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

The properties of relative entropy have been extensively studied in the past in many different contexts. The focus of this paper on its properties under mappings by stochastic matrices can provide more information on bounding the rates of convergence to equilibrium of ergodic Markov chains and Markov processes

2. Background

Preliminary Definitions

Definitions 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. We define

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

Definitions 1.2 (Matrices): A **(column) stochastic** $m \times n$ matrix is a matrix whos columns belong to N_m. 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 $m \times n$ matrix is called a **scrambling** matrix if any submatrix consisting of two columns has a row both elements of which are positive. Note, every row-positive matrix is scrambling, but not conversely.

Main Definitions

Definition 2.1 (Symmetric Relative Entropy): For any two positive d-vectors $x = (x_i)$ and $y = (y_i)$, whether or not x and y are probability vectors, we define the *relative entropy* 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, $x = (x_i), y = (y_i)$, whether or not x and y are probability vectores, we defind the *relativeϕ-entropy* as

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

and the *symmetric relative ϕ-entropy* as

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

Because $\tilde{\phi}(a, b) = \phi(a, b) + \phi(b, a)$ satisfies the conditions of 2 if ϕ does, and $J_\phi(x, y) = H_{\tilde{\phi}}(x, y)$, from this point on we will speak of J_ϕ as $H_{\tilde{\phi}}$

The function ϕ defined in definition 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 **??**. That is

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

Three main properties of relative entropy:

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

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

and also satisfies

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

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

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

Theorem 4.1: $0 \leq \eta_\phi(A) \leq \bar{\alpha}(A) \leq 1$
Proof: Given theorem 3.1, it's 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

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

From Theorem 3.1 (**ref this**), 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 (**ref this**) of relative entropy. Thus

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

$\square$

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:

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

Where $\Phi(x,v) = H_{\omega^2}(x, x+v) = \sum_j \frac{v_j^2}{x_j}$ 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 $n_{\omega^2}(A)$. It follows that $\eta_{w^2}(A) \leq \eta_g(A)$ $\square$

4. Relation to Course

- compare with what we've seen with DPI from KL-divergence and mutual information. Briefly, we have

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

  and also

$$\begin{aligned}
A \text{ is scrambling} &\iff \eta_\phi(A) < 1 \\
&\iff H_\phi(x,y) < H_\phi(Ax, Ay) \quad \text{with equality if } A \text{ is permutation}
\end{aligned}$$

  So

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

Seems like this connects easily to what we've seen if we define $P_{X|Y} \equiv A$ and $P_{Y|Z} \equiv A'$

- Compare relative g-entropy with f-divergence
  - Extend symmetry to f-divergence

$$J_\phi(x, y) = H_\phi(x, y) + H_\phi(y, x) \longrightarrow J_f(x, y) = D_f(x, y) + H_f(y, x)$$

5. Conclusion