The Rate–Distortion Region for Multiple Descriptions without Excess Rate

Rudolf Ahlswede

Abstract

During recent years there has been strong interest in a certain source coding problem, which some authors call the "problem of multiple descriptions". Old and new wringing techniques enable us to establish a single–letter characterization of the rate–distrotion region in the case of no excess rate for the joint description.

1 The Result

Since the origin of the problem of multiple descriptiona and motivations for its study have already been described in an extensive literature [1]–[9], we present our result immediately. It goes considerably beyond those of [17], where the reader also will find a detailed discussion of previously known results.

We are given the following.

1) A sequence $(X_t)_{t=1}^{\infty}$ of independent and identically distributed random variables with values in a finite set $\mathcal{X}$, that is, a discrete memoryless source (DMS).

2) Three finite reconstruction spaces $\hat{\mathcal{X}}_0$, $\hat{\mathcal{X}}_1$, and $\hat{\mathcal{X}}_2$, together with associated per–letter distortion measures

$$d_i : \mathcal{X} \times \hat{\mathcal{X}}_i \to \mathbb{R}^+, \text{ for } i = 0, 1, 2.$$

With $\tilde{d}_i = \max\{d(x, \hat{x}_i) : (x, \hat{X}_i) \in \mathcal{X} \times \hat{\mathcal{X}}_i\}$, $i = 0, 1, 2$, define

$$\mathcal{B} = [0, \log|\mathcal{X}|]^2 \times \prod_{i=0}^{2} [0, \tilde{d}_i].$$

The quintuple $(R_1, R_2, D_0, D_1, D_2) \in \mathcal{B}$ is called (α, β)–*achievable*, if for all large n there exist description functions $f_i : \mathcal{X}^n \to \mathcal{F}_i$ with

$$R_i + \alpha \geq \text{ rate } (f_i), \ i = 1, 2 \tag{1.1}$$

and reconstruction functions $g_i : \mathcal{F}_i \to \hat{\mathcal{X}}_i$ $(i = 1, 2)$, $g_0 : \mathcal{F}_1 \times \mathcal{F}_2 \to \hat{\mathcal{X}}_0$ such that for $\hat{X}_i^n = (\hat{X}_{i1}, \dots, \hat{X}_{in}) = \big(f_i(X^n)\big)$, $i = 1, 2$, and for $\hat{X}_0^n = (\hat{X}_{01}, \dots, \hat{X}_{on}) = \big(f_1(X^n), f_2(X^n)\big)$

$$(D_i + \beta)n \geq E \sum_{t=1}^{n} d_i(X_t, \hat{X}_{it}), \;\; i = 0, 1, 2. \tag{1.2}$$

$(R_1, R_2, D_0, D_1, D_2)$ is called *achievable*, if it is (α, β)achievable for all *positive* α and β.

The set $\mathcal{Q}$ of achievable quintuples is a compact subset of the five–dimensional Euclidean space.

The problem of multiple descriptions consists in finding a single–letter characterization for $\mathcal{Q}$. This is a difficult task and the literature shows that even very special cases are hard to handle. Even the uniform binary source is far from being completely understood. Therefore one studies first projections and cross sections of $\mathcal{Q}$.

Here we concentrate on a case of *no excess rate at* D_0,

$$R_1 + R_2 = R(D_0), \tag{1.3}$$

where R denotes Shannon's rate–distortion function. For any fixed $D_0 \geq d_0$, where

$$d_0 = \sum_{x \in \mathcal{X}} P_X(x) \min_{\hat{x}_0 \in \hat{\mathcal{X}}_0} d_0(x, \hat{x}_0),$$

define the set

$$\mathcal{Q}(D_0) = \big\{(R_1, R_2, D_0, D_1, D_2) \in \mathcal{Q} : \;\text{(1.3) holds)}\big\}. \tag{1.4}$$

El Gamal and Cover [6] have shown that $(R_1, R_2, D_0, D_1, D_2) \in \mathcal{B}$ is achievable, if there exist random variables $\hat{X}_1, \hat{X}_2, \hat{X}_0$ jointly distributed with a generic source variable X such that:

1) $R_i \geq I(X \wedge \hat{X}_i)$ for $i = 1, 2$;

2) $R_1 + R_2 \geq I(X \wedge \hat{X}_0\hat{X}_1\hat{X}_2) + I(\hat{X}_1 \wedge \hat{X}_2)$;

3) $D_i \geq Ed_i(X, \hat{X}_i)$ for $i = 0, 1, 2$.

If $\mathcal{C}$ denotes the set of these achievable quintuples, then time sharing yields that $\mathrm{conv}(\mathcal{C})$ is also an achievable region.

In a related paper, "New results in binary multiple descriptions", Z. Zhang and T. Berger have shown that in the case of an excess rate even for the very special binary symmetric source with the Hamming distortion measure, $\mathrm{conv}(\mathcal{C}) \neq \mathcal{Q}$.

For this source they have also established a bound on $\mathcal{Q}(D_0)$, which is not tight. It does, however, imply tightness of Witsenhausen's hyperbola bound. The proof for this result is entirely different and seemingly more complicated than the one we gave in [17]. In that paper we also proved tightness of the hyperbola bound for 0–distortion and arbitrarily small excess rate.

Now let $\mathcal{C}^*$ denote the subset $\mathcal{C}$ containing all those quintuples satisfying 1)–3) for random variables $\hat{X}_0, \hat{X}_1, \hat{X}_2$, with the additional property that $\hat{X}_1$ and $\hat{X}_2$ are independent. Since $I(\hat{X}_1 \wedge \hat{X}_2) = 0$ now, 2) can be replaced by

2*) $R_1 + R_2 \geq I(X \wedge \hat{X}_0, \hat{X}_1, \hat{X}_2)$.

Clearly, $\text{conv}(\mathcal{C}^*) \subset \text{conv}(\mathcal{C})$ is an achievable region and therefore for every $D_0 \geq d_0$ the region

$$\text{conv}(\mathcal{C}^*)(D_0) = \big\{(R_1, R_2, D_0, D_1, D_2 0) \in \text{conv}(\mathcal{C}^*) : R_1 + R_2 = R(D_0)\big\} \tag{1.5}$$

is achievable.

Theorem 1. *Let $(X_t)_{t=1}^{\infty}$ be a DMS. A solution of the multiple dexcription problem in case of no excess rate at D_0 is*

$$\mathcal{Q}(D_0) = conv(\mathcal{C}^*)(D_0), \ \ for\ D_0 \geq d_0.$$

It is clear from the foregoing explanations that only $\mathcal{Q}(D_0) \subset \text{conv}(\mathcal{C}^*)(D_0)$ remains to be proved.

In Section 2 we state elementary continuity properties of $\mathcal{C}^*$, and in Section 3 we discuss the ideas and present the techniques on which the proof in Section 4 is based.

2 Preliminaries

The elementary continuity properties to be stated in Propositions 1 and 2 are used in the proof of the theorem.

A. Definition of $\mathcal{F}^*$

Let $\mathcal{F}^*$ be the set of vectors $\overline{I} = (I_1, I_2, \dots, I_6)$ for which there exist independent random variables $\hat{X}_1, \hat{X}_2$ and a random variable $\hat{X}_0$ such that $I_i = I(X \wedge \hat{X}_i)$ $(i = 1, 2)$, $I_3 = I(X \wedge \hat{X}_0 \hat{X}_1 \hat{X}_2)$, $I_{i+4} = d_i(X, \hat{X}_i)$ $(i = 0, 1, 2)$. Clearly, $\mathcal{F}^*$ is compact in the topology $\mathcal{T}_0$, which is induced in the six–dimensional Euclidean space by its metric "dist".

For a finite set $\mathcal{Z}$ let $P(\mathcal{Z})$ be the set of probability distributions on $\mathcal{Z}$.

Since $\mathcal{F}^*$ depends on $P_X \in \mathcal{P}(\mathcal{X})$, we can interpret it as a map

$$\mathcal{F}^* : \mathcal{P}(\mathcal{X}) \to \text{comp}(\mathcal{B}^*), \tag{2.1}$$

where $\text{comp}(\mathcal{B}^*)$ is the set of compact subsets of $\mathcal{B}^* = [0, \log|\mathcal{X}|]^3 \times \prod_{i=0}^{2}[0, \tilde{d}_i]$.

B. Total–Variation, I–Divergence, and Hausdorff Metric

In $\mathcal{P}(\mathcal{Z})$ we have a topology $\mathcal{T}_1$ induced by the total variation

$$\|P-P'\|_1 = \sum_{z\in Z} |P(z)-P'(z)|,\ \ P,P'\in\mathcal{P}(\mathcal{Z}). \tag{2.2}$$

$\big(\mathcal{P}(\mathcal{Z}),\mathcal{T}_1\big)$ is compact.

Pinsker's inequality [16] relates the I–divergence to the $\|\cdot\|_1$–norm:

$$D(P\|P') \geq c\|P-P'\|_1^2,\ \ c \text{ a constant}. \tag{2.3}$$

In $\mathrm{comp}(\mathcal{B}^*)$ one has the Hausdorff metric

$$m(A,B) = \max\left(\max_{a\in A}\min_{b\in B}\mathrm{dist}(a,b), \max_{b\in B}\min_{a\in A}\mathrm{dist}(a,b)\right),\ \text{for } A,B\in\mathrm{comp}(\mathcal{B}^*). \tag{2.4}$$

The metric m induces a topology $\mathcal{T}_2$ and $\big(\mathrm{comp}(\mathcal{B}^*),\mathcal{T}_2\big)$ is a complete metric space.

Proposition 1. $\mathcal{F}^*: \mathcal{P}(\mathcal{X}) \to \mathrm{comp}(\mathcal{B}^*)$ is uniformly $(\|\cdot\|_1, m)$–continuous.

Proof. Write

$$\begin{aligned}
p(\hat{x}_0\hat{x}_1\hat{x}_2x) &= p(\hat{x}_1)p(\hat{x}_2)p(\hat{x}_0|\hat{x}_1\hat{x}_2)p(x|\hat{x}_0\hat{x}_1\hat{x}_2),\\
q(z) &= p(\hat{x}_1)p(\hat{x}_2)p(\hat{x}_0|\hat{x}_1\hat{x}_2),\\
z &= (\hat{x}_0\hat{x}_1\hat{x}_2),\\
w(x|z) &= p(x|\hat{x}_0\hat{x}_1\hat{x}_2).
\end{aligned}$$

Thus

$$\sum_z q(z)w(x|z) = p(x). \tag{2.5}$$

We now keep q fixed (and therefore also $\hat{X}_1,\hat{X}_2$ independent) and change P_X slightly, that is, we replace $p(x)$ by $p'(x)$ such that

$$\begin{aligned}
&p'(x) = p(x)+\varepsilon(x)\\
&|\varepsilon(x)| \leq \varepsilon,\ \ x\in\mathcal{X},
\end{aligned} \tag{2.6}$$

where $\varepsilon \leq 2^{-1}\mathbf{p}$ and $\mathbf{p} = \min\{p(x) : p(x) > 0\}$.

We show that there is a stochastic matrix w' with the following properties:

1) $\sum_z q(z)w'(x|z) = p'(x)$;
2) $\sum_{x\in\mathcal{X}} |w'(x|z) - w(x|z) \leq \varepsilon(1+2\mathbf{p}^{-1})|\mathcal{X}|$, for all $z\in Z$.

Since information and average distortion quantities are $\|\cdot\|_1$–continuous in the distributions, this implies the continuity of $\mathcal{F}^*$. Uniform continuity follows since $\mathcal{P}(\mathcal{X})$ is compact.

Now start with the identity

$$\sum_{z:q(z)>0} q(z)\big(w(x|z)+\varepsilon(x)\big) = p(x)+\varepsilon(x) = p'(x).$$

Since $w(x|z)+\varepsilon(x)$ need not be stochastic, we choose now instead

$$\begin{aligned} w'(x|z) &= \begin{cases} (1-\lambda)\big(w(x|z)+\varepsilon(x)\big)+\lambda p'(x), q(z)>0 \\ w(x|z), q(z)=0 \end{cases} \\ \lambda &= 2\varepsilon(2\varepsilon+\mathbf{p})^{-1} \end{aligned} \tag{2.7}$$

which obviously satisfies 1). Since

$$\begin{aligned} |w'(x|z)-w(x|z)| &\le |-\lambda w(x|z)+(1-\lambda)\varepsilon(x)+\lambda p'(x)| \\ &= |-\lambda w(x|z)+\varepsilon(x)+\lambda p(x)| \\ &= \varepsilon+\lambda = \varepsilon\big(1+2(2\varepsilon+\mathbf{p})^{-1}\big) \\ &\le \varepsilon(1+2\mathbf{p}^{-1}), \end{aligned}$$

2) also holds. It remains to be seen that w' is stochastic.

Since $\sum_x w'(x|z)=1$ we have to verify only

$$0 \le w'(x|z) \le 1, \text{ for } q(z)>0. \tag{2.8}$$

Case $p(x)=0$: Here $p'=\varepsilon(x)$ and $w(x|z)=0$. Therefore $w'(x|z)=(1-\lambda)\varepsilon(x)+\lambda p'(x)$ and thus (2.8) holds.

Case $p(x)=1$: Here $w(x|z)=1$ if $q(z)>0$; $-\frac{1}{2}\le\varepsilon(x)\le 0$; and therefore $0\le w'(x|z)=(1-\lambda)+(1-\lambda)\varepsilon(x)+\lambda p'(x)\le 1$.

Case $0<p(x)<1$: Since $|\varepsilon(x)|\le 2^{-1}\mathbf{p}$, here $p'(x)\ge 2^{-1}\mathbf{p}$ and $2^{-1}\ge\mathbf{p}$

$$\begin{aligned} (1-\lambda)\big(w(x|z)+\varepsilon(x)\big)+\lambda p'(x) &= \mathbf{p}(2\varepsilon+\mathbf{p})^{-1}\big(w(x|z)+\varepsilon(x)\big)+2\varepsilon(2\varepsilon+\mathbf{p})^{-1}p'(x) \\ &\ge (2\varepsilon+\mathbf{p})^{-1}\big(\mathbf{p}\varepsilon(x)+2\varepsilon p'(x)\big) \\ &\ge (2\varepsilon+\mathbf{p})^{-1}\big(\mathbf{p}\varepsilon(x)+\varepsilon\mathbf{p}\big)\ge 0. \end{aligned}$$

Finally,

$$(2\varepsilon+\mathbf{p})^{-1}\big(\mathbf{p}w(x|z)+\mathbf{p}\varepsilon(x)+2\varepsilon p'(x)\big)\le(2\varepsilon+\mathbf{p})^{-1}\big(\mathbf{p}+\mathbf{p}\varepsilon(x)+2\varepsilon p'(x)\big).$$

This is smaller than 1 if $\mathbf{p}\varepsilon(x)+2\varepsilon p'(x)\le 2\varepsilon$. Since $\varepsilon(x)\le\varepsilon$ and $\mathbf{p}2^{-1}+p'(x)\le\mathbf{p}2^{-1}+(1-\mathbf{p})+2^{-1}\mathbf{p}\le 1$, this is the case.

Replacement of the independence property in the definition of $\mathcal{F}^*$ by

$$I(\hat{X}_1 \wedge \hat{X}_2) \le \eta,\ \ \eta > 0 \tag{2.9}$$

leads to a set $\mathcal{F}_\eta, \mathcal{F}_\eta \supset \mathcal{F}^*$, respectively, a map

$$\mathcal{F}_\eta : \mathcal{P}(\mathcal{X}) \to \mathrm{comp}(\mathcal{B}^*). \tag{2.10}$$

Proposition 2. $\mathcal{F}_\eta(P_X) \overset{m}{\to} \mathcal{F}^*(P_X)$ uniformly in $P_X \in \mathcal{P}(\mathcal{X})$ as $\eta \to 0$.

Proof. Consider an $\overline{I} = \overline{I}(P_{X\hat{X}_0\hat{X}_1\hat{X}_2}) \in \mathcal{F}_\eta(P_X)$. Write $P_{X\hat{X}_o\hat{X}_1\hat{X}_2}(x\hat{x}_0\hat{x}_1\hat{x}_2) = w(x\hat{x}_0|\hat{x}_1\hat{x}_2)$ $P_{\hat{X}_1\hat{X}_2}(\hat{x}_1\hat{x}_2)$ and define random variables $X', \hat{X}_0', \hat{X}_2'$ with the distribution

$$P_{X'\hat{X}_0'\hat{X}_1'\hat{X}_2'}(x\hat{x}_0\hat{x}_1\hat{x}_2) = w(x\hat{x}_0|\hat{x}_1\hat{x}_2)P_{\hat{X}_1}(\hat{x}_1)P_{\hat{X}_2}(\hat{x}_2).$$

Since $I(\hat{X}_1 \wedge \hat{X}_2) = D(P_{\hat{X}_1\hat{X}_2} \| P_{\hat{X}_1} \times P_{\hat{X}_2}) \le \eta$, by Pinsker's inequality

$$\|P_{\hat{X}_1\hat{X}_2} - P_{\hat{X}_1} \times P_{\hat{X}_2}\| \le \sqrt{c^{-1}\eta} \tag{2.11}$$

and therefore also

$$\|P_{X\hat{X}_0\hat{X}_1\hat{X}_2} - P_{X'\hat{X}_0'\hat{X}_1'\hat{X}_2'}\|_1 \le \sqrt{c^{-1}\eta}, \tag{2.12}$$

$$\|P_X - P_{X'}\|_1 \le \sqrt{c^{-1}\eta}. \tag{2.13}$$

Since information quantities and average distortions are uniformly continuous on $\big(\mathcal{P}(\mathcal{X} \times \mathcal{X}_0 \times \mathcal{X}_1 \times \mathcal{X}_2), \| \cdot \|_1\big)$, (2.12) implies, for all $\overline{I} \in \mathcal{F}_\eta(P_X)$,

$$|I_i(P_{X\hat{X}_0\hat{X}_1\hat{X}_2}) - I_i(P_{X'\hat{X}_0'\hat{X}_1'\hat{X}_2'})| \le c_1(\eta), \tag{2.14}$$

where $c_1(\eta) \to 0 (\eta \to 0)$.

Also by Proposition 1 and (2.13)

$$m\big(\mathcal{F}^*(P_X), \mathcal{F}^*(P_{X'})\big) \le c_2(\eta), \tag{2.15}$$

where $c_2(\eta) \to 0 (\eta \to 0)$.

Finally, since $\overline{I}(P_{X'\hat{X}_0'\hat{X}_1'\hat{X}_2'}) \in \mathcal{F}^*(P_{X'})$ and since $\mathcal{F}_\eta(P_X) \supset \mathcal{F}^*(P_X)$, we conclude by (2.14) and (2.15) that

$$m\big(\mathcal{F}_\eta(P_X), \mathcal{F}^*(P_X)\big) \le c_1(\eta) + c_2(\eta).$$

The same bound holds for all $P_X \in \mathcal{P}(\mathcal{X})$ and the proof is complete.

3 Wringing Techniques

Suppose that $(R_1, R_2, D_0, D_1, D_2)$ is (α, β)–achievable and that f_i $(i = 1, 2)$, and g_i and $\hat{X}_i^n$, $i = 0, 1, 2$, satisfy (1.1) and (1.2). Then we have for $n \geq n(\alpha, \beta)$

$$\begin{aligned} n(R_i + \alpha) &\geq n \cdot \mathrm{rate}(f_i) \geq H(f_i) \geq H(\hat{X}_i^n) \\ &\geq I(X^n \wedge \hat{X}_i^n) \geq \sum_{t=1}^n I(X_t \wedge \hat{X}_{it}) \end{aligned} \tag{3.1}$$

(since the X_i are independent) for $i = 1, 2$.

$$\begin{aligned} n(R_1 + \alpha + R_2 + \alpha) &\geq H(f_1) + H(f_2) \geq H(f_1 f_2) \\ &\geq H(\hat{X}_0^n \hat{X}_1^n \hat{X}_2^n) \geq I(X^n \wedge \hat{X}_0^n \hat{X}_1^n \hat{X}_2^n) \\ &\geq \sum_{t=1}^n I(X_t \wedge \hat{X}_{0t} \hat{X}_{1t} \hat{X}_{2t}), \end{aligned} \tag{3.2}$$

$$\begin{aligned} n(D_i + \beta) &\geq E d_i(X_1^n, \hat{X}_i^n) \\ &\geq \sum_{t=1}^n E d_i(X_t, \hat{X}_{it}), \text{ for } i = 0, 1, 2. \end{aligned} \tag{3.3}$$

Since α, β can be made arbitrarily small and $\mathrm{conv}(\mathcal{C}^*)(D_0)$ is compact, these relations would imply $\mathcal{Q}(D_0) \subset \mathrm{conv}(\mathcal{C}^*)(D_0)$, if the random variables X_{1t}, X_{2t} were independent for every $t = 1, 2, \ldots, n$.

This is obviously not the case. However, we show next that if there is no excess rate at D_0, then $\hat{X}_1^n$ and $\hat{X}_2^n$ can be made to be close to independence. More precisely, for an arbitrary small positive number ε we can achieve

$$I(\hat{X}_1^n \wedge \hat{X}_2^n) \leq \varepsilon n \tag{3.4}$$

for n large enough. For this first notice that by the continuity of the rate–distortion function R we can choose α and β so small that

$$R(D_0) - R(D_0 + \beta) + 2\alpha \leq \varepsilon. \tag{3.5}$$

Since $R_1 + R_2 = R(D_0)$, (3.2) implies

$$n\big(R(D_0) + 2\alpha\big) \geq H(f_1) + H(f_2). \tag{3.6}$$

Furthermore, by the converse to the source coding theorem

$$\mathrm{rate}(\hat{X}_0^n) \geq \frac{1}{n} H(\hat{X}_0^n) \geq R(D_0 + \beta)$$

and since $\hat{X}_0$ is a function of f_1 and f_2

$$H(f_1, f_2) \geq nR(D_0 + \beta). \tag{3.7}$$

The inequalities (3.6), (3.7), and (3.5) imply

$$I(f_1 \wedge f_2) \leq n\big(R(D_0) + 2\alpha - R(D_0 + \beta)\big) \leq \varepsilon n,$$

and thus (3.4) follows by the data processing inequality.

Now notice that only independence properties of the component variables $\hat{X}_{1t}, \hat{X}_{2t}$ are actually relevant. If they were arbitrarily close to being independent, then by Proposition 2 this would be good enough. However, this is not the case.

We overcome this difficulty with the help of what we called in [15] a *wringing technique.* These techniques originated with [13] and were first made a powerful instrument for proving strong converse by Dueck [14]. (For a systematic analysis and improved forms see also [15]).

Wringing Lemma 1 [14]. *Let $U^n = (U_1, \dots, U_n)$, $V^n = (V_1, \dots, V_n)$ be random variables with values in $\mathcal{U}^n$ resp. $\mathcal{V}^n$.*

If $I(U^n \wedge V^n) \leq \sigma$ then for any $\delta > 0$ there exist $t_1, \dots, t_k \in \{1, \dots, n\}$, $k \leq \sigma/\delta$, such that

$$I(U_t \wedge V_t | U_{t_1} V_{t_1}, \dots, V_{t_k} V_{t_k}) \leq \delta, \text{ for } t = 1, 2, \dots, n.$$

This lemma says that *conditional on suitable, relatively few component variables* all corresponding component variables are much closer to independence than are U^n, V^n. The lemma was used by Dueck in [14] for strong converse proofs with σ, δ constant, but it turns out to be good enough also in the present rate–distortion situation, where $\sigma = \varepsilon n$ (ε arbitrarily small).

In addition, we need Wringing Lemma 2 to follow, which gives a wringing technique for distributions rather than just for information quantities (for another result in this direction compare also the quite different [15, Lemma ?]. It ensures that for "most" values of the conditioning variable the conditional distributions of the component variables X_t are close to the distribution of X_t for "most t".

Wringing Lemma 2. *Let $(X_t)_{t=1}^{\infty}$ be a DMS that has a joint distribution with a random variable Y taking values in a finite set $\mathcal{Y}$. Define for $\rho \in (0,1)$*

$$\mathcal{Y}(\rho, t) = \left\{ y \in \mathcal{Y} : \sum_x |P_{X_t|Y}(x|y) - P_{X_t}(x)| \leq \rho \right\}.$$

Then for any $\gamma > 0$ and every n there exists a set of components $\mathcal{N}(n,\gamma) \subset \{1, 2, \dots, n\}$ such that

$$|\mathcal{N}(n,\gamma)| \geq \gamma(1+\gamma)^{-1} n \tag{3.8a}$$

$$H(X_t|Y) \geq H(X_t) - (1+\gamma)n^{-1}H(Y), \text{ for all } t \in \mathcal{N}(n,\gamma) \tag{3.8b}$$

$$P_Y\big(\mathcal{Y}(\rho,t)\big) \geq 1 - \rho^{-1}\sqrt{(1+\gamma)c^{-1}n^{-1}H(Y)} \text{ for all } t \in \mathcal{N}(n,\gamma). \tag{3.8c}$$

Here c is the constant in Pinsker's inequality [16]. (independently, Csiszár, Kemperman, Kullback have shown that an optimal choice of c is $(2ln2)^{-1}$ (see [10]).

Proof. Choose $\mathcal{N}(n,\gamma) = \big\{t : H(X_t|Y) \geq H(X) - (1+\gamma)n^{-1}H(Y)\big\}$. Then (2.8b) holds by definition. Since

$$\sum_{t=1}^{n} H(X_t|Y) \geq H(X^n|Y) = nH(X) - H(Y)$$
$$H(X) \geq H(X_t|Y),$$

we also have

$$|\mathcal{N}(n,\gamma)|H(X)+(n-|\mathcal{N}(n,\gamma)|)\cdot\big(H(X)-(1+\gamma)n^{-1}H(Y)\big) \geq \sum_{t=1}^{n} H(X_t|Y) \geq nH(X)-H(Y).$$

Therefore, $|\mathcal{N}(n,\gamma)|(1+\gamma)n^{-1}H(Y) \geq \gamma H(Y)$ and thus (3.8a). Finally, we derive (3.8c) from (3.8b) by Pinsker's inequality.

We can restate (3.8b) as

$$(1+\gamma)n^{-1}H(Y) \geq I(X_t \wedge Y) = D(P_{X,Y}\|P_{X_t} \times P_Y),$$

and thus by (2.3)

$$\begin{aligned}(1+\gamma)n^{-1}H(Y) &\geq c\left(\sum_{x,y} |P_{X,Y}(x,y) - P_{X_t}(x)P_Y(y)|\right)^2 \\ &= c\left(\sum_{y}\left(\sum_{x} |P_{X_t|Y}(x|y) - P_{X_t}(x)|\right)P_Y(y)\right)^2,\end{aligned}$$

or, equivalently

$$\sqrt{(1+\gamma)n^{-1}H(Y)c^{-1}} \geq \sum_{y}\left(\sum_{x} |P_{X_t|Y}(x|y) - P_{X_t}(x)|\right)P_Y(y).$$

Application of Chebyshev's inequality gives (3.8c).

4 Proof of Theorem

We can assume that (3.1)–(3.4) hold. With the choices $U^n = \hat{X}_1^n$, $V^n = \hat{X}_2^n$, $\sigma = \varepsilon n$, and $\delta = \varepsilon l$, $1 \le l \le n$, Wringing Lemma 1 guarantees the existence of $t_1, t_2, \dots, t_k \in \{1, 2, \dots, n\}$ such that for $t = 1, 2, \dots, n$

$$I(\hat{X}_{1t} \wedge \hat{X}_{2t} | \hat{X}_{1t_1} \hat{X}_{2t_2} \dots \hat{X}_{1t_k} \hat{X}_{2t_k}) \le \varepsilon l, \;\; k < n l^{-1}. \tag{4.1}$$

We use the abbreviation $Y = \hat{X}_{1t} \hat{X}_{2t} \dots \hat{X}_{1t_k} \hat{X}_{2t_k}$ and $\mathcal{Y}$ for the range of this random variable. Since for $i = 1, 2$ $H(\hat{X}_i^n) + H(Y) \ge I(X^n \wedge \hat{X}_i^n Y)$, we have

$$\begin{aligned} n(R_i + \alpha) &\ge H(\hat{X}_i^n) \ge I(X^n \wedge \hat{X}_i^n Y) - H(Y) \\ &\ge \sum_{t=1}^n I(X_t \wedge \hat{X}_i^n Y) - H(Y) \\ &\ge \sum_{t=1}^n I(X_t \wedge \hat{X}_{it} Y) - H(Y), \end{aligned}$$

and thus

$$R_i + \alpha \ge \frac{1}{n} \sum_{t=1}^n I(X_t \wedge \hat{X}_{it} | Y) - \frac{1}{n} H(Y). \tag{4.2}$$

Similarly, since $H(\hat{X}_0^n \hat{X}_1^n \hat{X}_2^n) + H(Y) \ge I(X^n \wedge \hat{X}_0^n \hat{X}_1^n \hat{X}_2^n Y)$ from (3.2),

$$\begin{aligned} n(R_1 + \alpha + R_2 + \alpha) &\ge I(X^n \wedge \hat{X}_0^n \hat{X}_1^n \hat{X}_2^n Y) - H(Y) \\ &\ge \sum_{t=1}^n I(X_t \wedge \hat{X}_{0t} \hat{X}_{1t} \hat{X}_{2t} Y) - H(Y), \end{aligned}$$

and thus

$$R_1 + R_2 + 2\alpha \ge \frac{1}{n} \sum_{t=1}^n I(X_t \wedge \hat{X}_{0t} \hat{X}_{1t} \hat{X}_{2t} | Y) - \frac{1}{n} H(Y). \tag{4.3}$$

Define now for every $y \in \mathcal{Y}$ random variables $\hat{X}_{it}(y)$ and $X_t(y)$, $i = 0, 1, 2$, $1 \le t \le n$, with the distributions $\Pr\big(\hat{X}_{it}(y) = \hat{x}_i\big) = \Pr(\hat{X}_{it} = \hat{x}_i | Y = y)$, $\Pr\big(X_t(y) = x\big) = \Pr(X_t = x | Y = y)$ $(\hat{x}_i \in \hat{\mathcal{X}}_i, x \in \mathcal{X})$. Clearly, from (3.3) for $i = 0, 1, 2$

$$D_i + \beta \ge \frac{1}{n} E d_i(X^n, \hat{X}_i^n) = \sum_y \frac{1}{n} \sum_{t=1}^n E d_i\big(X_i(y), \hat{X}_{it}(y)\big) P_Y(y). \tag{4.4}$$

With the abbreviation $a = \log|\hat{\mathcal{X}}_1||\hat{X}_2|$ we have $H(Y) \leq k \cdot a$ and since $kn^{-1} < l^{-1}$ also

$$n^{-1}H(Y) \leq al^{-1}. \tag{4.5}$$

Now we inspect for a moment (4.1)–(4.5) and observe that, for instance, with the choice $l^{-1} = \varepsilon^{1/2}$ ($\varepsilon \to 0$) the desired result would follow if $X_t(y)$, $y \in \mathcal{Y}$ $1 \leq t \leq n$, had the same distribution as X. This is not the case. However, in the sense made precise in Wringing Lemma 2 it is approximately true, and by the continuity property of $\mathcal{C}$ described in Proposition 2 the desired result will follow.

Choose now $l = \varepsilon^{-2/5}$ and define

$$\mathcal{Y}'(\varepsilon, t) = \big\{y \in \mathcal{Y} : I(\hat{X}_{1t} \wedge \hat{X}_{2t}|Y = y) \leq \varepsilon^{1/5}\big\} \tag{4.6a}$$

$$\mathcal{Y}(\rho, t) = \left\{y \in \mathcal{Y} : \sum_x |P_{X_t|Y}(x|y) - P_{X_t}(x)| \leq \rho\right\} \tag{4.6b}$$

$$\mathcal{Y}(\varepsilon, \rho, t) = \mathcal{Y}'(\varepsilon, t) \cap \mathcal{Y}(\rho, t). \tag{4.6c}$$

By (4.1) and Chebyshev's inequality

$$P_Y\big(\mathcal{Y}'(\varepsilon, t)\big) \geq 1 - \varepsilon^{2/5}, \text{ for } t = 1, 2, \dots, n. \tag{4.7}$$

By Wringing Lemma 2 there is a set $\mathcal{N}(n, \gamma)$ with

$$|\mathcal{N}(n, \gamma)| \geq \gamma(1 + \gamma)^{-1}n \tag{4.8}$$

such that

$$P_Y\big(\mathcal{Y}(\rho, t)\big) \geq 1 - \rho^{-1}\sqrt{(1 + \gamma)c^{-1}n^{-1}H(Y)},\ t \in \mathcal{N}(n, \gamma). \tag{4.9}$$

Choose now $1 + \gamma = c \cdot a^{-1}\varepsilon^{-1/5}$ and insert this and (4.5) in (4.9). Thus

$$P_Y\big(\mathcal{Y}(\rho, t)\big) \geq 1 - \rho^{-1}\varepsilon^{1/10},\ t \in \mathcal{N}(n, \gamma), \tag{4.10}$$

and also from (4.8)

$$|\mathcal{N}\big(n, \gamma(\varepsilon)\big) \geq (1 - c^{-1}a\varepsilon^{1/5})n \triangleq \nu(\varepsilon)n. \tag{4.11}$$

Since $\varepsilon, \rho < 1$ (4.7) and (4.10) imply

$$P_Y\big(\mathcal{Y}(\varepsilon, \rho, t)\big) \geq 1 - 2\rho^{-1}\varepsilon^{1/10} \triangleq \mu(\varepsilon, \rho),\ t \in \mathcal{N}\big(n, \gamma(\varepsilon)\big). \tag{4.12}$$

Using (4.11), (4.12), and $n^{-1}H(Y) \leq a\varepsilon^{2/5}$, we derive from (4.2), (4.3), and (4.4)

$$
\begin{aligned}
(R_i + \alpha + a\varepsilon^{2/5})\nu(\varepsilon)^{-1}\mu(\varepsilon,\rho)^{-1} \geq {} & |\mathcal{N}\big(n,\gamma(\varepsilon)\big)|^{-1} \\
& \cdot \sum_{t\in\mathcal{N}\big(n,\gamma(\varepsilon)\big)} \sum_{y\in\mathcal{Y}(\varepsilon,\rho,t)} I\big(X_t(y)\wedge \hat{X}_{it}(y)\big)\cdot \frac{P_Y(y)}{P_Y\big(\mathcal{Y}(\varepsilon,\rho,t)\big)}
\end{aligned} \tag{4.13}
$$

$$
\begin{aligned}
(R_1 + R_2 + 2\alpha + a\varepsilon^{2/5})\nu(\varepsilon)^{-1}\mu(\varepsilon,\rho)^{-1} \geq {} & |\mathcal{N}\big(n,\gamma(\varepsilon)\big)|^{-1} \\
& \cdot \sum_{t\in\mathcal{N}\big(n,\gamma(\varepsilon)\big)} \sum_{y\in\mathcal{Y}(\varepsilon,\rho,t)} I\big(X_t(y)\wedge \hat{X}_{0t}(y)\cdot \hat{X}_{1t}(y)X_{2t}(y)\big) \\
& \cdot \frac{P_Y(y)}{P_Y\big(\mathcal{Y}(\varepsilon,\rho,t)\big)}
\end{aligned} \tag{4.13b}
$$

$$
\begin{aligned}
(D_i + \beta)\nu(\varepsilon)^{-1}\mu(\varepsilon,\rho)^{-1} \geq {} & |\mathcal{N}(n,\gamma)|^{-1} \\
& \cdot \sum_{t\in\mathcal{N}\big(n,\gamma(\varepsilon)\big)} \sum_{y\in\mathcal{Y}(\varepsilon,\rho,t)} d_i\big(X_t(y),\hat{X}_{it}(y)\big) \\
& \cdot \frac{P_Y(y)}{P_Y\big(\mathcal{Y}(\varepsilon\rho,t)\big)}.
\end{aligned} \tag{4.13c}
$$

The astute reader can see already that (4.13a)–(4.13c) imply the desired relation $(R_1,R_2,D_0,D_1,D_2)\in \text{conv}(\mathcal{C}^*)(D_0)$. A formal argument may proceed as follows. We can choose $\alpha=\alpha(\varepsilon)$, $\beta=\beta(\varepsilon\to 0)$, $(\varepsilon\to 0)$ so that for $n\geq n\big(\alpha(\varepsilon)\big)$ (3.1)–(3.4) and thus (4.13a)–(4.13c) hold.

Now since $\nu(\varepsilon)\to 1$, $\mu(\varepsilon,\rho)\to 1$, $\varepsilon\to 0$, the quantities on the left sides in (4.13a)–(4.13c) converge to $R_i (i=1,2)$, R_1+R_2, and D_i, $i=0,1,2$, as $\varepsilon\to 0$. Next recall that for $y\in\mathcal{Y}(\varepsilon,\rho,t)$, $t\in\mathcal{N}\big(n,\gamma(\varepsilon)\big)$ we have $\|P_{X_t(y)}-P_X\|_1\leq\rho$ and by Proposition 1 $\mathcal{F}^*(P_{X_t(y)})\to\mathcal{F}^*(P_X)$, $\rho\to 0$ uniformly in y,t,ε.

Furthermore, by Proposition 2 and (3.13a) uniformly in y,t,ρ

$$\mathcal{F}_{\varepsilon^{1/5}}(P_{X_t(y)})\to\mathcal{F}^*(P_{X_t(y)}),\ \varepsilon\to 0.$$

Thus for any $\delta\to 0$ we can choose ρ and $\varepsilon(\delta)$ so small that

$$m\big(\mathcal{F}_{\varepsilon^{1/5}}(P_{X_t(y)}),\mathcal{F}^*(P_X)\big)\leq\delta$$

for $y\in\mathcal{Y}(\varepsilon,\rho,t)$, $t\in\mathcal{N}\big(n,\gamma(\varepsilon)\big)$, $n\geq n\big(\alpha(\varepsilon),\beta(\varepsilon)\big)$.

By definition of the Hausdorff metric, therefore, for

$$\overline{\mathcal{F}}\triangleq\sum_{y\in\mathcal{Y}(\varepsilon,\rho,t)}\sum_{t\in\mathcal{N}\big(n,\gamma(\varepsilon)\big)} I_{\varepsilon^{1/5}}(P_{X_t(y)}),$$

we get $m\big(\overline{\mathcal{F}},\mathcal{F}^*(P_X)\big)\leq\delta$ and thus also

$$m\big(\text{conv}(\overline{\mathcal{F}}),\text{conv}(\mathcal{F}^*)\big)\leq\delta. \tag{4.14}$$

Since the vector having components as specified by the right sides of (4.13a)–(4.13c) is in conv($\overline{\mathcal{F}}$), letting δ and $\varepsilon(\delta)$, $\rho(\delta)$ tend to zero (4.14) implies the existence of a cluster point $I \in \mathrm{conv}(\mathcal{F}^*)$ of the right side vectors. Therefore,

$$\begin{aligned} R_i &\le I_i, \;\; i=1,2 \\ R_1 + R_2 &\le I_3 \\ D_i &\le I_{i+4}, \;\; i=0,1,2 \end{aligned}$$

and thus $(R_1, R_2, D_0, D_1, D_2) \in \mathrm{conv}(\mathcal{C}^*)(D_0)$.

5 References

[1] H.S. Witsenhausen, "On source networks with minimal breakdown degradation", Bell Syst. Tech. J., vol. 59, no. 6, 1083–1087, July–Aug. 1980.

[2] J.K. Wolf, A.D. Wyner, and J. Ziv, "Source coding for multiple descriptions", Bell Syst. Tech. J., vol. 59, no. 8, 1417–1426, Oct. 1980.

[3] H.S. Witsenhausen, "On team guessing with independent information", Math. Oper. Res., vol. 6, no. 2, 293–304, 1981.

[4] H.S. Witsenhausen, "Team guessing with lacunary information", Math. Oper. Res., vol. 8, no. 1, 110–121, Feb. 1983.

[5] H.S. Witzenhausen and A.D. Wyner, "Source coding for multiple descriptions II: A binary source", Bell. Syst. Tech. J., vol. 60, no. 10, 2281–2292, 1981.

[6] A.A. El Gamal and T.M. Cover, "Achievable rates for multiple descriptions", IEEE Trans. Inform. Theory, vol. IT–28, no. 6, 851–857, Nov. 1982.

[7] T. Berger and Z. Zhang, "Minimum breakdown degradation in binary source coding", IEEE Trans. Inform. Theory, vol. IT–29, no., 6, Nov. 1983.

[8] L. Ozarow, "On a source–coding problem with two channels and three receivers", Bell. Syst. Tech. J., vol. 59, no. 10, 1909–1921, 1980.

[9] T. Berger and Z. Zhang, "A theorem about diversity coding", in preparation.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1982.

[11] R. Ahlswede, "Coloring hypergraphs: a new approach to multi–user source coding", part I, J. Comb. Inform. Syst. Sci., vol. 4, no. 1, 76–115, 1979; part II ibid., vol. 5, no. 3, 220–268, 1980.

[12] R. Ahlswede and G. Dueck, "Good codes can be produced by a feq permutations", IEEE Trans. Inform. Theory, vol. IT–28, no. 3, 430–443, 1982.

[13] R. Ahlswede, "On two–way communication channels and a problem by Zarankiewic", in Trans. Sixth Prague Conf. Information Theory, Sept. 1971.

[14] G. Dueck, "The strong converse to the coding theorem for the multiple access channel", J. Comb. Inform. Syst. Sci., vol. 6, no. 3, 187–196, 1981.

[15] R. Ahlswede, "An elementary proof of the strong converse theorem for the multiple–access channel", J. Comb. Inform. Syst. Sci., vol. 7, no. 3, 216–230, 1982.

[16] M.S. Pinsker, "Information and information stability of random variables and processes", (in Russian), vol. 7, 1960: English translation: San Francisco: Holden–Day, 1964.

[17] R. Ahlswede, "The distortion region of a binary source for multiple descriptions without excess rate", IEEE Trans. Inform. Theory.