# Block Symmetry in Discrete Memoryless Channels

Jakob Bøje Pedersen and Flemming Topsøe
Copenhagen and Department of Mathematics
University of Copenhagen, Copenhagen, Denmark *

**Abstract**

Various notions of *block symmetry* for discrete, memoryless channels are introduced. Though block symmetry in itself does not guarantee easy calculation of channel capacity and easy determination of optimal distributions, usually, it does simplify matters, and also, known results in this direction – often based on rather strong concepts of symmetry – appear as simple corollaries to the results here presented. The findings appear to be relatively simple and natural ones which somehow were not considered or overlooked in the early development of information theory.

**Keywords**. Discrete Memoryless Channel, capacity, block symmetry.

## 1 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 Figure 1. Such a decomposition is induced by two decompositions – or equivalence relations if you wish – one of the input-, the other of the output alphabet. Assume that, within each block $\mathbf{P}^{ij}$, all row sums are equal and all column sums are equal. Assume further, that rows in the full matrix $\mathbf{P}$ which correspond to equvalent input letters have equal entropy. Then there exists an optiml 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. Furthermore,

*Flemming Topsøe is supported by the Danish Natural Science Research Council and by INTAS (project 00-738).

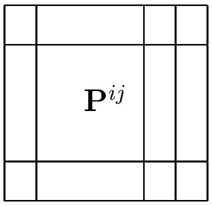


Figure 1

the optimal output distribution is consistent in a similar way, hence equivalent output letters are received with equal probability. This result may be viewed as our key result.

## 2 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 here 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 probability 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* $I(\vec{p};\mathbf{P})$, or just $I(\vec{p})$, can be expressed, using "D" for divergence, as

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

We need a refined notion of capacity, taking into regard 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 input symbols).

Consider a *benefit function* $\mathbf{a} : x \curvearrowright a_x$ which maps $X$ into the reals, and define the *modified capacity with benefit* $\mathbf{a}$, $C(\mathbf{P}, \mathbf{a})$, by

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

Here, 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, thereby generalizing the usual concepts (which correspond to the case with zero benefit).

**Lemma 1.** *Let $\vec{p}^*$ be an input distribution and denote by $\vec{q}^*$ the induced output distribution. A necessary and sufficient condition that $\vec{p}^*$ be an optimal* *input distribution for the modified problem with benefit* **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,* $C$ *is the modified capacity:* $C = C(\mathbf{P};\mathbf{a})$*.*

*Proof.* We show a simple proof of sufficiency, extending the reasoning in [8]. So assume that (3) and (4) hold. Employ the identity

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

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

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

Explicit formulas for calculation of optimal distributions and modified capacity pertaining to a general $2\times 2$ transition matrix have been worked out using Lemma 1.

# 3 The Basic Result

For any set $W$, $\mathrm{DEC}(W)$ denotes the lattice of decompositions of $W$, ordered by subdecompositions. Put $Z = X \times Y$. A *block decomposition* of $Z$ is a decomposition of $Z$ 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 $\mathrm{BDE}(Z)$. This set is a sublattice of $\mathrm{DEC}(Z)$.

Consider a block decomposition $\eta = \eta_X \times \eta_Y \in \mathrm{BDE}(X \times Y)$. As the block decomposition $\eta$ is seen in relation to $\mathbf{P}$, we write $\eta \in \mathrm{BDE}(\mathbf{P})$. The number of classes in $\eta_X$ and $\eta_Y$ are denoted $M$, respectively $N$. We put

$\eta_X = \{X_i \mid i \leq M\}$ and $\eta_Y = \{Y_j \mid j \leq N\}$. For $i \leq M, j \leq N$ 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 $\partial_\eta \mathbf{P} = (\sigma_-^{ij})_{i \leq M, j \leq N}$.

A general result (not shown here) can be developed which gives sufficient conditions for existence of optimal input- and output distributions for the DMC defined by $\mathbf{P}$ with certain prescribed conditional distributions, where conditioning is relative to the classes defined by $\eta_X$ and $\eta_Y$. The idea of proof is quite simple: We relate the original problem about $\mathbf{P}$ to one related to the derived transition matrix $\partial_\eta \mathbf{P}$. With appropriate assumptions it turns out that this reduction is possible if we introduce benefits. This situation is then handled via Lemma 1.

The results corresponding to uniform conditional distributions are those quoted in the preview and below. They lead to the following notions. A matrix is *weakly symmetric* if the rows are permutations of each other and if all column sums are equal. This terminology follows Cover and Thomas [3] (cf. p.190). We write $\eta \in \mathrm{BSD}(\mathbf{P})$ and call $\eta$ a *block symmetric decomposition* of $\mathbf{P}$ if all blocks in $\mathbf{P}$ are weakly symmetric. A *generalized block symmetric decomposition* is the one discussed in the preview. An even more general notion (not defined here) is preserved under the operation of adding rows to $\mathbf{P}$, a simple operation which destroys other notions of symmetry.

Notions of symmetry were studied by Shannon in [6] (cf. Sections 15 and 16), and appear in most textbooks. For $M = N = 1$, one may consult [3] as mentioned, and for $M = 1, N$ arbitrary, the notion is discussed in Gallager [4] (cf. p. 94, though there with the stronger requirement of equal columns modulo permutations).

**Theorem 1.** *For every generalized block symmetric decomposition of a given DMC, the optimal output distribution is consistent with the decomposition of the output alphabet and there exists an optimal input distribution which is consistent with the decomposition of the input alphabet.*

# 4 Discussion and further results, indications

Theorem 1 may not be all that informative. For instance if, given any DMC, you consider the finest block symmetric decomposition, the result actually contains no information. It is important to note that there is always, given any DMC, a coarsest block symmetric decomposition. Moreover, there is a simple algorithm to determine this most informative block symmetric decomposition (result obtained in discussion with Thomas Jakobsen). A similar result is not possible for the generalized notion of block symmetry.

As just one example of a generalized block symmetric deceomposition which is not a block symmetric decomposition we mention the $10 \times 5$ matrix $\mathbf{P}$ given, in natural notation, by $\mathbf{P} = (\mathbf{A}/\mathbf{B})$ where

$$\mathbf{A} = \frac{1}{8}\begin{pmatrix} 1 & 1 & 1 & 1 & 4 \\ 1 & 1 & 1 & 4 & 1 \\ 1 & 1 & 4 & 1 & 1 \\ 1 & 4 & 1 & 1 & 1 \\ 4 & 1 & 1 & 1 & 1 \end{pmatrix}, \qquad \mathbf{B} = \frac{1}{8}\begin{pmatrix} 2 & 2 & 2 & 2 & 0 \\ 2 & 2 & 2 & 0 & 2 \\ 2 & 2 & 0 & 2 & 2 \\ 2 & 0 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 & 2 \end{pmatrix}. \tag{5}$$

Thus, the uniform distribution $(\frac{1}{10}, \cdots, \frac{1}{10})$ is an optimal input distribution in this case.

In this example, the result could have been obtained by two successive applications of the simpler result which involves block symmetric decompositions. Possibly this kind of iterative procedure based on the simpler of the two notions discussed is the rule rather than the exception. A complete solution of this problem will involve 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. [1]

Theorem 1 can be viewed as a reduction of the problem of determining $C(\mathbf{P})$ – and associated optimal distributions – to a simpler problem involving the derived channel. However, the reduced problem cannot, generally speaking, be solved in closed form. One often has to turn to numerical methods, and here, the Arimoto-Blahut algorithm is the obvious choice, cf. Arimoto [1] and Blahut [2]. In this connection we point out that the algorithm can be modified without difficulty to the case when we allow for benefits. Theoretical results (cf. [1], [2]) and some 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 a direct approach employing the algorithm directly on the original problem.

[1] 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 $p_{22}$ and 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 $\alpha = p_{11}$ and $\beta = p_{22}$. This observation is the key fact used 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).

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

$$\mathbf{P} = \begin{pmatrix} \alpha & \beta & \gamma \\ \beta & \alpha & \gamma \\ \delta & \delta & \varepsilon \end{pmatrix} . \tag{6}$$

For instance, [7] has $\beta = \varepsilon = 0$.

With $\eta$ the obvious block symmetric decomposition, one can use our results to calculate the capacity and the optimal distributions.

In view of the result hinted at that a coarsest block symmetric decomposition always exists, we note that $\mathrm{BSD}(\mathbf{P})$ is not a lattice. Simple examples (e.g. with $\mathbf{P}$ containing the two row vectors $(0, \frac{1}{2}, 0, \frac{1}{2})$ and $(\frac{1}{2}, 0, \frac{1}{2}, 0)$) show that $\mathrm{BSD}(\mathbf{P})$ need not be closed under $\vee$.

# 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