Subadditivity of entropy for stochastic matrices

Wojciech Słomczyński[1]

PREPRINT 2001/16

Abstract

The entropy of irreducible stochastic matrix measures its mixing properties. We show that this quantity is subadditive and strongly subadditive.

[1]Instytut Matematyki
Uniwersytet Jagielloński
Reymonta 4, 30-059 Kraków, Poland
mail: slomczyn@im.uj.edu.pl

1 Preliminaries

Let $N \in \mathbb{N}$. We denote the simplex of probability vectors (finite probability distributions) by

$$\Delta_N = \left\{ p \in \mathbb{R}^N : p_i \geq 0 \text{ for } i = 1, \ldots, N \text{ and } \sum_{i=1}^N p_i = 1 \right\}$$

and its interior by

$$\Delta_N^\circ = \left\{ p \in \mathbb{R}^N : p_i > 0 \text{ for } i = 1, \ldots, N \text{ and } \sum_{i=1}^N p_i = 1 \right\} .$$

The extreme points of the simplex will be denoted by $e^k = (0, ..., \underset{k}{1}, ...0)$ $(k = 1, ..., N)$.

We define the function $\eta : \mathbb{R}^+ \to \mathbb{R}$ by the formula

$$\eta(x) = \begin{cases} -x \ln x & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases} .$$

Then η is continuous and strictly concave. Moreover $\eta|_{[0,1]}$ is nonnegative.

The *Boltzmann-Shannon entropy* $h : \Delta_N \to \mathbb{R}^+$ is defined by the formula

$$h(p) = \sum_{i=1}^N \eta(p_i)$$

for $p \in \Delta_N$. The quantity $h(p)$ can be interpreted as a measure of uncertainty of p and it changes from 0 for e^k $(k = 1, ..., N)$ to $\ln N$ for the uniform vector $(1/N, \ldots, 1/N)$. It is easy to check that h is continuous and concave.

We call a nonnegative matrix $P = (p_{ij})_{i,j=1,\ldots,N}$ *stochastic* iff $\sum_{j=1}^N p_{ij} = 1$ for each $i = 1, \ldots, N$. We say that a stochastic matrix $P = (p_{ij})_{i,j=1,\ldots,N}$ is *irreducible* iff for every $i, j = 1, \ldots, N$ there exists $k \in \mathbb{N}$ such that $\left(P^k\right)_{ij} > 0$. We denote the set of all irreducible stochastic matrices by I_N. It is well known (see e.g. [1]) that a stochastic matrix P is irreducible, iff it is *ergodic*, i.e., iff there exists a unique probability (row) vector $p \in \Delta_N^\circ$ such that $pP = p$, where $(pP)_j := \sum_{i=1}^N p_i p_{ij}$ for $j = 1, \ldots, N$ (i.e. p is the normalized left eigenvector of P corresponding to the eigenvalue 1). We call such a vector *stationary*.

2 Definitions

To measure the amount of chaos generated by an irreducible stochastic matrix P one may consider its entropy.

Definition 1 *For $P = (p_{ij})_{i,j=1,\ldots,N} \in I_N$ we define the **entropy of** P by the formula*

$$H(P) := -\sum_{i,j=1}^{N} p_i p_{ij} \ln p_{ij} \ ,$$

where $p \in \Delta_N^{\circ}$ is the stationary vector for P.

Note that this quantity is equal to the Kolmogorov-Sinai entropy of the Markov shift generated by P and p (see [2]), and varies from 0 for permutation matrices to $\ln N$ for the matrix all of whose elements are $1/N$. To estimate the entropy increase we will also use another quantity - the weighted entropy of a stochastic matrix with respect to a probability vector.

Definition 2 *Let $P = (p_{ij})_{i,j=1,\ldots,N}$ be a stochastic matrix, $y \in \Delta_N$. We define the **weighted entropy of** P **with respect to** y by the formula*

$$H_y(P) := \textstyle\sum_{k=1}^{N} y_k h\left(e^k P\right) \ .$$

If $P \in I_N$ and p is its stationary vector, then $H(P) = H_p(P)$.

3 Results

Our main result is the following

Theorem 1 (subadditivity of entropy) *Let B and C be two irreducible stochastic matrices with common stationary vector p. Then*

$$H(C) \le H(BC) \le H(B) + H(C) \ . \tag{1}$$

The elementary proof of the theorem is based on several auxiliary lemmas and propositions.

Lemma 1 *Let $\lambda_i \geq 0$, $a_i > 0$, and $d_i \in \mathbb{R}$ for $i = 1, ..., N$. Then*

$$\left(\sum_{i=1}^{N} \lambda_i d_i\right)^2 \leq \left(\sum_{i=1}^{N} \lambda_i a_i\right) \cdot \left(\sum_{i=1}^{N} \lambda_i \frac{(d_i)^2}{a_i}\right) . \tag{2}$$

Proof. Put $\alpha_i = \sqrt{\lambda_i a_i}$ and $\beta_i = d_i \sqrt{\lambda_i / a_i}$ for $i = 1, ..., N$. Then the inequality (2) follows from the Cauchy-Schwartz inequality $\left(\sum_{i=1}^{N} \alpha_i \beta_i\right)^2 \leq \left(\sum_{i=1}^{N} \alpha_i^2\right) \cdot \left(\sum_{i=1}^{N} \beta_i^2\right)$. ■

Lemma 2 *Let $C = (c_{ij})_{i,j=1,...,N} \in I_N$, $y \in \Delta_N^\circ$, $x \in \mathbb{R}^N$. Then*

$$\sum_{j=1}^{N} \frac{(xC)_j^2}{(yC)_j} \leq \sum_{j=1}^{N} \frac{x_j^2}{y_j} \tag{3}$$

(note that $(yC)_j \neq 0$ for $j = 1, ..., N$ due to the irreducibility of matrix C).

Proof. Let $j = 1, ..., N$. Put $\lambda_i = c_{ij}$, $a_i = y_i$, and $d_i = x_i$. Applying Lemma 1 we get $(xC)_j^2 / (yC)_j \leq \sum_{i=1}^{N} c_{ij} x_i^2 / y_i$. Summing these inequalities over $j = 1, ..., N$ and using the identities $\sum_{j=1}^{N} c_{ij} = 1$ ($i = 1, ..., N$) we obtain the desired conclusion. ■

The above inequality has a simple geometrical meaning. Let us recall that the natural *Fisher-Mahalonobis-Bhattacharyya metric* on Δ_N° is given by the Riemannian form $g_{ij}(p) = \delta_{ij} p_j^{-1}$ for $p \in \Delta_N^\circ$ and $i, j = 1, ..., N$ (see e.g. [3], some authors call it also the *Shahshahani metric* [4]). The Riemannian geometry of Δ_N° is that of $1/2^N$ portion of the sphere of radius two in $\mathbb{R}^N$. Using the Riemannian form we can rewrite the inequality (3) as

$$\|xC\|_{yC} \leq \|x\|_y \tag{4}$$

for $y \in \Delta_N^\circ$ and $x \in T_y \Delta_N^\circ \sim \left\{x \in \mathbb{R}^N : \sum_{i=1}^{N} x_i = 0\right\}$. Consequently, we obtain

Proposition 1 *If $C \in I_N$, then the map $\Delta_N^\circ \ni x \to xC \in \Delta_N^\circ$ is a contraction on Δ_N° endowed with the Fisher-Mahalonobis-Bhattacharyya metric.*

In fact, Chentsov [5] proved that this property characterizes the Fisher-Mahalonobis-Bhattacharyya metric among all Riemannian metrics on Δ_N° (see also [6] and [7]).

From Lemma 2 we deduce

Proposition 2 *Let $C=(c_{ij})_{i,j=1,\ldots,N}$ be a stochastic matrix. Then the function $F_C:\Delta_N\to\mathbb{R}$ given by $F_C(y)=h(yC)-h(y)$ for $y\in\Delta_N$ is convex.*

Proof. Observe first that the function $(C,y)\to F_C(y)$ is continuous in both variables. We first assume that C is irreducible. Let $y\in\Delta_N^{\circ}$. Then, it is easy to check that $\frac{\partial F_C}{\partial y_k}(y)=\ln y_k-\sum_{j=1}^{N}c_{kj}\ln(yC)_j$ for $k=1,...,N$, and so

$$\frac{\partial^2 F_C}{\partial y_k\partial y_m}(y)=\sum_{j=1}^{N}\frac{(e^k)_j\,(e^m)_j}{y_j}-\sum_{j=1}^{N}\frac{(e^kC)_j\,(e^mC)_j}{(yC)_j}$$

for $k,m=1,...,N$. Hence the quadratic form related to the second order differential of F_C at y is given by

$$\left(D_y^2F_C\right)(x,x)=\sum_{j=1}^{N}\frac{x_j^2}{y_j}-\sum_{j=1}^{N}\frac{(xC)_j^2}{(yC)_j} \tag{5}$$

for $x\in\mathbb{R}^N$. Applying (3) and (5) we see that the Hessian matrix $\left(\frac{\partial^2F_C}{\partial y_k\partial y_m}(y)\right)_{k,m=1,\ldots,N}$ is nonnegative definite. Consequently, F_C is convex on Δ_N°, and so on Δ_N. If C is not irreducible, then one can approximate it by irreducible matrices. ■

Our next claim is that the weighted entropy estimates the entropy increase from above (the result was announced without proof in [8]). Namely, we have

Proposition 3 *Let $(C_{ij})_{i,j=1,\ldots,N}$ be a stochastic matrix, $y\in\Delta_N$. Then*

$$H_y(C)\le h(yC)\le H_y(C)+h(y)\ . \tag{6}$$

Proof. Let $y\in\Delta_N$. Clearly, $y=\sum_{k=1}^{N}y_ke^k$. The first inequality follows from the fact that the function $\Delta_N\ni y\to h(yC)\in\mathbb{R}$ is concave. The second inequality follows from Proposition 2 and from the equalities $h\left(e^k\right)=0$ for $k=1,\ldots,N$. ■

We are now in a position to show the main theorem.

Proof of the subadditivity of entropy. Let $n = 1, \ldots, N$. Put $(y_n)_k = b_{nk}$ for $k = 1, \ldots, N$. Then $y_n \in \Delta_N$. Applying Proposition 3 we get

$$H_{y_n}(C) \leq h(y_n C) \leq H_{y_n}(C) + h(y_n) \ .$$

Hence

$$\textstyle\sum_{k=1}^{N} b_{nk} \sum_{j=1}^{N} \eta(c_{jk}) \leq \sum_{k=1}^{N} \eta\left(\sum_{j=1}^{N} b_{nj} c_{jk}\right)$$

and

$$\textstyle\sum_{k=1}^{N} \eta\left(\sum_{j=1}^{N} b_{nj} c_{jk}\right) \leq \sum_{k=1}^{N} b_{nk} \sum_{j=1}^{N} \eta(c_{kj}) + \sum_{k=1}^{N} \eta(b_{nk}) \ .$$

Multiplying these inequalities by p_n and summing over $n = 1, \ldots, N$ we get the required result. ∎

Analogously, one can show even stronger result.

Theorem 2 (strong subadditivity of entropy) *Let A, B, and C be irreducible stochastic matrices with common stationary vector p. Then*

$$H(ABC) + H(B) \leq H(AB) + H(BC) \ . \tag{7}$$

Proof. From Proposition 2 it follows that the function $F_{B,C} : \Delta_N \to \mathbb{R}$ given by $F_{B,C}(y) = h(yBC) - h(yB)$ for $y \in \Delta_N$ is convex. Hence, as in the proof of Proposition 3, we get the inequality $h(yBC) - h(yB) \leq H_y(BC) - H_y(B)$ for $y \in \Delta_N$. Then we proceed in the same manner as in the proof of Theorem 1. ∎

Remark 1 *Theorems 1 and 2 still hold if we drop the assumption of the irreducibility of matrices. However, in this case the entropy depends both on a matrix and on a stationary vector, which need not be unique.*

Remark 2 *We call a stochastic matrix $P = (p_{ij})_{i,j=1,\ldots,N}$* bistochastic *(or* doubly stochastic*) iff $\sum_{i=1}^{N} p_{ij} = 1$ for each $j = 1, \ldots, N$. If a bistochastic matrix is irreducible, then its only stationary vector is the uniform vector $(1/N, \ldots, 1/N)$. It is well known that the action of bistochastic P on probability vectors increases the uncertainty, i.e., $h(p) \leq h(pP)$ for $p \in \Delta_N$ - this fact is the first step in the proof of the famous H-theorem [9]. From this inequality, as in the proof of Theorem 1, we get the inequality $H(B) \leq H(BC)$ for irreducible bistochastic matrices B and C.*

Acknowledgments. I am indebted to K. Życzkowski for suggesting the problem and for stimulating conversations, and to A. Urbański for helpful remarks. Financial support from Polski Komitet Badań Naukowych Grant no 2P 03B 07219 and British-Polish Joint Research Collaboration Programme Grant no WAR 341 204 is gratefully acknowledged.

References

[1] Berman, A. and Plemmons, R. J., *Nonnegative Matrices in the Mathematical Sciences* (Academic Press, New York, 1979).

[2] Walters, P., *An Introduction to Ergodic Theory* (Springer-Verlag, New York, 1982).

[3] Antonelli, P. L., The information dynamics of the diffusion processes of population genetics. *Open Syst. Inf. Dyn.* **4**, 101-117 (1997).

[4] Akin, E., *The Geometry of Population Genetics* (Springer-Verlag, Berlin, 1979).

[5] Čencov, N. N., Statisticheskie reshayushchie pravila i optimal'nye vyvody (Nauka, Moscow, 1972) (in Russian), English transl.: *Statistical decision rules and optimal inferences.* Translation of Math. Monog. 53 (Amer. Math. Soc., Providence, R. I., 1982).

[6] Čencov, N. N., On basic concepts of mathematical statistics, in: Mathematical statistics. Banach Center Publ., vol. 6 (PWN, Warsaw, 1980), pp. 85-94.

[7] Campbell, L. L., An extended Čencov characterization of the information metric. *Proc. Amer. Math. Soc.* **98**, 135-141 (1986).

[8] Życzkowski, K., Słomczyński, W., and Kuś, M., Random unistochastic matrices, *to appear.*

[9] Lasota, A. and Mackey M. C., *Chaos, Fractals, and Noise. Stochastic Aspects of Dynamics* (Springer-Verlag, New York, 1994).