

An Elaboration on

Relative Entropy Under Mappings by Stochastic Matrices

Joel E. Cohen, Yoh Iwasa, Gh. Rautu, Mary Beth Ruskai, Eugene Seneta, Gh. Zbaganu

By Aaron Kelley

1 Introduction

Relative entropy has been extensively studied in the past in many different contexts and under a variety of different names, much of which had to do with the change over time between two distributions. However, the main focus of this paper concerns its properties under mappings by stochastic matrices. The results can be used to provide more information on bounding the rates of convergence to equilibrium of ergodic Markov chains and Markov processes. After reviewing some definitions, we will see theorem 3.1 (omitting the majority of the proof) which will be used to prove the two main results selected for this elaboration: theorems 4.1 and 5.4. Finally, we will see some of the ideas of this paper in the context of the what we saw throughout the semester.

2 Background

The following definitions are necessary in order to lay the groundwork for how we will talk about probability vectors and stochastic matrices.

Definition 1.1 (Vectors): Let m, n , and d be finite positive integers. Vectors that are $n \times 1$ or $d \times 1$ will be called **n** and **d** vectors respectively. \mathbf{N}_d will be the set of probability vectors and \mathbf{P}_d the positive ones. That is,

$$N_d = \{x \in R^d : x_i \geq 0, \sum_i x_i = 1\} \quad P_d = \{x \in N_d : x_i > 0, \forall i\}$$

Definition 1.2 (Matrices): A **(column) stochastic** $d \times n$ matrix is a matrix whose columns belong to N_d . A nonnegative matrix is called **row-allowable** if each row contains at least one positive element. A matrix with at least one positive row (all elements of a row positive) is called **row-positive**. A column-stochastic row-positive matrix is called a **Markov matrix**. A nonnegative $d \times d$ matrix A is called **primitive** if A^k is positive for some positive integer k . A column-stochastic $d \times n$ matrix is called a **scrambling** matrix if any submatrix consisting of two columns has a row both elements of which are positive.

The following are the principal definitions and immediate consequences that will be referenced throughout the rest of the paper.

Definition 2.1 (Symmetric Relative Entropy): For any two positive d -vectors $x = (x_i)$ and $y = (y_i)$ (not necessarily probability vectors), the *relative entropy* is defined as

$$H(x, y) = \sum_i x_i \log(x_i/y_i)$$

and they *symmetric relative entropy* as

$$J(x, y) = H(x, y) + H(y, x) = \sum_i (x_i - y_i) \log \frac{x_i}{y_i}$$

Definition 2.2 (relative ϕ -entropy): Let ϕ be a continuous real-valued function on

$(0, \infty) \times (0, \infty)$ that is homogeneous and jointly convex in its arguments, and satisfies $\phi(1, 1) = 0$. For any two positive d -vectors, $x = (x_i), y = (y_i)$ (not necessarily probability vectors) the *relative ϕ -entropy* is defined as

$$H_\phi = \sum_i \phi(x_i, y_i)$$

and the *symmetric relative ϕ -entropy* as

$$J_\phi(x, y) = H_\phi(x, y) + H_\phi(y, x)^2$$

The function ϕ defined in definition 2.2 is jointly convex in both arguments if and only if $g(t) \equiv \phi(1, 1+t)$ is convex for $t \in (-1, \infty)$. Therefore any continuous real-valued convex function $g(t)$ on $(-1, \infty)$ such that $g(0) = 0$ defines a relative ϕ -entropy via the assumptions that $\phi(1, 1+t) = g(t)$ and ϕ is homogeneous. So the relative ϕ -entropy and related quantities can be indexed by both ϕ and/or g . That's to say

$$H_\phi(x, y) = \sum_i \phi(x_i, y_i) \iff H_g(x, y) = \sum_i x_i g(y_i/x_i - 1)$$

Keep in mind that in all cases H_{\log} denotes the relative entropy in definition 2.1. That is

$$H_{\log} = H_g, \quad \text{when } g(t) = -\log(1+t)$$

Here are a few fundamental properties of relative entropy, many of which will be referenced in the later sections:

1. H_ϕ is a continuous, real-valued function that is homogeneous, jointly convex in (x, y) for any positive d -vectors x and y , subadditive, and such that $H_\phi(x, x) = 0$.
2. For any $x, y \in P_d$, $H_\phi(x, y) \geq 0$; and if $\phi(1, t)$ is strictly convex for $t \in (0, \infty)$, then $H_\phi(x, y) = 0$ if and only if $x = y$.
3. For any positive d -vectors x, y and positive n -vectors x', y' any permutation matrices Q_1, Q_2 of size $m \times m$ and $n \times n$, respectively, and any row-allowable $m \times n$ matrix A , there exists positive n -vectors x', y' such that

$$\frac{H_\phi(Q_1 A Q_2 x, Q_1 A Q_2 y)}{H_\phi(x, y)} = \frac{H_\phi(Ax', Ay')}{H_\phi(x', y')}$$

4. If A is a column-stochastic, row-allowable $m \times d$ matrix and x, y are positive d -vectors, and $\phi(1, \cdot)$ convex, then $H_\phi(Ax, Ay) \leq H_\phi(x, y)$

Definition 2.3 (Dobrushin's Ergodicity Coefficient): For any $m \times n$ matrix A , Dobrushin's coefficient of ergodicity is

$$\alpha(A) = \min_{j,k} \sum_{i=1}^m \min(a_{ij}, a_{ik})$$

We will see that the complement, $1 - \alpha(A)$, is a bit more interesting with respect to the conclusions that we arrive at. Here we note that

$$\bar{\alpha}(A) \equiv 1 - \alpha(A) = \frac{1}{2} \max_{j,k} \sum_{i=1}^m |a_{ij} - a_{ik}|$$

²because $\tilde{\phi}(a, b) = \phi(a, b) + \phi(b, a)$ satisfies the conditions of definition 2.2 if ϕ does, and $J_\phi(x, y) = H_{\tilde{\phi}}(x, y)$, we can speak of J_ϕ and $H_{\tilde{\phi}}$ interchangeably.

and also satisfies

$$\bar{\alpha}(A) = \sup \left\{ \frac{\|A(x-y)\|_1}{\|x-y\|_1} : x \text{ and } y \text{ are positive } n\text{-vectors such that } x \neq y, \|x\|_1 = \|y\|_1 \right\}$$

Definition 2.4 (ϕ -entropy contraction coefficient): Let A be a column-stochastic, row-allowable $m \times n$ matrix. We define the **relative ϕ -entropy contraction coefficient**

$$\eta_\phi(A) = \sup \left\{ \frac{H_\phi(Ax, Ay)}{H_\phi(x, y)} : x, y \in P_n, x \neq y \right\}$$

3 Principal Results

For this section, assume that $\phi(1, \cdot)$ is strictly convex on $(0, \infty)$, $x, y \in P_n$, $x \neq y$.

Theorem 3.1: Let A be a column-stochastic, row-allowable $m \times n$ matrix, and let $x, y \in P_n$. Then

$$H_\phi(Ax, Ay) \leq \bar{\alpha}(A) H_\phi(x, y)$$

Proof Sketch:

Theorem 4.1: $0 \leq \eta_\phi(A) \leq \bar{\alpha}(A) \leq 1$

Proof: Given theorem 3.1, it's fairly trivial. Under the assumption of strict convexity of $\phi(1, \cdot)$, by property 2 we have that $0 < H_\phi(x, y)$ and as A is a nonnegative matrix, Ax and Ay are nonnegative (possibly equal) d -vectors, so $0 \leq H_\phi(Ax, Ay)$. Clearly

$$0 \leq H_\phi(Ax, Ay), 0 < H_\phi(x, y) \implies 0 \leq \frac{H_\phi(Ax, Ay)}{H_\phi(x, y)}$$

From Theorem 3.1, we have

$$\frac{H_\phi(Ax, Ay)}{H_\phi(x, y)} \leq \bar{\alpha}(A) \implies \sup \left\{ \frac{H_\phi(Ax, Ay)}{H_\phi(x, y)} \right\} \leq \bar{\alpha}(A)$$

$\bar{\alpha}(A) \leq 1$ follows directly from property 4 of relative entropy. Thus

$$0 \leq \sup \left\{ \frac{H_\phi(Ax, Ay)}{H_\phi(x, y)} \right\} \leq \bar{\alpha}(A) \leq 1$$

□

It follows that for all A and for all strictly convex functions g , $\eta_g(A) \leq \eta_{|\omega|} = \bar{\alpha}(A)$, where $g_{|\omega|} = |\omega|$ meaning $H_g(x, y) = H(x, y)_{|\omega|} = \sum_i x_i \left| \frac{y_i}{x_i} - 1 \right|$

Theorem 5.4: If $g(w)$ is thrice differentiable in a neighborhood of 0 and $g''(0) > 0$, then $\eta_{w^2}(A) \leq \eta_g(A)$; in particular, $\eta_{w^2}(A) \leq \eta_{\log}(A)$

Proof: Because $g(\omega) = \omega^2$ is homogeneous of degree two,

$$\begin{aligned} \eta_{w^2}(A) &= \sup_{\substack{x \neq y \\ x, y \in P_n}} \frac{H_{w^2}(Ax, Ay)}{H_{w^2}(x, y)} \\ &= \sup_{\substack{x \in P_n \\ v^\top 1 = 0}} \frac{H_{w^2}(Ax, Ax + Av)}{H_{w^2}(x, x + v)} \equiv \sup_{\substack{x \in P_n \\ v^\top 1 = 0}} \frac{\Phi(Ax, Av)}{\Phi(x, v)} \end{aligned} \quad (1)$$

Where $\Phi(x, v) = H_{w^2}(x, x + v) = \sum_j \frac{v_j^2}{x_j}$ and $v \neq 0$. Now we use the fact that g is thrice

differentiable to expand it in a Taylor's series about $w = 0$.

$$\phi(s, t) = sg \left(\frac{t}{s} - 1 \right) = (t - s)g'(0) + \frac{(t - s)^2}{s} \frac{g''(0)}{2} + \frac{1}{s^2} O \left((t - s)^3 \right)$$

Pick ϵ sufficiently small and let $y_\epsilon = x + \epsilon v$. Then

$$H_g(x, y_\epsilon) = \frac{g''(0)}{2} \epsilon^2 \Phi(x, v) + O(\epsilon^3)$$

and

$$H_g(Ax, Ay_\epsilon) = \frac{g''(0)}{2} \epsilon^2 \Phi(Ax, Av) + O(\epsilon^3)$$

therefore,

$$\eta_g(A) \geq \frac{H_g(Ax, Ay_\epsilon)}{H_g(x, y_\epsilon)} = \frac{\Phi(Ax, Av)}{\Phi(x, v)} + O(\epsilon)$$

Because of equation (1), we can choose x, v and ϵ such that $\frac{\Phi(Ax, Av)}{\Phi(x, v)} + O(\epsilon)$ is arbitrarily close to $\eta_{w^2}(A)$. It follows that $\eta_{w^2}(A) \leq \eta_g(A)$ \square

4 Relation to Course

In the course, we first defined and studied the properties of KL-divergence (in fact, this is the relative entropy defined in Definition 2.1, also called H_{\log} throughout the paper), and generalized it to a family of divergences called f -divergence. This paper also forms a generalization of KL-divergence (under the name of relative entropy), but instead abstracts it to relative ϕ -entropy. These two methods of generalization can be connected, and we'll see that relative entropy is a type of f -divergence.

Let $f : (0, \infty) \rightarrow \mathbb{R}$ be convex and $f(1) = 0$. Furthermore, let P and Q be probability distributions such that $P \ll Q$. Define **f -divergence** to be

$$D_f(P||Q) := \int f \left(\frac{dP}{dQ} \right) dQ \xrightarrow{\text{discrete}} \sum_i f \left(\frac{p_i}{q_i} \right) q_i$$

where $p = (p_i)$ and $q = (q_i)$ are two d -vectors. Thus f -divergence (in the discrete case) is equivalent to the relative g -entropy if and only if $f(z) = zg \left(\frac{1}{z} - 1 \right)^3$, because

$$D_f(P||Q) = \sum_i f \left(\frac{p_i}{q_i} \right) q_i = \sum_i \frac{p_i}{q_i} g \left(\frac{1}{p_i/q_i} \right) = \sum_i p_i g \left(\frac{q_i}{p_i} - 1 \right) = H_g(P, Q)$$

Note that we slightly abuse notation thinking about P and Q as vectors of their respective probabilities indexed at i . For $z \in (0, \infty)$, $\frac{1}{z} - 1 \in (-1, \infty)$ so we have that $z \mapsto zg \left(\frac{1}{z} - 1 \right)$ is convex. And of course, $f(1) = 1g(1 - 1) = g(0) = 0$; thus, everything is well defined.

On a related note, we can also extend f divergence to be symmetric in a similar way to how the symmetric relative entropy is defined, $J_f(x, y) = D_f(x, y) + D_f(y, x)$. However, this still does not permit us to think about divergence as a distance metric, as it still does not satisfy the triangle inequality.

We were able to define a contraction coefficient for f -divergence, and we can do the same for relative ϕ -entropy. This leads to some parallels about the (strong) Data Processing Inequality that we've seen in class. In this paper we see that A is a scrambling matrix if and only if

³ g is not necessarily invertible

$\eta_\phi(A) < 1$; furthermore, we have that $\eta_\phi(A) = 1$ if A is a permutation matrix. Clearly, by definition of $\eta_\phi(A)$ we see that $H_\phi(Ax, Ay) \leq H_\phi(x, y)$ with equality if and only if A is a permutation matrix. In class we saw the DPI in the context of divergence,

$$D(P_{Y|X}P_X||P_{Y|X}Q_X) \leq D(P_X||Q_X)$$

Because of how the contraction coefficient was defined in this paper, we can easily set $P_{X|Y} = A$ and $P_{Y|Z} = A$, then using the notation from class,

$$X \xrightarrow{A} Y \xrightarrow{A'} Z \implies H_\phi(X, Y) \leq H_\phi(X, Z)$$

where the equality only holds if both A and A' are permutation matrices.

5 Conclusion

We learned KL-divergence as a way to extend the concept of Shannon entropy. To expand this idea even further, we utilize a continuous homogeneous function $\phi(x, y)$ convex in both arguments satisfying $\phi(1, 1) = 0$. Then the calculation is simple, $\sum_i \phi(x_i, y_i)$. The idea is that $\phi(x_i, y_i)$ tells us something meaningful about the difference between each pair of elements in the probability vectors P and Q . If we think about the function H_ϕ in the equivalent sense, $H_g = \sum_i x_i g\left(\frac{y_i}{x_i} - 1\right)$, the relative entropy is really just an average with respect to P , weighted by $g\left(\frac{y_i}{x_i} - 1\right)$. Properties of ϕ such as continuity, homogeneity, and convexity allows for many of the properties of the classical relative entropy still hold under this generalization, which allows for a meaningful contraction coefficient, η_ϕ . This allows us to speak of data processing inequalities with respect to different measures of relative entropy. In this paper, we were able to set upper and lower bounds on η_ϕ with respect to a stochastic matrix A . One can imagine that with an upper and lower bound on the contraction coefficient, the effect of A on probability vectors x, y can be assessed in much a more meaningful fashion. One example would be bounding the rates of convergence to equilibrium of Markov processes, as mentioned in the introduction.