Extremal Properties of Rate–Distortion Functions

Rudolf Ahlswede

Abstract

At a recent workshop (Hakone, Japan, July 1–2, 1988) I. Csiszár stated that inspite of strong efforts the following question had been waiting for an answer for many years. It is true that for fixed distortion level Δ the rate–distortion function $R(P,\Delta)$ has in the distribution P no local maxima with value different from the global maximum? We show that in general the answer is negative. However, the answer is positive for Hamming distortion measures. Moreover, there R is Schur–concave.

1 Basic Concepts and Auxiliary Results

Let $(X_t)_{t=1}^{\infty}$ be a discrete memoryless source (DMS), that is, a sequence of independent and identically distributed random variables (RV's) with values in a finite set $\mathcal{X}$. We are also given a finite reconstruction space $\hat{\mathcal{X}}$ and a per letter distortion measure

$$d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+. \tag{1.1}$$

For a function F defined on a product space $\mathcal{Y}^n$ we use the notation

$$\text{rate}(F) = \frac{1}{n} \log_2 \|F\|, \ \|F\| = \text{ the cardinality of the range of } F. \tag{1.2}$$

For an encoding function $f_n : \mathcal{X}^n \to \hat{\mathcal{X}}^n$ we consider the reproduction

$$\hat{X}^n = (\hat{X}_1, \dots, \hat{X}_n) = f_n(X^n) \tag{1.3}$$

and the average distortion $\text{dist}(f_n)$, defined by

$$\text{dist}(f_n) = \mathbb{E}\frac{1}{n}\sum_{t=1}^{n} d(X_t, \hat{X}_t). \tag{1.4}$$

ρ is said to be an ε–achievable rate for distortion level Δ, if for all large n there are encoding functions f_n with

$$\text{rate}(f_n) \le \rho, \ \text{dist}(f_n) \le \Delta + \varepsilon. \tag{1.5}$$

Thus

$$R(\Delta) = \inf\left(\bigcap_{\varepsilon>0}\{\rho : \rho \text{ is } \varepsilon\text{–achievable for } \Delta\}\right) \tag{1.6}$$

is the smallest rate which for arbitrarily small ε is ε–achievable for Δ. This quantity depends on the generic distribution P_X of the source. Thus we get a function

$$R : \mathcal{P}(\mathcal{X}) \times \mathbb{R}_+ \to \mathbb{R}_+,\ \mathcal{P}(\mathcal{X}) = \text{ set of probability distribution (PD's) on } \mathcal{X}, \tag{1.7}$$

of two variables, which was introduced by Shannon [1] and is called rate–distortion function. Using time–sharing one readily verifies with the forgoing definitions that $R(P,\Delta)$ is convex in Δ for every $P \in \mathcal{P}(\mathcal{X})$. A remarkable characterization of R is due to Shannon [1].

Rate–Distortion Theorem. *Let $(X,\hat{X})$ denote a pair of RV's with values in $\mathcal{X} \times \hat{\mathcal{X}}$ and $I(X \wedge \hat{X})$ their mutual information. Then*

$$R(P,\Delta) = \min_{(X,\hat{X}):P_X=P,\mathbb{E}d(X,\hat{X})\leq\Delta} I(X,\hat{X}) \textit{ for every } \Delta \in \mathbb{R}_+ \textit{ and } P \in \mathcal{P}(\mathcal{X}).$$

This formula has been used to derive several analytical properties of R, mainly as function of Δ. Whereas those properties can be found in many textbooks, the properties of R as function of P are not as well understood.

An exception seems to be the following basic result of Gallager (Theorem 9.5.1 and its extension on pages 467–469 in [2]), which he derives from Shannon's formula. For its formulation we need some notation.

Let the elements in $\mathcal{X}$ be labelled such that for fixed distribution Q

$$Q(0) \geq Q(1) \geq \cdots \geq Q(\alpha-1), \sum_{i=0}^{\alpha-1} Q(i) = 1. \tag{1.8}$$

For the fixed distortion level Δ let m be the *smallest* non–negative integer with

$$m\ Q(m) + \sum_{k=m+1}^{\alpha-1} Q(K) \leq \Delta \leq (m-1)Q(m-1) + \sum_{k=m}^{\alpha-1} Q(k). \tag{1.9}$$

The conventional understanding of (1.9) is for $m=0$

$$1 - Q(0) \leq \Delta \leq 1 \tag{1.10}$$

and for $m = \alpha$

$$0 \leq \Delta \leq (\alpha-1)Q(\alpha-1). \tag{1.11}$$

We also use the abbreviations

$$S_m = \sum_{k=0}^{m-1} Q(k),\ T_m = 1 - S_m \tag{1.12}$$

and the quantities

$$Q_m = \left(\frac{Q(0)}{S_m}, \dots, \frac{Q(m-1)}{S_m}\right),\ \Delta^* = \frac{\Delta - T_m}{S_m}. \tag{1.13}$$

Theorem G. *For the Hamming distortion measure the rate–distortion function $R : \mathcal{P}(\mathcal{X}) \times [0,1] \to \mathbb{R}_+$ is of the form*

$$R(Q,\Delta) = \begin{cases} 0 & \textit{for } m = 0 \\ S_m\big[H(Q_m) - h(\Delta^*) - \Delta^* \log(m-1)\big] & \textit{otherwise} \\ H(Q) - h(\Delta) - \Delta \log(\alpha - 1) & \textit{for } m = \alpha. \end{cases}$$

By the result in case $m = \alpha$ one has

$$R(Q,0) = H(Q) \tag{1.14}$$

and one may therefore expect that R has similar convexity properties than H. Already the case $\mathcal{X} = \{0,1\}$ is instructive. One readily verifies that

$$R(Q,\Delta) = \begin{cases} 0 & \text{if } Q(1) \le \Delta \le 1 \\ h\big(Q(1)\big) - h(\Delta) & \text{if } 0 \le \Delta \le Q(1) \le \frac{1}{2}. \end{cases} \tag{1.15}$$

Therefore R is *not concave* in Q. However, our main result says that in the Hamming case even for general alphabets another basic convexity of entropy generalizes to R, namely the property *Schur–concavity.*

We recall this concept. For the distributions $P = \big(P(0), \dots, P(\alpha-1)\big)$ and $Q = \big(Q(0), \dots, Q(\alpha-1)\big)$ we say that P *majorizes* Q and write $P \succ Q$, if

$$\sum_{i=0}^{k} P[i] \ge \sum_{i=0}^{k} Q[i] \text{ for } k = 0, 1, \dots, \alpha - 1, \tag{1.16}$$

where $P[i]$ (resp. $Q[i]$) is the i–th largest component of P (resp. Q). A function $\varphi : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$ is Schur–concave, if

$$P \succ Q \text{ implies } \varphi(P) \le \varphi(Q), \tag{1.17}$$

and it is *strictly* Schur–concave, if

$$P \succ Q \text{ and } P \ne Q \text{ imply } \varphi(P) < \varphi(Q). \tag{1.18}$$

2 The Key Ideas and a Basic Inequality

It was also mentioned by Csiszár that the truth of the main statement

(A) R has no local maxima in P with values different from the global maximum

would follow from the truth of anyone of the following statements:

(B) R is quasi–concave in P

(C) R is concave in P on $\mathcal{P}(\mathcal{X}, \Delta) = \{P \in \mathcal{P}(\mathcal{X}) : R(P, \Delta) > 0\}$.

Recall that quasi–concavity means that the level sets $\{P \in \mathcal{P}(\mathcal{X}) : R(P, \Delta) > \rho\}$ are convex for all $\rho \geq 0$, Clearly, (C) implies (B). In order to better understand (A) we studied first (B). We explain now our ideas to disprove (B). Those lead naturally to the counterexample to (A) in Section 4. To find distributions P_1 and P_2 with

$$R(P_i, \Delta) \geq \rho \text{ for } i = 1, 2; R\left(\frac{1}{2}P_1 + \frac{1}{2}P_2, \Delta\right) \leq \rho - \delta \tag{2.1}$$

we consider a new DMS with generic distribution

$$\overline{P} = \frac{1}{2}P_1 + \frac{1}{2}P_2. \tag{2.2}$$

Clearly, its n–variate distribution is

$$\overline{P}^n = \prod_1^n \overline{P}. \tag{2.3}$$

In order to get a link to sources involving P_1 and P_2 *we view* $\overline{P}^n$ *as an average over the arbitrarily varying source* (AVS) defined by the set of distributions $\{P_1, P_2\}$. They have been intensively studied in [4].

The set on n–variate distributions is $\mathcal{A}_n = \{P(\cdot|s^n) : s^n \in \{1, 2\}^n\}$, where

$$P(x^n|s^n) = \prod_{t=1}^n P_{s_t}(x_t) \text{ for } s^n = (s_1, \ldots, s_n) \in \{1, 2\}^n \text{ and } x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n. \tag{2.4}$$

For the distribution q on $\{1, 2\}$ with $q(1) = q(2)$ we clearly have for $q^n = \prod_1^n q$

$$\overline{P}(x^n) = \sum_{s^n} p(x^n|s^n)q^n(s^n). \tag{2.5}$$

In the light of this formula one can view the operation of $\overline{P}^n$ as producing with "high probability" a string s^n whose number $< 1|s^n >$ of 1's and number $< 2|s^n >$ of 2's is "approximately equal" to $\frac{n}{2}$.

Therefore $p(\cdot|s^n)$ is typically of the form $\prod_1^{n/2} P_1 \times \prod_1^{n/2} P_2$.

What is the rate–distortion function for such a non–stationary source? More generally, if λ is the fraction of P_1's in the product, the answer is in obvious notation

$$R(P_1, P_2, \lambda, \Delta) = \inf_{|\gamma|\leq\Delta} \big[\lambda R(P_1, \Delta+\gamma) + (1-\lambda) R(P_2, \Delta-\gamma)\big], \tag{2.6}$$

which is smaller than $\lambda R(P_1, \Delta) + (1-\lambda)R(P_2, \Delta)$.

The exact difference

$$V(P_1, P_2, \lambda, \Delta) = \lambda R(P_1, \Delta) + (1-\lambda) R(P_2, \Delta) - R(P_1, P_2, \lambda, \Delta) \tag{2.7}$$

depends on the slopes of $R(P_1, \Delta)$ and $R(P_2, \Delta)$ as functions of Δ. There is no reason why this difference cannot become arbitrary big by proper choices of P_1, P_2 and the distortion measure d. But then this fact can be exploited for $\overline{P}$ via the AVS by a simple trick. Let us consider only P_1, P_2 with disjoint supports. Then the encoder can identify s^n and he can inform the decoder with n bits, that is, at a rate 1. Quasi–concavity is then disproved if the difference mentioned can be made bigger than 1. Of course our idea to estimate the rate–distortion function via AVS's works for general convex combinations

$$\overline{P} = \sum_{i=1}^{\ell} \lambda_i P_i, \;\; \lambda = (\lambda_1, \dots, \lambda_\ell). \tag{2.8}$$

Viewing $\overline{P}$ as a λ–average over the AVS with generic distributions $\mathcal{A} = \{P_i : 1 \leq i \leq \ell\}$ for any $\eta > 0$ we can find by Chebyshev's inequality a $c(\eta) > 0$ and an $n(\eta)$ such that for $n > n(\eta)$ with a probability greater than $1 - e^{-c(\eta)n}$ the relative frequencies $\lambda_i + \varepsilon_i$ for P_i satisfy $|\varepsilon_i| \leq \eta$. Therefore, as $\eta \to 0$ and $n \to \infty$ we conclude

$$R(\overline{P}, \Delta) \geq \min\left\{\sum_{i=1}^{\ell} \lambda_i R(P_i, \Delta + \varepsilon_i) : \Sigma\varepsilon_i = 0, |\varepsilon_i| \leq \Delta\right\}. \tag{2.9}$$

Even more importantly, we have also an inequality upper bounding $R(\overline{P}, \Delta)$. For $\mathcal{X} = \{0, 1, \dots, \alpha-1\}$ let $\mathcal{X}(i) = \big\{(0,i), \dots, (\alpha-1, i)\big\}$ $(i = 1, \dots, \ell)$ be ℓ copies of $\mathcal{X}$, which are disjoint sets.

Define a distortion measure d^* on $\mathcal{X}^* = \bigcup_{i=1}^{\ell} \mathcal{X}(i)$ by

$$d^*\big((j,i), (j',i')\big) = d(j,j'), \tag{2.10}$$

where d is our original distortion measure on $\mathcal{X}$. Also define P^* on $\mathcal{X}^*$ by

$$P^*\big(j,i)\big) = \lambda_i P_i(j) \tag{2.11}$$

and the AVS $\mathcal{A}^* = \{P_i^* : 1 \leq i \leq \ell\}$, where $P_i^*\big((j,i)\big) = P_i(j)$ for $j = 0, \dots, \alpha-1$. As earlier the encoder can now record s^n. Here by the source coding theorem a rate $H(\lambda)$ suffices for a recording which is false with arbitrarily small probability.

Thus we have derived

$$R(P^*,\Delta) \le \min\left\{\sum_{i=1}^{\ell} \lambda_i R(P_i,\Delta+\varepsilon_i) : \Sigma\varepsilon_i = 0, |\varepsilon_i| \le \Delta\right\} + H(\lambda). \tag{2.12}$$

Identification of all points (j,i) $(i = 1,\dots,\ell\}$ with j transforms d^* into d and P^* into P. Therefore a code for the P^*–source gets transformed into a code for the $\overline{P}$–source with at least the same distortion level quaranteed. We summarize our findings.

Basic Inequality. For any distortion measure d and any $\Delta \in \mathbb{R}_+$ for $\overline{P} = \sum_{i=1}^{\ell} \lambda_i P_i$

(i) $R(\overline{P},\Delta) \ge \min\left\{\sum_{i=1}^{\ell} \lambda_i R(P_i,\Delta+\varepsilon_i) : \Sigma\varepsilon_i = 0, |\varepsilon_i| \le \Delta\right\}$

(ii) $R(\overline{P},\Delta) \le \min\left\{\sum_{i=1}^{\ell} \lambda_i R(P_i,\Delta+\varepsilon_i) : \Sigma\varepsilon_i = 0, |\varepsilon_i| \le \Delta\right\} + H(\lambda)$.

3 Schur–Concavity in the Hamming Case

We consider the case $\mathcal{X} = \hat{\mathcal{X}}, d = d_H$, where

$$d_H(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \ne \hat{x}. \end{cases} \tag{3.1}$$

For fixed $\Delta \in [0,1]$, $P \in \mathcal{P}(\mathcal{X})$ is a local maximum of R if for some neighborhood $\mathcal{U}(P)$

$$R(P,\Delta) \ge R(Q,\Delta) \text{ for all } Q \in \mathcal{U}(P). \tag{3.2}$$

The local maximum is called *strict*, if

$$R(P,\Delta) > R(Q,\Delta) \text{ for all } Q \in \mathcal{U}(P) - \{P\}. \tag{3.3}$$

Theorem 1. *In the Hamming case for no $\Delta \in [0,1]$ R has a strict local maximum other than the global maximum at $\left(\frac{1}{\alpha},\dots,\frac{1}{\alpha}\right)$.*

Proof: We don't use here Gallager's or even Shannon's result from Section 1. As in [4], Part II, Section 5 we exploit *symmetry.* Let Π be the symmetric group (the group of permutations) acting on $\mathcal{X} = \{0,1,\dots,\alpha-1\}$.
For $\pi \in \Pi$ and $P = \big(P(0),\dots,P(\alpha-1)\big) \in \mathcal{P}(\mathcal{X})$ we denote $\Big(P\big(\pi(0)\big),\dots,P\big(\pi(\alpha-1)\big)\Big)$ by πP. Obviously, since $d_H(x,\hat{x}) = d_H\big(\pi(x),\pi(\hat{x})\big)$

$$R(P,\Delta) = R(\pi P,\Delta) \text{ for } \Delta \in [0,1]. \tag{3.4}$$

Now for any $P' \in \operatorname{conv}\{\pi P : \pi \in \Pi\}$ we have

$$P' = \sum_{\pi \in \Pi} \lambda(\pi) \cdot \pi P, \;\; \lambda(\pi) \geq 0, \sum_{\pi \in \Pi} \lambda(\pi) = 1. \tag{3.5}$$

Consider the AVS with $\mathcal{A} = \{\pi P : \pi \in \Pi\}$ and the DMS with generic distribution P'.

From the basic inequality (i)

$$R(P', \Delta) \geq \min \left\{ \sum_{\pi \in \Pi} \lambda(\pi) \big(R(\pi P, \Delta + \gamma(\pi)\big) : \sum_{\pi} \gamma(\pi) = 0, |\gamma(\pi)| \leq \Delta \right\}. \tag{3.6}$$

However, by (3.4) the quantity at the right–hand side in (3.6) equals $R(P, \Delta)$.

Since for $P \neq \left(\frac{1}{\alpha}, \dots, \frac{1}{\alpha}\right)$ P' can be choosen arbitrarily close to but different from P, there cannot be a strict local maximum other than at $\left(\frac{1}{\alpha}, \dots, \frac{1}{\alpha}\right)$. Clearly, since for every $P\left(\frac{1}{\alpha}, \dots, \frac{1}{\alpha}\right) \in \operatorname{conv}\{\pi P : \pi \in \Pi\}$, the global maximum is assumed at $\left(\frac{1}{\alpha}, \dots, \frac{1}{\alpha}\right)$. A stronger conclusion can be drawn from our main result.

Theorem 2. *In the Hamming case*

(i) *R is Schur–concave on $\mathcal{P}(\mathcal{X})$ for every $\Delta \in [0,1]$.*

(ii) *$R(P, \Delta) = 0$ for all $P \in \mathcal{P}(\mathcal{X})$ and $\Delta \in \left[\frac{\alpha-1}{\alpha}, 1\right]$. For $\Delta \in \left[0, \frac{\alpha-1}{\alpha}\right]$ R has no local maximum in P other than the global maximum at $\left(\frac{1}{\alpha}, \dots, \frac{1}{\alpha}\right)$.*

(iii) *Statement (A) is true.*

Proof: (iii) obviously follows from (ii). The first part of (ii) follows from Theorem G. While we prove (i) we also prove the second statement in (ii).

Since $R(\pi P, \Delta) = R(P, \Delta)$ it suffices to consider $P, Q \in \mathcal{P}(Q)$ with $Q \succ P$ and $P(0) \geq P(1) \geq \dots \geq P(\alpha-1)$, $Q(0) \geq Q(1) \geq \dots \geq Q(\alpha-1)$. It is well–known (see [5]) that there exist finitely many *transfers* $U_{i,\varepsilon} : Q \to Q_{i,\varepsilon}$, where

$$Q_{i,\varepsilon} = \big(Q(0), Q(1), \dots, Q(i) - \varepsilon, Q(i+1) + \varepsilon, \dots\big) \tag{3.7}$$

$$Q(i) - \varepsilon \geq Q(i+1) + \varepsilon, \tag{3.8}$$

whose successive applications finally transform Q into P.

We have to show that such a transfer increases R on $\mathcal{P}(\mathcal{X}, \Delta)$, that is,

$$R(Q_{i,\varepsilon}, \Delta) \geq R(Q, \Delta) \text{ for } \varepsilon > 0. \tag{3.9}$$

Since $R(Q, \Delta) = 0$ for $Q \notin \mathcal{P}(\mathcal{X}, \Delta)$ and always $R(P, \Delta) \geq 0$, only Schur–concavity on $\mathcal{P}(\mathcal{X}, \Delta)$ remains to be shown. That is the m defined in (1.9) may have values between 2 and α. By the assumed minimality of m we have

$$m\, Q(m) + \sum_{k=m+1}^{\alpha-1} Q(k) \leq \Delta < (m-1) Q(m-1) + \sum_{i=m}^{\alpha-1} Q(k) \tag{3.10}$$

for these values of m.

Now notice that transfers for $i = 0, 1, \ldots, m-3, m+1, \ldots, \alpha - 1$ have no effect on either side in (3.10). They also leave S_m, T_m and Q_m invariant. By Theorem G *also R does not change.*

For $i = m-2$ (resp. $i = m$) the right (resp. left) hand side in (3.10) increases (resp. decreases), but as before S_m, T_m and finally R don't change.

Only the case $i = m-1$ needs serious consideration. We show that for transfers $U_{m-1,\beta}$ with

$$\Delta < (m-1)\big(Q(m-1) - \beta\big) + \big(Q(m) + \beta\big) + \sum_{i=m+1}^{\alpha-1} Q(k) \tag{3.11}$$

$R(U_{m-1,\beta}Q, \Delta) > R(Q, \Delta)$.

For this we calculate first the change in rate.

$$\begin{aligned} &R(U_{m-1,\beta}Q, \Delta) - R(Q, \Delta) = (S_m - \beta)\left[H\left(\frac{Q(0)}{S_m - \beta}, \ldots, \frac{Q(m-1) - \beta}{S_m - \beta}\right)\right. \\ &\quad \left. - h\left(\frac{\Delta - (T_m + \beta)}{S_m - \beta}\right) - \frac{\Delta - (T_m + \beta)}{S_m - \beta}\log(m-1)\right] \\ &\quad - S_m\left[H\left(\frac{Q_0}{S_m}, \ldots, \frac{Q(m-1)}{S_m}\right) - h\left(\frac{\Delta - T_m}{S_m}\right) - \frac{\Delta - T_m}{S_m}\log(m-1)\right] \\ &= +(S_m - \beta)\log(S_m - \beta) - S_m \log S_m - \big(Q(m-1) - \beta\big)\log\big(Q(m-1) - \beta\big) \\ &\quad + Q(m-1)\log Q(m-1) - (S_m - \beta)h\left(\frac{1 - \Delta}{S_m - \beta}\right) + S_m h\left(\frac{1 - \Delta}{S_m}\right) + \beta\log(m-1). \end{aligned}$$

It suffices now to show that

$$\omega(\beta) = \frac{d}{d\beta}\big[R(U_{m-1,\beta}Q, \Delta) - R(Q, \Delta)\big] > 0$$

for β satisfying (3.11).

We use that for $f(\beta) = (x - \beta)\log(x - \beta)$

$$f'(\beta) = -\log(x - \beta) - 1 \tag{3.12}$$

and that

$$h'(y) = \log\frac{1 - y}{y}. \tag{3.13}$$

Now

$$\omega(\beta) = -\log(S_m - \beta) + \log\big(Q(m-1) - \beta\big) + \log(m-1)$$

$$+ h\left(\frac{1-\Delta}{S_m - \beta}\right) - \frac{1-\Delta}{S_m-\beta}\left[\log\frac{S_m-\beta}{1-\Delta} - 1\right] = \log\frac{(m-1)\big(Q(m-1)-\beta\big)}{S_m-\beta}$$

$$- \frac{1-\Delta}{S_m-\beta}\log\frac{1-\Delta}{S_m-\beta} - \left(1 - \frac{1-\Delta}{S_m-\beta}\right)\log\left(1-\frac{1-\Delta}{S_m-\beta}\right) - \frac{1-\Delta}{S_m-\beta}\log\frac{S_m-\beta-(1-\Delta)}{1-\Delta}$$

$$= \log\frac{(m-1)\big(Q(m-1)-\beta\big)}{S_m-\beta} + \log\frac{S_m-\beta}{S_m-\beta-(1-\Delta)}$$

$$= \log\frac{(m-1)\big(Q(m-1)-\beta\big)}{S_m-\beta-(1-\Delta)}.$$

Since $\Delta < (m-1)\big(Q(m-1)-\beta\big)+T_m+\beta$, we have $(m-1)\big(Q(m-1)-\beta\big) > \Delta - T_m - \beta = S_m - \beta - (1-\Delta)$ and thus $\omega(\beta) > 0$.

Notice that there cannot be a local maximum at Q unless $m = \alpha$ and thus $A < (\alpha - 1)Q(\alpha-1)$.

In this case the transfer is not defined.

However, since $Q \succ P$ implies $P(\alpha-1) \geq Q(\alpha-1)$ we have also $\Delta < (\alpha-1)P(\alpha-1)$ and by Theorem G not only

$$R(P,\Delta) = H(P) - h(\Delta) - \Delta\log(\alpha-1)$$

but also

$$R(P,\Delta) = H(P) - h(\Delta) - \Delta\log(\alpha-1)$$

holds. The fact that the entropy function H is *strictly* Schur–concave completes the proof.

Remark: The proof shows that R stays constant for transfers $U_{i,\varepsilon}(i \neq m-1)$. Therefore (C) is not true even in the Hamming case. We have not decided upon (B). It is worth knowing that symmetry and quasi–concavity imply Schur–concavity, but not conversely (see [5], page 69).

4 The Counterexample

Let us consider a simple distortion measure τ, which does not equal the Hamming distortion measure.

$\mathcal{X}$ is partitioned into two sets $\mathcal{Y}$ and $\mathcal{Z}$ and $\hat{\mathcal{X}} = \hat{\mathcal{Y}} \cup \hat{\mathcal{Z}}$, where $\hat{\mathcal{Y}} = \mathcal{Y}$, $\hat{\mathcal{Z}} = \mathcal{Z}$.

$$\tau(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \text{ and } x,\hat{x} \in \mathcal{Y} \\ a & \text{if } x \neq \hat{x} \text{ and } x,\hat{x} \in \mathcal{Z} \\ b & \text{otherwise.} \end{cases} \tag{4.1}$$

Basic in our analysis are the uniform distributions Q_1 and Q_2 on $\mathcal{Y}$ resp. $\mathcal{Z}$, that is,

$$\begin{aligned} Q_1(x) &= |\mathcal{Y}|^{-1} \text{ for } x \in \mathcal{Y} \\ Q_2(x) &= |\mathcal{Z}|^{-1} \text{ for } x \in \mathcal{Z}. \end{aligned} \tag{4.2}$$

The reduction to these distributions proceeds as follows. Any distribution $P \in \mathcal{P}(\mathcal{X})$ can be written as $\lambda P_1 + (1-\lambda)P_2$, where

$$P_1(x) = P(x) \left(\sum_{x \in \mathcal{Y}} P(x) \right)^{-1} \quad \text{for } x \in \mathcal{Y} \tag{4.3}$$

$$P_2(x) = P(x) \left(\sum_{x \in \mathcal{Z}} P(x) \right)^{-1} \quad \text{for } x \in \mathcal{Z} \tag{4.4}$$

$$\lambda = \sum_{x \in \mathcal{Y}} P(x) \tag{4.5}$$

and therefore $P_1 \in \mathcal{P}(\mathcal{Y})$, $P_2 \in \mathcal{P}(\mathcal{Z})$.

Lemma 1. *For the distortion measure τ, any Δ and any $P = \lambda P_1 + (1-\lambda)P_2 \in \mathcal{P}(\mathcal{X})$, $P_1 \in \mathcal{P}(\mathcal{Y})$, $P_2 \in \mathcal{P}(\mathcal{Z})$ the inequality*

$$R\big(\lambda Q_1 + (1-\lambda)Q_2, \Delta\big) \geq R(P, \Delta) \tag{4.6}$$

holds.

Proof: We follow the idea in the proof of Theorem 1. The only difference is that now we have the groups of permutations Π on $\mathcal{Y}$ and Σ on $\mathcal{Z}$ and we define

$$(\pi, \sigma)P = \lambda \pi P_1 + (1-\lambda)\tau P_2. \tag{4.7}$$

Notice that

$$\sum_{(\pi,\sigma)\in\Pi\times\Sigma} (\pi,\sigma)P|\Pi|^{-1}|\Sigma|^{-1} = \lambda Q_1 + (1-\lambda)Q_2$$

and as explained in earlier sections (4.6) holds.

In order to show that there are two local maxima with different values we define three sets of distributions with the property that every continuous path from the first set to the third has to meet the second.

For $c \in \left(0, \frac{1}{2}\right)$ the sets are

$$\begin{aligned} &\mathcal{Q}_1^c = \big\{\lambda Q_1 + (1-\lambda)Q_2 : \lambda \geq 1-c\big\}, \\ &\mathcal{B}^c = \big\{\lambda P_1 + (1-\lambda)P_2 : P_1 \in P(\mathcal{Y}), P_2 \in \mathcal{P}(\mathcal{Z}), c \leq \lambda \leq 1-c\big\} \\ &\mathcal{Q}_2^c = \big\{\lambda Q_1 \in (1-\lambda)Q_2 : \lambda \leq c\big\}. \end{aligned}$$

We also use

$$\mathcal{Q}^c = \big\{ P \in \mathcal{B}^c \text{ of the form } \lambda Q_1 + (1-\lambda) Q_2 \big\}.$$

By Lemma 1 global maxima in $\mathcal{Q}_1^c$ and $\mathcal{Q}_2^c$ are local maxima in $\mathcal{P}(\mathcal{X})$ with the desired properties, if we can show that

$$\max_{P \in \mathcal{Q}^c} R(P, \Delta) < \max_{P \in \mathcal{Q}_1^c} R(P, \Delta) < \max_{P \in \mathcal{Q}_2^c} R(P, \Delta). \tag{4.8}$$

We consider the ranges

$$\gamma = |\mathcal{Z}| > \beta = |\mathcal{Y}|, a < 1 \ll b, \Delta \leq \frac{\beta - 1}{\beta}, \frac{\Delta}{a} \leq \frac{\gamma - 1}{\gamma}. \tag{4.9}$$

By Theorem G

$$\begin{aligned} R(Q_1, \Delta) &= \log \beta - h(\Delta) - \Delta \log(\beta - 1) \\ R(Q_2, \Delta) &= \log \gamma - h\left(\frac{\Delta}{a}\right) - \frac{\Delta}{a} \log(\gamma - 1). \end{aligned}$$

In order to simplify calculations we give up some freedom in the choice of parameters by requiring

$$\Delta + \frac{\Delta}{a} = 1. \tag{4.10}$$

Under this constraint we choose the other parameters so that $R(Q_1, \Delta)$ and $R(Q_2, \Delta)$ are approximately equal, that is,

$$\log \beta - \Delta \log(\beta - 1) \sim \log \gamma - \frac{\Delta}{a} \log(\gamma - 1)$$

or

$$\beta(\beta - 1)^{\Delta} \sim \gamma(\gamma - 1)^{-(1-\Delta)}. \tag{4.11}$$

Choose now

$$\gamma = \beta^3, \Delta = \frac{1}{4} + \varepsilon, a = \frac{1+\varepsilon}{3-\varepsilon}. \tag{4.12}$$

Thus

$$\beta(\beta - 1)^{\Delta} \sim \beta^{\frac{3}{4} - \varepsilon} \tag{4.13}$$

$$\gamma(\gamma - 1)^{-(1-\Delta)} \sim \beta^{\frac{3}{4} + 3\varepsilon}. \tag{4.14}$$

Application of our basic inequality (ii) to $P = \lambda Q_1 + (1-\lambda)Q_2$ yields

$$R(P,\Delta) \leq \min_{|\delta|\leq\Delta} \big(\lambda R(Q_1, \Delta+\delta) + (1-\lambda)R(Q_2, \Delta-\delta)\big) + \log 2$$

and for $\lambda = \frac{1}{2}$, $\Delta = \frac{1}{4}$, $a = \frac{1}{3}$

$$R\left(P, \frac{1}{4}\right) \leq 1 + \frac{1}{2}\left(R\left(Q_1, \frac{1}{6}+\eta\right) + R\left(Q_2, \frac{1}{3}-\eta\right)\right),$$

where η is introduced to insure $\left(\frac{1}{3}-\eta\right)a^{-1} \leq \frac{\gamma-1}{\gamma}$. Since $\frac{\gamma-1}{\gamma}$ goes to 1 very fast η is ignorably small.

Just notice that

$$\begin{aligned}
(\log\beta)^{-1} R\left(Q_1, \frac{1}{4}\right) &\to \frac{1}{3} \text{ as } \beta\to\infty \\
(\log\beta)^{-1} R\left(Q_2, \frac{1}{4}\right) &\to \frac{3}{4} \text{ as } \beta\to\infty \\
(\log\beta)^{-1} R\left(Q_1, \frac{1}{6}+\eta\right) &\to \frac{5}{6}-\eta \text{ as } \beta\to\infty \\
(\log\beta)^{-1} R\left(Q_2, \frac{1}{3}-\eta\right) &\to 9\eta \text{ as } \beta\to\infty
\end{aligned}$$

and thus

$$(\log\beta)^{-1} R\left(P, \frac{1}{4}\right) \to\leq \frac{1}{2}\left(\frac{5}{6}-\eta+9\eta\right) = \frac{5}{12}+4\eta < \frac{3}{4} \text{ as } \beta\to\infty.$$

For c sufficiently close to $\frac{1}{2}$

$$\max_{P\in\mathcal{Q}^c} R\left(P, \frac{1}{4}\right) \leq \frac{1}{2} < \frac{3}{4}.$$

If by chance we should have

$$\max_{P\in\mathcal{Q}_1^c} R\left(P, \frac{1}{4}\right) = \max_{P\in\mathcal{Q}_2^c} R\left(P, \frac{1}{4}\right)$$

by continuity of R in all parameters a slight change in them would make the maxima different.

5 A Consequence for Error Exponents

Our results seem to be relevant to the theory of universal source coding. We present now an immediate consequence in another direction. Marton [3] has investigated and found the error exponent $e(P,\rho,\Delta)$ for the following refined source coding problem:

for encoding functions $f_n : \mathcal{X}^n \to \hat{\mathcal{X}}^n$ with $\lim\limits_{n\to\infty} \mathrm{rate}(f_n) = \rho > R(P,\Delta)$ what is the value of

$$e(P,\rho,\Delta) = -\underline{\lim}_{n\to\infty} \frac{1}{n} \min_{f_n} \log \mathrm{Prob}\big(d\big(x^n, f_n(x^n)\big) > \Delta\big)? \tag{5.1}$$

Her answer is

Theorem M

$$e(P,\rho,\Delta) = \begin{cases} \min\limits_{Q:R(Q,\Delta)>\rho} D(Q\|P), & \text{if } R(Q,\Delta) > \rho \text{ for some } Q \\ 0 & \text{otherwise.} \end{cases}$$

Since $D(Q\|P)$ is continuous in P and $R(Q,\Delta)$ is continuous in Δ $e(P,\rho,\Delta)$ is continuous in P and in Δ.

Csiszár mentioned his interest in the question whether the exponent is also continuous in ρ.

He also mentioned the following observation.

Lemma *$e(P,\rho,\Delta)$ is continuous in ρ for all P,Δ exactly if (A) holds for R.*

Proof: If (A) does not hold, then for some Δ there are two local maxima at P_1 and P_2, say with

$$R(P_1,\Delta) > R(P_2,\Delta).$$

For $P = P_2$ and $\rho = R(P_2,\Delta) - \varepsilon$ we have

$$\min_{Q:R(Q,\Delta)>\rho} D(Q\|P_2) = D(P_2\|P_2) = 0.$$

However, if $\varepsilon = 0$, then there is no Q in the neighbourhood of P_2 with $R(Q,\Delta) > \rho$, but there is one far away in the neighbourhood of P_1. The exponent is discontinuous at ρ. Conversely, if (A) holds no such jump can occur. This result, Theorem 2 and the example give valuable information about e.

Corollary 2. *Marton's exponent $e(P,\rho,\Delta)$ is continuous in ρ for Hamming distortion measures. For general distortion measures it can jump.*

References

[1] C.E. Shannon, “Coding theorems for a discrete source with a fidelity criterion”, IRE Nat. Conv. Rec. Part 4, 142–163, 1959.

[2] R.G. Gallager, “Information Theory and Reliable Communication”, J. Wiley & Sons., Inc. New York 1968.

[3] K. Marton, “Error exponent for source coding with a fidelity criterion”, IEEE Trans. Inform. Theory 20, 197–199, 1974.

[4] R. Ahlswede, “Coloring hypergraphs: A new approach to multi–user source coding”, Part I, J. Combin. Inform. System Sci., Vol. 4, 76–115, 1979; Part II, ibid. Vol. 5, 220–268, 1980.

[5] A.W. Marshall, I. Olkin, “Inequalities: Theory of Majorization and its Applications”, Vol. 143 in Mathematics in Science and Engineering, Academic 1979.