\hat{l}k}$. This triple, and in particular k, is known to the transmitter as well as all the receivers because of feedback. This completes the first stage.

If $k \in [K_n]$, then the transmitter sends the sequence $\tilde{\mathbf{c}}_j$ in the second stage, where $j = \mathcal{A}(a;m,l,k)$ is the element in row a and column (m,l,k) of the array $\mathcal{A}$; correspondingly, there is a unique $\hat{j} \in [J]$ such that the output sequence falls in $\tilde{\mathcal{C}}_{\hat{j}}$. On the other hand, if $k = e$, both transmitter and receiver declare an error, and the transmitter sends a dummy sequence of length t, say $\tilde{\mathbf{c}}_e$, in the second stage. We have thus implicitly defined $(n+t)$-step strategies $F_{a,m}$ for each $(a,m) \in [N_n] \times [M_n]$.

We will now describe what receiver a' ($a' \in [N_n]$) does at the end of $n+t$ steps: it simply checks if $\mathcal{A}(a';\hat{m},\hat{l},k) = \hat{j}$. If so, it assumes that it is indeed the intended recipient, and that the transmitted message is $\hat{m}$. Otherwise, i.e. if $\mathcal{A}(a';\hat{m},\hat{l},k) \neq \hat{j}$, it decides that the message is not intended for it. Formally, this means that the decoding region $\mathcal{D}_{a,m} \subset \mathcal{Y}^{n+t}$ equals $\bigcup_{(l,k,j)} \mathcal{C}_{mlk} \times \tilde{\mathcal{C}}_j$, the union extending over all those triples (l,k,j) for which $\mathcal{A}(a;m,l,k) = j$.

3.1.3 Analysis

We will now bound the error probabilities of the IT code just defined. Suppose the transmitter attempts to convey message m to receiver a. First note that if $(\hat{m},\hat{l}) = (m,l)$, $k \neq e$, and $\hat{j} = j$, then receiver a will recognize that the message is intended for it, and decode it correctly as m. Therefore, by a union bound,

$$\begin{aligned} 1 - Q_{F_{a,m}}(\mathcal{D}_{a,m}) &\leq \alpha_n + \beta_n + \epsilon \\ &< 3\epsilon \quad \text{for all large } n. \end{aligned}$$

Next, consider any receiver $a' \neq a$. Note that if $(\hat{m}, \hat{l}) = (m, l)$, $k \neq e$, $\hat{j} = j$, and $\mathcal{A}(a'; m, l, k) \neq \mathcal{A}(a; m, l, k)$, then receiver a' will correctly recognize that the transmitted message is not intended for it. As before, the probability that either $(\hat{m}, \hat{l}) \neq (m, l)$, or $k = e$, or $\hat{j} \neq j$ is at most $\alpha_n + \beta_n + \epsilon$. Further, given that $(\hat{m}, \hat{l}) = (m, l)$, $k \neq e$, and $\hat{j} = j$, the probability that $\mathcal{A}(a'; m, l, k) = \mathcal{A}(a; m, l, k)$ is equal to $(L_n K_n)^{-1}$ times $E_m(a, a')$, where $E_m(a, a')$ is the number of pairs (l, k) such that $\mathcal{A}(a'; m, l, k) = \mathcal{A}(a; m, l, k)$. This is because l is chosen with a uniform distribution over $[L_n]$, and, conditional on $k \neq e$, k has a uniform distribution on $[K_n]$ for all values of (m, l). Thus, we have

$$Q_{F_{a,m}}(\mathcal{D}_{a'}) \leq \alpha_n + \beta_n + \epsilon + \frac{E_m(a, a')}{L_n K_n},$$

so that

$$\begin{aligned} \frac{1}{M_n} \sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_{a'}) &\leq \alpha_n + \beta_n + \epsilon + \frac{1}{M_n} \sum_{m=1}^{M_n} \frac{E_m(a, a')}{L_n K_n} \\ &< 4\epsilon \qquad \text{for all large } n, \end{aligned}$$

because $\sum_m E_m(a, a')$ is just the number of positions in which rows a and a' of the array $\mathcal{A}$ agree. This completes the proof.

3.2 The deterministic case

The achievability proof in the deterministic case is very similar. Arguing as in the general case, it suffices to prove that any rate-pair (R_1, R_2) satisfying

$$0 \leq R_1 < C - \delta \quad \text{and} \quad 0 \leq R_2 < R_1 + \max_{P: I(P;W) \geq R_1 + \delta} H(W|P) - \delta$$

for some $\delta > 0$ is in $\mathcal{R}(4\epsilon, 4\epsilon)$, where $\epsilon = (1/4)\min(\lambda, \mu)$, and $\lambda > 0$, $\mu > 0$ are given.

The above statement will be proved if we show that for all sufficiently large n there exists an $(n+t, N_n, M_n, 4\epsilon, 4\epsilon)$ deterministic IT code, with $M_n = \lfloor \exp(nR_1) \rfloor$ and $N_n = \lfloor \exp[\exp(nR_2)] \rfloor$ (t being a constant, as before).

We have a proof of the existence of such codes if we simply change three sentences in the proof for the general case, starting from Section 3.1.1. These are the first sentences of paragraphs 1, 2, and 4 of Section 3.1.1. The first of these must be changed to "Let P^* be any p.d. on $\mathcal{X}$ that maximizes $H(W|P)$ subject to the constraint $I(P; W) \geq R_1 + \delta$"; the second to "Let $L_n = 1$"; and the third to "Note that $\lim n^{-1} \log(M_n L_n K_n) = R_1 + H(W|P^*) - \delta/2 > R_2$, by assumption."

But for these changes, the proof in the general case carries over word-for-word. The resulting sequence of IT codes is indeed deterministic because $L_n = 1$ here; an inspection of the previous proof shows that the transmitter needs randomization *only* to generate a random $l \in [L_n]$.

4 Proofs of the converse parts

Let $\lambda > 0$, $\mu > 0$, and $\lambda + \mu < 1$. Consider any sequence of $(n, N_n, M_n, \lambda, \mu)$ IT codes $\{(F_{a,m}, \mathcal{D}_{a,m})\}$ achieving the rate-pair (R_1, R_2). (To avoid cumbersome notation, we have suppressed the dependence of $F_{a,m}$ and $\mathcal{D}_{a,m}$ on n.) We will now outline the ideas for bounding the transmission rate R_1 and the identification rate R_2. We will treat the general case and the deterministic case in parallel.

To begin with, note that $\{(F_{1,m}, \mathcal{D}_{1,m}) : m \in [M_n]\}$ is a sequence of (n, M_n) transmission codes with average error probability λ, for the DMC W with feedback. The encoding may involve randomization, but this of course does not help at all in the context of transmission codes. Since $\lambda < 1$, the strong converse to Shannon's theorem for DMCs with feedback (first proved by Kemperman [6]) yields $\limsup n^{-1} \log M_n \leq C$. This implies that $R_1 = \liminf n^{-1} \log M_n \leq C$, which is the required bound on the transmission rate both for general and deterministic IT codes. However, it turns out that we can prove Kemperman's result here with very little additional effort, and will therefore not appeal to it directly. Our proof of this result is different from the original one.

The idea for bounding the identification rate is similar to that in [1]. Pick any $\gamma \in (\mu, 1 - \lambda)$, say $\gamma = (1 - \lambda + \mu)/2$. Suppose we could find subsets $\mathcal{D}_a^* \subseteq \mathcal{D}_a$ of the decoding regions such that

$$\frac{1}{M_n} \sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_a^*) > \gamma \quad \text{and} \quad |\mathcal{D}_a^*| \leq K_n \qquad \forall\ a \in [N_n]. \tag{6}$$

Then, since

$$\begin{aligned} \frac{1}{M_n} \sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_{a'}^*) &\leq \frac{1}{M_n} \sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_{a'}) \\ &< \gamma \qquad \text{for all } a' \neq a, \end{aligned}$$

the sets $\mathcal{D}_a^*$, $a \in [N_n]$, would have to be distinct. But this would imply that N_n is no bigger than the number of distinct subsets of $\mathcal{Y}^n$ whose size is at most K_n. The latter, in turn, is bounded above by $|\mathcal{Y}^n|^{K_n}$, so that we would have

$$n^{-1} \log \log N_n \leq n^{-1} \log K_n + o(1). \tag{7}$$

We will prove that if n is large enough then there exist subsets $\mathcal{D}_a^* \subseteq \mathcal{D}_a$ for which (6) holds, with the uniform bound K_n satisfying

$$n^{-1} \log K_n = \max_{P : I(P;W) \geq R_1 - \delta} H(PW) + o(1) \tag{8}$$

in the general case, and

$$n^{-1} \log K_n = n^{-1} \log M_n + \max_{P : I(P;W) \geq R_1 - \delta} H(W|P) + o(1) \tag{9}$$

in the deterministic case. Here, δ is an arbitrary positive number. The required bounds on the identification rate $R_2 = \liminf n^{-1} \log\log N_n$ in the two cases will then follow from (7), (8), (9), and the continuity in R of $\max_{I(P;W)\geq R} H(PW)$ and $\max_{I(P;W)\geq R} H(W|P)$.

Definition 4.1 *Let F be an n-step strategy.*

1. *If $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$, then $P_k^F(x|x^{k-1}, y^{k-1})$ is the probability that $X_k = x$ given $(X^{k-1}, Y^{k-1}) = (x^{k-1}, y^{k-1})$, under the strategy F. $P^F_{x^n,y^n}$ is the p.d. on $\mathcal{X}$ given by*

$$P^F_{x^n,y^n}(x) = n^{-1} \sum_{k=1}^{n} P_k^F(x|x^{k-1}, y^{k-1}).$$

2. *The "typical set" $\mathcal{E}(F)$ for the strategy F is the set of (x^n, y^n) such that $W_F(x^n, y^n) > 0$, and*

$$\left|N(x, y|x^n, y^n) - nP^F_{x^n,y^n}(x)W(y|x)\right| \leq n^{3/4}\sqrt{W(y|x)} \quad \forall\ (x, y).$$

   Here, $N(x, y|x^n, y^n) = |\{k : (x_k, y_k) = (x, y)\}|$. The "section" of $\mathcal{E}(F)$ at y^n is $\mathcal{E}_{y^n}(F) = \{x^n : (x^n, y^n) \in \mathcal{E}(F)\}$.

3. $\mathcal{B}_\alpha(F) = \left\{y^n : \exists\ x^n \in \mathcal{E}_{y^n}(F) \text{ such that } I(P^F_{x^n,y^n}; W) > \alpha\right\}$.

Remarks:

1) If the probability that $(X^{k-1}, Y^{k-1}) = (x^{k-1}, y^{k-1})$ under F is zero, then $P_k^F(\cdot|x^{k-1}, y^{k-1})$ is not well-defined, but this will be of no consequence (in this case, $P_k^F(\cdot|x^{k-1}, y^{k-1})$ can be arbitrarily defined).

2) The condition $W_F(x^n, y^n) > 0$ in the definition of $\mathcal{E}(F)$ ensures that $P_k^F(\cdot|x^{k-1}, y^{k-1})$ is well-defined for all k, and hence that $P^F_{x^n,y^n}$ is well-defined. Also, the term $n^{3/4}$ appearing in the definition can be replaced by any n^γ, $1/2 < \gamma < 1$, without essentially affecting the results.

Not surprisingly, the typical set $\mathcal{E}(F)$ carries most of the probability under the p.d. W_F (see (10) in Lemma 4.1), and sequences in $\mathcal{E}(F)$ are "well-behaved" in various ways (see Parts (a), (b), (c) of Lemma 4.1).

3) Later, we will choose the set $\mathcal{D}_a^*$ to be $\bigcup_m \mathcal{D}_{a,m} \cap \mathcal{B}_\alpha(F_{a,m})$ for each $a \in [N_n]$, with $\alpha = R_1 - \delta$. This amounts to throwing away all sequences y^n in the decoding region $\mathcal{D}_{a,m}$ which are, roughly speaking, either "atypical" (i.e., $\mathcal{E}_{y^n}(F_{a,m})$ is empty) or have "low mutual information" (i.e., $I(P^{F_{a,m}}_{x^n,y^n}; W) \leq R_1 - \delta$ for all $x^n \in \mathcal{E}_{y^n}(F_{a,m})$). The intuition is that, if n is large, such sequences cannot contribute significantly to the probability of $\mathcal{D}_a$ because the IT code is required to transmit messages at rate R_1 (see (11) in Lemma 4.2). The removal of such sequences also trims down $\mathcal{D}_a$ to the right size (see (12) and (13) in Lemma 4.2).

Lemma 4.1 *Let $d = |\mathcal{X}|\,|\mathcal{Y}|$. Then, for any n-step strategy F,*

$$1 - W_F(\mathcal{E}(F)) \;\leq\; dn^{-1/2}. \tag{10}$$

If $(x^n, y^n) \in \mathcal{E}(F)$, and y^n has type Q, then

(a) $\left|\log Q^n(y^n) + nH(P^F_{x^n,y^n}W)\right| \;\leq\; dn^{7/8}.$

(b) $\left|\log W^n(y^n|x^n) + nH(W|P^F_{x^n,y^n})\right| \;\leq\; dn^{7/8}.$

(c) $\left|\log\left(\frac{W^n(y^n|x^n)}{Q^n(y^n)}\right) - nI(P^F_{x^n,y^n};W)\right| \;\leq\; 2dn^{7/8}.$

Proof: Appendix. □

Lemma 4.2 *Let $F_1, \ldots, F_M$ be n-step strategies, and $\mathcal{D}_1, \ldots, \mathcal{D}_M$ pairwise disjoint subsets of $\mathcal{Y}^n$. Then,*

$$\frac{1}{M}\sum_{m=1}^{M} Q_{F_m}(\mathcal{D}_m \cap \mathcal{B}^c_\alpha(F_m)) \;\leq\; \frac{\exp\{n\alpha + 2dn^{7/8}\}}{M}(n+1)^d \;+\; dn^{-1/2}. \tag{11}$$

Here, $\mathcal{B}^c_\alpha(F_m) = \mathcal{Y}^n - \mathcal{B}_\alpha(F_m)$. Further,

$$\left|\bigcup_{m=1}^{M} \mathcal{D}_m \cap \mathcal{B}_\alpha(F_m)\right| \;\leq\; (n+1)^d \exp\left\{n\left[\max_{P:I(P;W)\geq\alpha} H(PW)\right] \;+\; dn^{7/8}\right\}. \tag{12}$$

If $F_1, \ldots, F_M$ are deterministic, then

$$\left|\bigcup_{m=1}^{M} \mathcal{D}_m \cap \mathcal{B}_\alpha(F_m)\right| \;\leq\; M \cdot \exp\left\{n\left[\max_{P:I(P;W)\geq\alpha} H(W|P)\right] \;+\; dn^{7/8}\right\}. \tag{13}$$

Proof: Appendix. □

We will now return to the sequence of $(n, N_n, M_n, \lambda, \mu)$ IT codes $\{(F_{a,m}, \mathcal{D}_{a,m})\}$ at the beginning of this section, and complete the proofs of the converses. First, note that if $\alpha = C$ then $\mathcal{B}_\alpha(F_{1,m})$ is empty for all $m \in [M_n]$, so that $\mathcal{D}_{1,m} = \mathcal{D}_{1,m} \cap \mathcal{B}^c_\alpha(F_{1,m})$. Therefore,

$$\begin{aligned} 1 - \lambda \;&<\; \frac{1}{M_n}\sum_{m=1}^{M_n} Q_{F_{1,m}}(\mathcal{D}_{1,m}) \\ &\leq\; \frac{\exp[nC + o(n)]}{M_n} \;+\; o(1), \end{aligned} \tag{14}$$

where the second inequality is by (11), applied with $\alpha = C$. Since $\lambda < 1$, (14) implies that $\limsup n^{-1}\log M_n \leq C$, hence that $R_1 \leq C$. We have now proved Kemperman's strong converse for DMCs with feedback. It has already been

pointed out that the extra generality of possible randomization in the encoding is superficial.

Pick any $\delta > 0$. As mentioned before, we will define the sets $\mathcal{D}_a^*$ in the following way:

$$\mathcal{D}_a^* = \bigcup_{m=1}^{M_n} \mathcal{D}_{a,m} \cap \mathcal{B}_\alpha(F_{a,m}), \qquad \alpha = R_1 - \delta.$$

Then, $|\mathcal{D}_a^*|$ is bounded above by the RHS of (12) in the general case, and by the RHS of (13) in the deterministic case, with α replaced by $R_1 - \delta$ and M by M_n. Moreover, for any $a \in [N_n]$,

$$\begin{aligned}
\frac{1}{M_n}\sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_a^*) &\geq \frac{1}{M_n}\sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_{a,m} \cap \mathcal{B}_\alpha(F_{a,m})) \\
&= \frac{1}{M_n}\sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_{a,m}) - \frac{1}{M_n}\sum_{m=1}^{M_n} Q_{F_{a,m}}(\mathcal{D}_{a,m} \cap \mathcal{B}_\alpha^c(F_{a,m})) \\
&> 1 - \lambda - \frac{\exp[n(R_1 - \delta) + o(n)]}{M_n} - o(1) \\
&> \gamma \qquad \text{for all large } n.
\end{aligned}$$

In the second inequality above, we have once again used (11), this time with $\alpha = R_1 - \delta$. Thus, the sets $\mathcal{D}_a^*$ have all the properties postulated earlier in (6), (8), and (9), and the converses are proved.

Appendix

Proof of Lemma 3.1: As mentioned earlier, the first part of the lemma is a standard result in channel coding. A proof can be found in [4] (Theorem 5.2 on p. 165). We will therefore only sketch the proof of the second part. In the course of the proof, we will make use of some simple results on types, without explicit reference (all of them can be found in [4]).

Let $\mathcal{W}_n(P)$ be the set of those DMCs V (with alphabets $\mathcal{X}$ and $\mathcal{Y}$) such that $nP(x)V(y|x)$ is an integer for all x, y. For any such V, define $\mathcal{T}_V(\mathbf{c})$ to be the set of y^n such that $N(x, y|\mathbf{c}, y^n) = nP(x)V(y|x)$ for all x, y. Here, $N(x, y|\mathbf{c}, y^n)$ is the number of occurences of the pair (x, y) in $(\mathbf{c}, y^n)$.

For each $V \in \mathcal{W}_n(P)$, construct K pairwise disjoint subsets of $\mathcal{C} \cap \mathcal{T}_V(\mathbf{c})$, say $\mathcal{C}_1(V), \ldots, \mathcal{C}_K(V)$, each of size exactly $\lfloor |\mathcal{C} \cap \mathcal{T}_V(\mathbf{c})|/K \rfloor$ (the subsets are otherwise arbitrary). Let

$$\mathcal{C}_k = \bigcup_{V \in \mathcal{W}_n(P)} \mathcal{C}_k(V), \quad k \in [K].$$

Then $W^n(\mathcal{C}_k|\mathbf{c})$ is the same for all $k \in [K]$, since the number of sequences in $\mathcal{C}_k$ of a given conditional type w.r.t. $\mathbf{c}$ is the same for all k. It remains to upper bound $W^n(\mathcal{C}_\epsilon|\mathbf{c})$, where $\mathcal{C}_\epsilon$ is the set of those sequences in $\mathcal{C}$ that are not in any of the $\mathcal{C}_k$, $k \in [K]$. Note that $\mathcal{C}_\epsilon$ contains exactly $|\mathcal{C} \cap \mathcal{T}_V(\mathbf{c})| \bmod K$ sequences from $\mathcal{T}_V(\mathbf{c})$. Since $K \leq \exp(nR'')$ and $|\mathcal{T}_V(\mathbf{c})| \leq \exp[nH(V|P)]$, we have

$$\begin{aligned} |\mathcal{C} \cap \mathcal{T}_V(\mathbf{c})| \bmod K &\leq \min\{K, |\mathcal{T}_V(\mathbf{c})|\} \\ &\leq \min\{\exp(nR''), \exp[nH(V|P)]\} \\ &= \exp\left\{n\left[H(V|P) - (H(V|P) - R'')^+\right]\right\}. \end{aligned}$$

For any $y^n \in \mathcal{T}_V(\mathbf{c})$, $W^n(y^n|\mathbf{c}) = \exp\{-n[H(V|P) + D(V\|W|P)]\}$. Therefore,

$$\begin{aligned} W^n(\mathcal{C}_\epsilon|\mathbf{c}) &= \sum_{V \in \mathcal{W}_n(P)} [|\mathcal{C} \cap \mathcal{T}_V(\mathbf{c})| \bmod K] \cdot \exp\{-n[H(V|P) + D(V\|W|P)]\} \\ &\leq |\mathcal{W}_n(P)| \max_{V \in \mathcal{W}_n(P)} \exp\left\{-n\left[D(V\|W|P) + (H(V|P) - R'')^+\right]\right\} \\ &\leq (n+1)^d \exp[-nE''(R'', P)]. \end{aligned}$$

The stated properties of $E'(R', \cdot)$ and $E''(R'', \cdot)$ are easy to establish. □

Proof of Lemma 3.2: We will actually prove the following stronger statement: if $0 < \epsilon < 1$, $J > 1/\epsilon$, $S \geq 1$, and $(N-1)\exp\{-S \cdot D(\epsilon\|1/J)\} < 1$, then there exists an $N \times S$ array with entries from $[J]$, in which any two rows differ in more than $(1-\epsilon)S$ positions. This will prove the lemma because $D(p\|q) \geq 2(p-q)^2$ for all p, q in $[0,1]$ (see Lemma 12.6.1 on p. 300 of [3]), so that if $J \geq 2/\epsilon$ then $D(\epsilon\|1/J) \geq \epsilon^2/2$, and $(N-1)\exp\{-S \cdot D(\epsilon\|1/J)\} \leq N\exp(-S\epsilon^2/2)$.

Let the first row of the array be arbitrary. Then, choose a random second row, by picking each element independently and equiprobably from $[J]$. Since $\epsilon > 1/J$, the probability that the second row matches the first at least in ϵS positions is, by a Chernoff bound, no greater than $\exp\{-S \cdot D(\epsilon\|1/J)\} < 1$. Hence, there exists a "good" $2 \times S$ array.

In general, if there exists a "good" $L \times S$ array for some $L \geq 2$, and we pick an $(L+1)^{th}$ row randomly as above, then the probability that this row matches any of the other L rows at least in ϵS positions is bounded by $L \cdot \exp\{-S \cdot D(\epsilon\|1/J)\}$, which is < 1 if $L < N$. This proves the existence of a "good" $N \times S$ array. □

Proof of Lemma 4.1: Let (X^n, Y^n) be the random pair of input and output sequences when the n-step strategy F is used. Then,

$$\begin{aligned} W_F(x^n, y^n) &= Pr[(X^n, Y^n) = (x^n, y^n)] \\ &= \prod_{k=1}^{n} Pr\left[(X_k, Y_k) = (x_k, y_k) \,\middle|\, (X^{k-1}, Y^{k-1}) = (x^{k-1}, y^{k-1})\right] \end{aligned}$$

$$= \prod_{k=1}^{n} P_k^F(x_k|x^{k-1}, y^{k-1}) W(y_k|x_k)$$
$$= \left(\prod_{k=1}^{n} P_k^F(x_k|x^{k-1}, y^{k-1})\right) W^n(y^n|x^n). \tag{15}$$

Fix a pair (x, y). For $1 \leq k \leq n$, let $A_k = 1$ if $(X_k, Y_k) = (x, y)$, and 0 otherwise. Then, $E[A_k|X^{k-1}, Y^{k-1}] = P_k^F(x|X^{k-1}, Y^{k-1})W(y|x)$. Thus, if $\tilde{A}_k = A_k - E[A_k|X^{k-1}, Y^{k-1}]$, then

$$\sum_{k=1}^{n} \tilde{A}_k = N(x, y|X^n, Y^n) - nP_{X^n,Y^n}^F(x)W(y|x). \tag{16}$$

It can be verified easily that $\mathrm{Var}(\tilde{A}_k) \leq W(y|x)$, and that the $\tilde{A}_k$'s are pairwise uncorrelated. Hence, by Chebyshev's inequality,

$$Pr\left[\left|\sum_{k=1}^{n} \tilde{A}_k\right| > n^{3/4}\sqrt{W(y|x)}\right] \leq n^{-1/2}. \tag{17}$$

By (16), (17), and a union bound over all (x, y), we have $1 - W_F(\mathcal{E}(F)) \leq dn^{-1/2}$. This proves (10).

Suppose $(x^n, y^n) \in \mathcal{E}(F)$ and y^n has type Q. Then, for any $y \in \mathcal{Y}$,

$$\begin{aligned}\left|Q(y) - P_{x^n,y^n}^F W(y)\right| &= n^{-1}\left|\sum_x \left[N(x, y|x^n, y^n) - nP_{x^n,y^n}^F(x)W(y|x)\right]\right| \\ &\leq n^{-1}\sum_x n^{3/4}\sqrt{W(y|x)} \\ &\leq |\mathcal{X}|n^{-1/4}. \end{aligned} \tag{18}$$

Now, if P_1 and P_2 are probability distributions on a finite set $\mathcal{Z}$, and $|P_1(z) - P_2(z)| \leq \beta$ for all $z \in \mathcal{Z}$, then $|H(P_1) - H(P_2)| \leq |\mathcal{Z}|\sqrt{\beta}$. For,

$$\begin{aligned}|H(P_1) - H(P_2)| &\leq \sum_{z\in\mathcal{Z}} |-P_1(z)\log P_1(z) + P_2(z)\log P_2(z)| \\ &\leq |\mathcal{Z}|G(\beta),\end{aligned}$$

where $G(\beta) = \max|-x\log x + y\log y|$ subject to $0 \leq x, y \leq 1$ and $|x - y| \leq \beta$. By elementary calculus methods, it can be shown that $G(\beta) = -\beta\log\beta$ if $0 \leq \beta \leq 1/e$, and $G(\beta) = 1/e$ if $\beta > 1/e$. Since $-\beta\log\beta \leq \sqrt{\beta}$ on $[0, 1/e]$, and $1/e \leq \sqrt{\beta}$ on $[1/e, \infty)$, we have $G(\beta) \leq \sqrt{\beta}$ for all $\beta \geq 0$, which proves the claim made earlier. Applying this result to (18), we have

$$\begin{aligned}\left|H(Q) - H(P_{x^n,y^n}^F W)\right| &\leq |\mathcal{Y}|\sqrt{|\mathcal{X}|n^{-1/4}} \\ &\leq dn^{-1/8}.\end{aligned}$$

Since $\log Q^n(y^n) = -nH(Q)$, Part (a) of the Lemma is proved. Next,

$$\begin{aligned}
&\left|\log W^n(y^n|x^n) + nH(W|P^F_{x^n,y^n})\right| \\
&\quad = \left|\sum_{x,y}\left[N(x,y|x^n,y^n) - nP^F_{x^n,y^n}(x)W(y|x)\right]\log W(y|x)\right| \\
&\quad \le \sum_{x,y} n^{3/4}\sqrt{W(y|x)}\,|\log W(y|x)| \\
&\quad \le dn^{7/8}, \qquad (19)
\end{aligned}$$

since $|\sqrt{z}\log z| \le 1$ if $0 \le z \le 1$, and $n^{3/4} \le n^{7/8}$. This proves Part (b). Part (c) is an obvious consequence of Parts (a) and (b). □

Proof of Lemma 4.2: Let F be any n-step strategy, $\mathcal{D} \subseteq \mathcal{Y}^n$, and $\mathcal{D}' = \mathcal{D} \cap \mathcal{B}^c_\alpha(F)$. Then,

$$\begin{aligned}
Q_F(\mathcal{D}') &= W_F(\mathcal{X}^n \times \mathcal{D}') \\
&\le W_F([\mathcal{X}^n \times \mathcal{D}'] \cap \mathcal{E}(F)) + dn^{-1/2} \\
&= \sum_{Q\in\mathcal{P}_n(\mathcal{Y})}\sum_{y^n\in\mathcal{D}'\cap\mathcal{T}_Q} W_F(\mathcal{E}_{y^n}(F)\times\{y^n\}) \;+\; dn^{-1/2}, \qquad (20)
\end{aligned}$$

where $\mathcal{P}_n(\mathcal{Y})$ is the set of n-types on $\mathcal{Y}$, and $\mathcal{T}_Q$ the set of y^n with type Q (the inequality above is by (10)). Now, for any $y^n \in \mathcal{D}' \cap \mathcal{T}_Q$,

$$\begin{aligned}
&W_F(\mathcal{E}_{y^n}(F)\times\{y^n\}) \\
&\quad = \sum_{x^n\in\mathcal{E}_{y^n}(F)}\left(\prod_{k=1}^n P^F_k(x_k|x^{k-1},y^{k-1})\right)W^n(y^n|x^n) \\
&\quad \le \sum_{x^n\in\mathcal{E}_{y^n}(F)}\left(\prod_{k=1}^n P^F_k(x_k|x^{k-1},y^{k-1})\right)Q^n(y^n)\exp\{n\alpha + 2dn^{7/8}\} \\
&\quad \le Q^n(y^n)\exp\{n\alpha + 2dn^{7/8}\}. \qquad (21)
\end{aligned}$$

Here, the first equality is by (15), and the first inequality is by Part (c) of Lemma 4.1, together with the fact that if $y^n \in \mathcal{B}^c_\alpha(F)$ then $I(P^F_{x^n,y^n}; W) \le \alpha$ for all $x^n \in \mathcal{E}_{y^n}(F)$. From (20) and (21),

$$\begin{aligned}
Q_F(\mathcal{D}') &\le \sum_{Q\in\mathcal{P}_n(\mathcal{Y})} Q^n(\mathcal{D}'\cap\mathcal{T}_Q)\exp\{n\alpha + 2dn^{7/8}\} \;+\; dn^{-1/2} \\
&\le \sum_{Q\in\mathcal{P}_n(\mathcal{Y})} Q^n(\mathcal{D})\exp\{n\alpha + 2dn^{7/8}\} \;+\; dn^{-1/2}.
\end{aligned}$$

Thus, if $F_1, \ldots, F_M$ are n-step strategies, and $\mathcal{D}_1, \ldots, \mathcal{D}_M$ are pairwise disjoint subsets of $\mathcal{Y}^n$,

$$\begin{aligned}
&\frac{1}{M}\sum_{m=1}^{M} Q_{F_m}\left(\mathcal{D}_m \cap \mathcal{B}_\alpha^c(F_m)\right) \\
&\quad \le \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{Q\in\mathcal{P}_n(\mathcal{Y})} Q^n(\mathcal{D}_m)\exp\{n\alpha + 2dn^{7/8}\} + dn^{-1/2}\right] \\
&\quad = \frac{\exp\{n\alpha + 2dn^{7/8}\}}{M}\sum_{Q\in\mathcal{P}_n(\mathcal{Y})} Q^n\left(\bigcup_{m=1}^{M}\mathcal{D}_m\right) + dn^{-1/2} \\
&\quad \le \frac{\exp\{n\alpha + 2dn^{7/8}\}}{M}(n+1)^d + dn^{-1/2},
\end{aligned}$$

since $|\mathcal{P}_n(\mathcal{Y})| \le (n+1)^d$. This proves (11).

Next, we will prove that

$$\left|\bigcup_F \mathcal{B}_\alpha(F)\right| \le (n+1)^d \exp\left\{n\left[\max_{P:I(P;W)\ge\alpha} H(PW)\right] + dn^{7/8}\right\}, \qquad (22)$$

where the union is over all n-step strategies F. This will obviously imply (12). Now, if $y^n \in \mathcal{B}_\alpha(F)$, then there exists $x^n \in \mathcal{E}_{y^n}(F)$ such that $I(P^F_{x^n,y^n}; W) > \alpha$. Therefore, if y^n has type Q, then

$$\begin{aligned}
n^{-1}\log Q^n(y^n) &\ge -H(P^F_{x^n,y^n}W) - dn^{-1/8} \\
&\ge -\max_{P:I(P;W)\ge\alpha} H(PW) - dn^{-1/8},
\end{aligned}$$

where the first inequality is by Part (a) of Lemma 4.1. Hence,

$$\left|\mathcal{T}_Q \cap \bigcup_F \mathcal{B}_\alpha(F)\right| \le \exp\left\{n\left[\max_{P:I(P;W)\ge\alpha} H(PW)\right] + dn^{7/8}\right\}. \qquad (23)$$

Since $|\bigcup_F \mathcal{B}_\alpha(F)| = \sum_Q |\mathcal{T}_Q \cap \bigcup_F \mathcal{B}_\alpha(F)|$, and the number of types Q is at most $(n+1)^d$, (22) is proved.

Finally, we will prove that if F is a deterministic n-step strategy, then

$$|\mathcal{B}_\alpha(F)| \le \exp\left\{n\left[\max_{P:I(P;W)\ge\alpha} H(W|P)\right] + dn^{7/8}\right\}. \qquad (24)$$

From this and the disjointness of the sets $\mathcal{D}_m$, (13) will follow. Suppose $F(f) = 1$. Then, $(x^n, y^n) \in \mathcal{E}(F)$ implies that $x^n = f(y^n)$ (because of the condition

$W_F(x^n, y^n) > 0$ in the definition of $\mathcal{E}(F)$). Thus, if $y^n \in \mathcal{B}_\alpha(F)$, we must have $I(P^F_{f(y^n),y^n}; W) > \alpha$, so that

$$\begin{aligned} Q_F(y^n) &= W^n(y^n|f(y^n)) \\ &\geq \exp\left\{-nH(W|P^F_{f(y^n),y^n}) - dn^{7/8}\right\} \\ &\geq \exp\left\{-n\left[\max_{P:I(P;W)\geq\alpha} H(W|P)\right] - dn^{7/8}\right\}. \end{aligned} \tag{25}$$

Here, the first inequality is by Part (b) of Lemma 4.1. From (25), we obviously have (24). □

References

[1] R. Ahlswede and G. Dueck. Identification in the presence of feedback - a discovery of new capacity formulas. *IEEE Transactions on Information Theory*, Vol. 35(No. 1), January 1989.

[2] R. Ahlswede and G. Dueck. Identification via channels. *IEEE Transactions on Information Theory*, Vol. 35(No. 1), January 1989.

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

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

[5] T.S. Han and S. Verdú. New results in the theory of identification via channels. *IEEE Transactions on Information Theory*, Vol. 38(No. 1), January 1992.

[6] J.H.B. Kemperman. Strong converses for a general memoryless channel with feedback. *Trans. 6th Prague Conf. Information Theory, Stat. Dec. Fct's and Rand. Proc.*, 1973.