Block Symmetry in Discrete Memoryless Channels

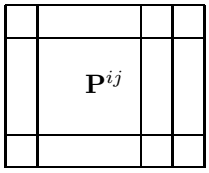
Jakob Bøje Pedersen
Frederiksborg Gymnasium
Hillerød
Denmark.
e-mail: jakobbp@worldonline.dk

Flemming Topsøe[1]
Department of Mathematics
University of Copenhagen
Denmark.
e-mail: topsoe@math.ku.dk

Abstract — **Notions of *block symmetry* for discrete, memoryless channels are introduced. The results deal with capacity and optimal distributions and appear to be simple and natural ones which somehow were overlooked or not considered in the early development of information theory.**

I. Preview

Let $\mathbf{P}$ be the transition matrix for a discrete memoryless channel (DMC) and consider a *block decomposition* $(\mathbf{P}^{ij})_{i,j}$ of $\mathbf{P}$ as indicated in the figure. Such a decomposition is induced by two decompositions – or equivalence relations – one of the input-, the other of the output alphabet. Assume that, within each block $\mathbf{P}^{ij}$, all row sums are equal and also that all column sums are equal. Assume further, that rows in the full matrix $\mathbf{P}$ which correspond to equivalent input letters have equal entropy. Then there exists an optimal input distribution – i.e. one for which the transmission rate reaches capacity – which is consistent with the decomposition of the input alphabet in the sense that equivalent input letters are sent with equal probability. The optimal output distribution is consistent in a similar way. This is the key result, stated in detail in Theorem 2.



II. DMC's with Benefit

Let $\mathbf{P} = (p_{xy})_{x\in X, y\in Y}$ be a stochastic matrix, fixed in the sequal. We view $\mathbf{P}$ as the transition matrix of a DMC. The sets X and Y, the *input-* and *output alphabets*, are assumed to be finite. For $x \in X$, $\vec{q}_x$ denotes the x'th row vector in $\mathbf{P}$. An *input distribution* is a distribution $\vec{p} = (p_x)_{x\in X}$ over X. The *induced output distribution* is the mixture $\vec{q} = \sum p_x \vec{q}_x$. The *information transmission rate* $I(\vec{p})$ can be expressed, using "D" for Kullback-Leibler divergence, as

$$I(\vec{p}) = \sum_{x\in X} p_x D(\vec{q}_x \| \vec{q}) \,. \tag{1}$$

[1] Supported by the Danish Natural Science Research Council and by INTAS, project 00-738.

We need a refined notion of capacity, which allows that the sending of an input symbol may be associated with a certain benefit. This idea, and the basic result connected with it, has been considered before, cf. Blahut [2, Theorem 9], where it was found more natural to associate a "cost" with the transmission of an input symbol. Consider a *benefit function* $\mathbf{a} : x \curvearrowright a_x$ which maps X into the reals, and define the *modified capacity with benefit* $\mathbf{a}$, by

$$C(\mathbf{P}; \mathbf{a}) = \sup_{\vec{p}} \left(I(\vec{p}) + \langle \mathbf{a}, \vec{p} \rangle \right) \,. \tag{2}$$

The bracket notation indicates mean value: $\langle \mathbf{a}, \vec{p} \rangle = \sum_x p_x a_x$. Clearly, the supremum in (2) is attained, and we are led to consider optimal input- and output distributions for the modified problem.

Theorem 1 (Kuhn-Tucker conditions). *Let $\vec{p}^*$ be an input distribution and $\vec{q}^*$ the induced output distribution. A necessary and sufficient condition that $\vec{p}^*$ be optimal for the modified problem with benefit $\mathbf{a}$ is that, for some constant C, the following two conditions hold:*

$$D(\vec{q}_x \| \vec{q}^*) + a_x \le C \quad \textit{for all } x \,, \tag{3}$$

$$D(\vec{q}_x \| \vec{q}^*) + a_x = C \quad \textit{for all } x \quad \textit{with } p^*_x > 0 \,. \tag{4}$$

If these conditions are satisfied, $C = C(\mathbf{P}; \mathbf{a})$.

Proof. We present a simple proof based on [8]. First assume that (3) and (4) hold. Employ the identity [1]

$$I(\vec{p}) + D(\vec{q} \| \vec{q}^*) = \sum_{x\in X} p_x D(\vec{q}_x \| \vec{q}^*) \,,$$

valid for any input distribution $\vec{p}$ with induced output distribution $\vec{q}$, to conclude that for any such $\vec{p}$,

$$I(\vec{p}) + \langle \mathbf{a}, \vec{p} \rangle \le \sum_{x\in X} p_x \Big(D(\vec{q}_x \| \vec{q}^*) + a_x \Big) \le C \,.$$

It readily follows that $C(\mathbf{P}; \mathbf{a}) = I(\vec{p}^*) + \langle \mathbf{a}, \vec{p}^* \rangle$.

To prove necessity, assume that the stated conditions fail. Denote the left hand side of (3) by K_x. Choose x_0 with K_{x_0} maximal. Then

$$K_{x_0} > \sum_{x\in X} p^*_x K_x \,. \tag{5}$$

[1] This is the "compensation identity", cf. [9].

For $0 \le \varepsilon \le 1$, put $\vec{p}_\varepsilon = (1-\varepsilon)\vec{p}^* + \varepsilon \vec{p}_{x_0}$ with $\vec{p}_{x_0}$ representing a unit mass at x_0. Let $\vec{q}_\varepsilon$ be the distribution induced by $\vec{p}_\varepsilon$. Then, by the compensation identity,

$$(1-\varepsilon)I(\vec{p}^*) + (1-\varepsilon)D(\vec{q}^*\|\vec{q}_\varepsilon) + \varepsilon D(\vec{q}_{x_0}\|\vec{q}_\varepsilon) = I(\vec{p}_\varepsilon)\,.$$

Writing $I(\vec{p}^*) + \langle \mathbf{a}, \vec{p}^*\rangle$ as $\sum p_x^* K_x$, it follows from this that $I(\vec{p}_\varepsilon) + \langle \mathbf{a}, \vec{p}_\varepsilon\rangle$ is lower bounded by

$$\sum_{x\in X} p_x^* K_x + \varepsilon\left(D(\vec{q}_{x_0}\|\vec{q}_\varepsilon) + a_{x_0} - \sum_{x\in X} p_x^* K_x\right).$$

By (5) and lower semi-continuity of divergence, it follows that for ε sufficiently small, but positive, $I(\vec{p}_\varepsilon) + \langle \mathbf{a}, \vec{p}_\varepsilon\rangle$ is strictly larger than $I(\vec{p}^*) + \langle \mathbf{a}, \vec{p}^*\rangle$, hence $\vec{p}^*$ is not optimal. □

Though some parts of the proof can be modified to cover cases with X or Y infinite, care has to be taken for several reasons, e.g. there may be no optimal distribution at all (consider $\mathbf{P}$ with rows $(1,0)$, $(\frac{1}{2},\frac{1}{2})$, $(\frac{1}{4},\frac{3}{4}),\cdots$).

Every input distribution $\vec{p}^*$ with all p_x^* positive is optimal for the modified problem for some choice of benefit function. Indeed, with benefits $a_x = -D(\vec{q}_x\|\vec{q}^*)$, $\vec{p}^*$ is optimal and the modified capacity is 0. In this setting, it is more natural to consider $D(\vec{q}_x\|\vec{q}^*)$ as a *cost* associated with the transmission of x. This remark illuminates the very definition of optimal distributions.

Whereas there may be several optimal input distributions, the optimal output distribution is unique, as in the case with zero benefit. This follows by concavity of $\vec{p} \curvearrowright I(\vec{p}) + \langle \mathbf{a}, \vec{p}\rangle$.

Explicit formulas for the calculation of optimal distributions and modified capacity do not exist in general, and even when they do, they become complicated. In the appendix we develop formulas pertaining to a 2×2 matrix.

III. The Basic Result

For any set W, $\mathrm{DEC}(W)$ denotes the set of decompositions of W, ordered by subdecomposition. Put $Z = X\times Y$ corresponding to given finite sets X and Y. A *block decomposition* of Z is a decomposition of the form $\eta_X\times\eta_Y = \{A\times B \mid A\in\eta_X\,,\, B\in\eta_Y\}$ with $\eta_X \in \mathrm{DEC}(X)$, and $\eta_Y \in \mathrm{DEC}(Y)$. A set $A\times B\in\eta_X\times\eta_Y$ is called a *block* of the decomposition. The set of block decompositions of Z is denoted $\mathrm{BDE}(Z)$.

Consider $\eta = \eta_X\times\eta_Y \in \mathrm{BDE}(Z)$, to be fixed from now on. As η is seen in conjunction with $\mathbf{P}$, we write $\eta\in\mathrm{BDE}(\mathbf{P})$. The number of classes in η_X and η_Y are denoted by M, respectively N. We put $\eta_X = \{X_i \mid i\le M\}$ and $\eta_Y = \{Y_j \mid j\le N\}$. We denote by $\mathbf{P}^{ij}$ the ij'th block in $\mathbf{P}$, i.e. $\mathbf{P}^{ij} = (p_{xy})_{x\in X_i, j\in Y_j}$. We write $\eta\in\mathrm{BDE}(\mathbf{P};\sigma_-)$ if, within each block $\mathbf{P}^{ij}$, the row sums are equal, say $=\sigma_-^{ij}$. If $\eta\in\mathrm{BDE}(\mathbf{P};\sigma_-)$, we define the *derived* DMC as the DMC with transition matrix $\partial_\eta\mathbf{P} = (\sigma_-^{ij})_{i\le M, j\le N}$ and we denote the i'th row in $\partial_\eta\mathbf{P}$ by $\vec{\sigma}_i$. We write $\eta\in\mathrm{BDE}(\mathbf{P};\sigma_+)$ if $\eta\in\mathrm{BDE}(\mathbf{P};\sigma_-)$ and if, within each block $\mathbf{P}^{ij}$, the column sums are equal, say $=\sigma_+^{ij}$.

We denote by m_i (n_j) the number of elements in X_i (Y_j) and by $\vec{u}^i$ ($\vec{v}^j$) the uniform distribution over X_i (Y_j). As indicated in section I, a *consistent* (or η-*consistent*) *input distribution* is a distribution $\vec{p}$ for which, given $i\le M$, p_x is independent of $x\in X_i$. Clearly, an input distribution $\vec{p}$ is consistent if and only if it is a convex combination $\vec{p} = \sum_{i\le M}\alpha_i\vec{u}^i$ of the $\vec{u}^i$. Similarly, we consider consistent *output distributions* which are convex combinations of the form $\vec{q} = \sum_{j\le N}\beta_j\vec{v}^j$.

Lemma 1. *Assume that $\eta\in\mathrm{BDE}(\mathbf{P};\sigma_+)$. Let $\vec{p} = \sum_{i\le M}\alpha_i\vec{u}^i$ be a consistent input distribution and $\vec{q}$ the induced output distribution. Let $\vec{\beta} = (\beta_j)_{j\le N}$ be the output distribution for $\partial_\eta\mathbf{P}$ induced by $\vec{\alpha} = (\alpha_i)_{i\le M}$. Then $\vec{q} = \sum_{j\le N}\beta_j\vec{v}^j$. In particular, $\vec{q}$ is consistent.*

Proof. This follows by simple computation, relying also on the relation $n_j\sigma_+^{ij} = m_i\sigma_-^{ij}$. □

The connection between $\mathbf{P}$ and $\partial_\eta\mathbf{P}$ will be further exploited. For this we strengthen the conditions on η. We say that η is a *generalized block symmetric decomposition*, and write $\eta\in\mathrm{GBSD}(\mathbf{P})$, if $\eta\in\mathrm{BDE}(\mathbf{P};\sigma_+)$ and if, for each $i\le M$, $H(\vec{q}_x)$ is independent of x for $x\in X_i$.

For each $x\in X$ we consider the η_Y-*conditional divergence* of $\vec{q}_x$ w.r.t. $(\vec{v}^j)_{j\le N}$, a quantity defined by first determining i such that $x\in X_i$ and then setting

$$D^\eta(\vec{q}_x\|\cdot) = \sum_{j\le N}\sigma_-^{ij} D(\vec{q}_x|Y_j\|\vec{v}^j) \tag{6}$$

with $\vec{q}_x|Y_j$ the usual conditional distribution of $\vec{q}_x$ given Y_j. If the output distribution $\vec{q}$ is consistent, say $\vec{q} = \sum_{j\le N}\beta_j\vec{v}_j$ then, by a simple datareduction identity,

$$D(\vec{q}_x\|\vec{q}) = D(\vec{\sigma}_i\|\vec{\beta}) + D^\eta(\vec{q}_x\|\cdot)\,. \tag{7}$$

Here, we still assume that $x\in X_i$. For more on identities like (7), see [9]. We can now state the main result.

Theorem 2. *Assume that $\eta\in\mathrm{BDE}(\mathbf{P};\sigma_+)$.*
(i). If $\eta\in\mathrm{GBSD}(\mathbf{P})$, there exists an optimal consistent input distribution. In this case, $a_i = D^\eta(\vec{q}_x\|\cdot)$ is independent of x for $x\in X_i$ and $C(\mathbf{P}) = C(\partial_\eta\mathbf{P};\mathbf{a})$ with benefit vector $\mathbf{a} = (a_i)_{i\le M}$.
(ii). If an optimal consistent input distribution exists with all point probabilities positive, then $\eta\in\mathrm{GBSD}(\mathbf{P})$.

Proof. (i) follows by Theorem 1 (applied both to $\mathbf{P}$ without benefit and to $\partial_\eta\mathbf{P}$ with benefit $\mathbf{a}$) in conjunction with Lemma 1 and (7).

To prove (ii) let $\vec{p}$ be an optimal consistent input distribution with positive point probabilities and let $\vec{q} = \sum_{j\le N}\beta_j\vec{v}^j$ be the induced output distribution. By Theorem 1, $D(\vec{q}_x\|\vec{q})$ is indepent of x. By (7), $D^\eta(\vec{q}_x\|\cdot)$ is independent of x when restricting x to a class X_i. As

$$D^\eta(\vec{q}_x\|\cdot) = -H(\vec{q}_x) + \sum_{j\le N}\sigma_-^{ij}\log\frac{n_j}{\sigma_-^{ij}}\,,$$

it follows that $H(\vec{q}_x)$ is independent of x for $x\in X_i$. □

It is a bit surprising that the condition of equal entropies for the $\vec{p}_i$'s plays a central role. It does not have much of the flavour of a symmetry condition. Previous results work with stronger but more "clean" conditions of symmetry. Such notions were studied already by Shannon, cf. [6, Sections 15 and 16] and appear in most textbooks. For our purposes, a matrix is called *weakly symmetric* if the rows are permutations of each other and if all column sums are equal. This terminology is in consistency with Cover and Thomas [3, p. 190]. We call η a *block symmetric decomposition* of $\mathbf{P}$ if all blocks $\mathbf{P}^{ij}$ are weakly symmetric. Notationally we write $\eta \in \mathrm{BSD}(\mathbf{P})$. Clearly, if $\eta \in \mathrm{BSD}(\mathbf{P})$, then $\eta \in \mathrm{GBSD}(\mathbf{P})$, hence:

Corollary 1. *If $\eta \in \mathrm{BSD}(\mathbf{P})$, there exist consistent optimal input- and output distributions.*

The case $M = N = 1$ is dealt with in [3]. In Gallager [4, p. 94] one finds a variant of the corollary corresponding to the case $M = 1, N$ arbitrary, though there with the stronger requirement of equal columns modulo permutations.

Example 1. Theorem 2 is stronger than Corollary 1. This is seen from the example of the 10×5 matrix $\mathbf{P}$ which is $\frac{1}{8}$ times the transpose of the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 4 & 2 & 2 & 2 & 2 & 0 \\ 1 & 1 & 1 & 4 & 1 & 2 & 2 & 2 & 0 & 2 \\ 1 & 1 & 4 & 1 & 1 & 2 & 2 & 0 & 2 & 2 \\ 1 & 4 & 1 & 1 & 1 & 2 & 0 & 2 & 2 & 2 \\ 4 & 1 & 1 & 1 & 1 & 0 & 2 & 2 & 2 & 2 \end{pmatrix}.$$

By Theorem 2, the uniform distribution $(\frac{1}{10}, \cdots, \frac{1}{10})$ is an optimal input distribution. This result is not covered by a direct application of Corollary 1, but could have been obtained by two successive applications of the corollary. Possibly, this kind of iterative procedure can be applied more generally. A complete answer to the questions this raises will contain the characterization of channels with the uniform distribution as an optimal input distribution. Even the simple case of a binary channel is not entirely trivial. [2]

IV. Coarsest decompositions, symmetry profiles

Theorem 2 may not be all that informative. For instance, in case η is the finest block symmetric decomposition (consisting of singletons), it contains no information. The coarser η is, the more informative is the result. It is important that, given any DMC $\mathbf{P}$, a coarsest block symmetric decomposition exists. Moreover, there is a simple algorithm to determine this most informative block symmetric decomposition. A similar result does not hold for the generalized notion of block symmetry.

Theorem 3. *Any DMC $\mathbf{P} = (p_{xy})_{x\in X, y\in Y}$ has a coarsest block symmetric decomposition.*

Proof. The key point is to show that $\mathrm{BSD}(\mathbf{P})$ is closed under the lattice operation $\wedge$. [3]

Assume that $\eta' = \eta'_X \times \eta'_Y$ and $\eta'' = \eta''_X \times \eta''_Y$ are in $\mathrm{BSD}(\mathbf{P})$ and put $\eta = \eta' \wedge \eta'' = \eta_X \times \eta_Y$, say. Consider an η-class $A \times B$ and let $\mathbf{Q} = (p_{xy})_{x\in A, y\in B}$. We shall show that $\mathbf{Q}$ is weakly symmetric. This involves a condition on the rows and a condition on the columns in $\mathbf{Q}$. Consider first the rows.

As $\eta_X = \eta'_X \wedge \eta''_x$, we can pass from one element of A to another by a finite number of equivalences, each one being either under η'_X (denoted $\equiv'$) or under η''_X (denoted $\equiv''$). Note that if $a_1, a_2 \in A$ and $a_1 \equiv' a_2$, then the a_1-row and the a_2-row of $\mathbf{Q}$ are decomposed into parts corresponding to the decomposition of B into η'_Y-classes. Mutually, these parts are permutations of each other. Then so are the a_1- and a_2 rows of $\mathbf{Q}$. A similar argument applies if $a_1 \equiv'' a_2$. Applying this reasoning a finite number of times, we conclude that the rows in $\mathbf{Q}$ are indeed permutations of each other.

A similar analysis applied to the columns of $\mathbf{Q}$ show that the corresponding column sums are equal.

We conclude, that each η-block $\mathbf{Q}$ of $\mathbf{P}$ is weakly symmetric, hence $\eta \in \mathrm{BSD}(\mathbf{P})$. □

Theorem 3 gives rise to the following concept. Consider a map τ which, to every stochastic matrix $\mathbf{P}$, associates a subset $\tau(\mathbf{P})$ of $\mathrm{BDE}(\mathbf{P})$. We use the notation $\eta_\tau(\mathbf{P})$ to denote the coarsest decomposition in $\tau(\mathbf{P})$, provided this is well defined. We call $\eta_\tau(\mathbf{P})$ the *τ-profile* of $\mathbf{P}$. In this terminology, Theorem 3 asserts that the BSD-*profile* exists for every $\mathbf{P}$. We write $\eta_\alpha(\mathbf{P})$ for $\eta_{\mathrm{BSD}}(\mathbf{P})$. The *strong symmetry profile* of $\mathbf{P}$, denoted $\eta_\beta(\mathbf{P})$, is defined as the profile corresponding to the subset $\mathrm{SBS}(\mathbf{P})$ of *strong block symmetric decomposition* of $\mathbf{P}$. This set consists of $\eta \in \mathrm{BDE}(\mathbf{P})$ such that, for each η-block of $\mathbf{P}$, not only the rows, but also the columns are permutations of each other. Applying the same technique as in the proof above, we see that also $\eta_\beta(\mathbf{P})$ exists for every $\mathbf{P}$. We always have $\eta_\alpha(\mathbf{P}) \le \eta_\beta(\mathbf{P})$. In general, the profiles are different. [4]

It lies nearby to ask for an effective algorithm which determines $\eta_\alpha(\mathbf{P})$. In fact, such an algorithm exists.[5] Briefly, this works as follows. Firstly, $\eta_1 \in \mathrm{BDE}(\mathbf{P})$ is constructed corresponding to the two equivalence relations "corresponding rows in $\mathbf{P}$ are permutations of each

[2] This case may be discussed by considering the function $D(\vec{q}_1\|\vec{q}) - D(\vec{q}_2\|\vec{q})$ with $\vec{q} = \frac{1}{2}\vec{q}_1 + \frac{1}{2}\vec{q}_2$ as a function of p_{11} and p_{22}, observing that the determinant of the Hessian has a simple factorization; indeed, the determinant in question is

$$\frac{-(1-\alpha-\beta)^4}{\alpha\beta(1-\alpha)(1-\beta)(1-\beta+\alpha)^2(1-\alpha+\beta)^2}$$

where $\alpha = p_{11}$, $\beta = p_{22}$. This is the key fact needed to show that the function only vanishes on the diagonals of the unit square (for this argument, we acknowledge discussions with J. P. R. Christensen).

[3] The set $\mathrm{BDE}(Z)$ is a sublattice of $\mathrm{DEC}(Z)$, but $\mathrm{BSD}(\mathbf{P})$ is not a further sublattice. To see this, consider $\mathbf{P}$ consisting of the two rows $(0, \frac{1}{2}, 0, \frac{1}{2})$ and $(\frac{1}{2}, 0, \frac{1}{2}, 0)$ and note that $\mathrm{BSD}(\mathbf{P})$ is not closed under the lattice operation $\vee$.

[4] Consider the 2×3 matrix with rows $(0, \frac{1}{3}, \frac{2}{3})$ and $(\frac{2}{3}, \frac{1}{3}, 0)$.

[5] This observation resulted from discussions with Mr. Thomas Jakobsen, the Danish Technical University, who also implemented the algorithm.

other", and, "corresponding column sums in $\mathbf{P}$ are equal". Each of the η_1-blocks of $\mathbf{P}$ are then handled one by one in a similar way. Each step introduces a block decomposition of $\mathbf{P}$, finer than the previous one. The process continues until no block gives rise to a finer decomposition. A finite sequence $0 \leq \eta_1 \leq \eta_2 \leq \cdots \leq \eta_k$ is then constructed with $\eta_k = \eta_\alpha(\mathbf{P})$. The algorithm needs approximately $\nu^3 \log \nu$ steps where $\nu = \max(m, n)$.

V. Discussion

Theorem 2 reduces the problem to determine capacity and optimal distributions from $\mathbf{P}$ to $\partial_\eta \mathbf{P}$ (e.g. with $\eta = \eta_\alpha(\mathbf{P})$). However, the reduced problem cannot, in general, be solved in closed form. One often has to turn to numerical methods, and here, the Arimoto-Blahut algorithm is the obvious choice, cf. [1] and [2]. Note that this algorithm can be modified without difficulty to the case when we allow benefits. Theoretical results and numerical experiments have shown the feasibility of this approach but, at the same time, indicated that there is little or no saving in using the reduction provided by our results as compared to an approach where the Arimoto-Blahut algorithm is employed directly to the original problem.

In the literature (Silverman [7], [3], [4], [5] etc.), non-trivial concrete examples of DMC's are often pointed out with $\mathbf{P}$ a 3×3 matrix. They all have a non-trivial block symmetric decomposition, hence are of the form

$$\mathbf{P} = \begin{pmatrix} \alpha & \beta & \gamma \\ \beta & \alpha & \gamma \\ \delta & \delta & \varepsilon \end{pmatrix} . \tag{8}$$

For instance, [7] has $\beta = \varepsilon = 0$. Applying our results leads to substantial simplifications.

VI. Appendix

Consider the DMC $\mathbf{P}$ given by

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \tag{9}$$

and an associated benefit vector $\mathbf{a} = (a_1, a_2)$.

Below, i and j are either 1 or 2 and $i \neq j$. Denote by d the determinant of $\mathbf{P}$, by H_i the entropy of $\vec{q}_i$ and by D_{ij} the divergence $D(\vec{q}_i \| \vec{q}_j)$. Put $a_{ij} = a_i - a_j$, $H_{ij} = H_i - H_j$ and define, for $d \neq 0$, quantities h_i, h_{ij}, and d_{ij} by $h_i = d^{-1}(p_{jj}(a_i - H_i) - p_{ij}(a_j - H_j))$, $h_{ij} = h_i - h_j = d^{-1}(a_{ij} - H_{ij})$ and $d_{ij} = d^{-1}(D_{ij} + a_{ij})$.

For $d \neq 0$ consider $\vec{q}^* = (q_1^*, q_2^*)$ given by

$$q_i^* = (1 + e^{h_{ji}})^{-1}$$

and consider the unique signed (!) input distribution $\vec{p}^*$ which induces $\vec{q}^*$ i.e.

$$p_i^* = d^{-1}(q_i^* - p_{ji}).$$

Elementary calculations lead to the formulas:

$$q_i^* = \frac{p_{ii}}{p_{ii} + p_{ij} e^{d_{ji}}} = \frac{p_{ji}}{p_{ji} + p_{jj} e^{-d_{ij}}} ,$$

$$p_i^* = \frac{q_i^*}{d} \left(p_{jj} - p_{ji} e^{h_{ji}} \right) = \frac{q_j^*}{d} \left(p_{jj} e^{h_{ij}} - p_{ji} \right) ,$$

$$p_i^* = \frac{p_{ji} q_j^*}{d} \left(e^{d_{ij}} - 1 \right) = \frac{p_{jj} q_i^*}{d} \left(1 - e^{-d_{ij}} \right) ,$$

which apply when $d \neq 0$. Continuity considerations may have to be taken into account, however, we can always select formulas to avoid this. Applying Theorem 1, one finds:

Theorem 4. *Consider $\mathbf{P}$ given by (9) and with associated benefit vector $\mathbf{a} = (a_1, a_2)$. If $d = 0$ and $a_1 > a_2$, then $(1, 0)$ is the unique optimal input distribution and $C(\mathbf{P}; \mathbf{a}) = a_1$; if $d = 0$ and $a_2 > a_1$, a similar statement holds, and if $d = 0$ and $a_1 = a_2$, any input distribution is optimal and $C(\mathbf{P}; \mathbf{a}) = a_1$.*

If $d \neq 0$, the optimal input distribution is unique, and a necessary and sufficient condition that this distribution is non-trivial (i.e. both input probabilities are positive) is that

$$-D_{21} < a_2 - a_1 < D_{12} \tag{10}$$

When (10) holds, the optimal input- and output distributions are $\vec{p}^*$ and $\vec{q}^*$ given by the formulas above and the modified capacity can be determined from Theorem 1 or from the formulas

$$C(\mathbf{P}; \mathbf{a}) = h_1 - \log q_1^* = h_2 - \log q_2^* = \log \left(e^{h_1} + e^{h_2} \right) .$$

If $a_2 - a_1 \leq -D_{21}$, the optimal input distribution is $(1, 0)$ and $C(\mathbf{P}; \mathbf{a}) = a_1$, and if $a_2 - a_1 \geq D_{12}$, the optimal input distribution is $(0, 1)$ and $C(\mathbf{P}; \mathbf{a}) = a_2$.

By Theorem 1, this is easily checked.

References

[1] S. Arimoto "An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels", *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 14–20, Jan. 1972.

[2] R. E. Blahut, "Computation of Channel Capacity and Rate-Distortion Functions", *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.

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

[4] R. C. Gallager, *Information Theory and Reliable Communication*. New York: Wiley 1968.

[5] C. M. Goldie and R. G. E. Pinch, *Communication Theory*, Cambridge: Cambridge University Press 1991.

[6] C. Shannon, "A mathematical theory of communication", *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[7] R. A. Silverman, "On Binary Channels and their Cascades", *IRE Trans. on Inform. Theory*, vol. IT-1, pp. 19–27, Dec. 1955.

[8] F. Topsøe, "A New Proof of a Result Concerning Computation of the Capacity for a Discrete Channel", *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 22, pp. 166–168, 1972.

[9] F. Topsøe, "Basic Concepts, Identities and Inequalities – the Toolkit of Information Theory", *Entropy*, vol. 3, pp. 162–190, 2001, http://www.unibas.ch/mdpi/entropy/ [ONLINE].