

CONVERGENCE OF MARKOV CHAINS IN INFORMATION DIVERGENCE

P. HARREMOËS AND K. K. HOLST

ABSTRACT. Information theoretic methods are used to prove convergence in information divergence of reversible Markov chains. Also some ergodic theorems for information divergence are proved.

1. INTRODUCTION AND PRELIMINARIES

Relating results from probability theory and information theory is not a new idea. Some convergence theorems in probability theory can be reformulated as "the entropy converges to its maximum". A. Rényi [Rén61] used information divergence to prove convergence of Markov chains to equilibrium on a finite state space. Later I. Csiszár [Csi63] and Kendall [Ken64] extended Rényi's method to provide convergence on countable state spaces. Their proofs use basically that information divergence is an Csiszár f -divergence. Later Fritz [Fri73] used information theoretic arguments to establish convergence of reversible Markov chains in total variation. Recently Barron [Bar00] improved Fritz' method and proved convergence in information. Other limit theorems have been proved using information theoretic methods. The Central Limit Theorem was treated by Linnik and A. Barron [Bar86], the Local Central Limit Theorem was treated by S. Takano [Tak87] and Poisson's law was treated by P. Harremoës [Har01]. There has also been work strengthening weak or strong convergence to convergence in information divergence. All the above mentioned papers have results of this kind, but also work by A. Barron [Bar00] should be mentioned. Some work has also been done where the limit of a sequence is identified as an information projection. The most important paper in this direction is due to I. Csiszár [Csi84].

Here we shall establish convergence in information divergence for a large class of Markov chains. The result will include the results of Rényi, Csiszár, Kendall and Fritz. The case where no convergence takes place will be studied from the point of view of ergodic theory, and some of the classical results of Birkhoff will be derived. The basic result is that information divergence is continuous under the formation of the intersection of a decreasing sequence of σ -algebras. The same technique can be used to obtain a classical result of Pinsker [Pin60] about continuity under an increasing sequence of σ -algebras.

Date: July 25, 2007.

2000 Mathematics Subject Classification. Primary 60J10, 94A15; Secondary 60B11, 60F15.

Key words and phrases. Information divergence, increasing information, decreasing information, Markov chain, reversible Markov chain, ergodic theorems.

P. Harremoës has supported by the Villum Kann Rasmussen Foundation, The Danish Natural Research Council and INTAS (project 738-00).

Let P and Q be probability measures. Then the *information divergence from P to Q* is defined by

$$D(P\|Q) = \begin{cases} \int \log \frac{dP}{dQ} dP & \text{if } P \ll Q, \\ \infty & \text{otherwise.} \end{cases}$$

This quantity is also called the *Kullback-Leibler discrimination* or *relative entropy*. Information divergence does not define a metric, but is related to total variation via *Pinsker's inequality* $\frac{1}{2} \|P - Q\|^2 \leq D(P\|Q)$ proved by I. Csiszár [Csi67] and others. If $(P_n)_{n \in \mathbb{N}}$ is a sequence of probability distributions, we say that $(P_n)_{n \in \mathbb{N}}$ converges to Q in information if $D(P_n\|Q) \rightarrow 0$ for $n \rightarrow \infty$. Pinsker's inequality shows that convergence in information is a stronger condition than convergence in total variation. See [Har07] for details about topologies related to information divergence.

The following proposition was first formulated by F. Topsøe in 1967 [Top74] and has a purely computational proof.

Proposition 1. *Let P_1, P_2, \dots, P_n be distributions and let (p_1, p_2, \dots, p_n) be a probability vector. Then*

$$\sum p_i D(P_i\|Q) = D\left(\sum p_i P_i\|Q\right) + \sum p_i D\left(P_i\|\sum p_i P_i\right).$$

The function $P \rightarrow D(P\|Q)$ is strict convex in the first variable.

Corollary 1. *Let P_1, P_2, \dots, P_n be distributions and let (p_1, p_2, \dots, p_n) be a probability vector. Then*

$$D\left(\sum p_i P_i\|Q\right) \geq \sum p_i D(P_i\|Q) - H(p_1, p_2, \dots, p_n),$$

and

$$D(P_1\|Q) \leq \frac{D(\sum p_i P_i\|Q) + H(p_1, 1 - p_1)}{p_1}.$$

Proof. We have to prove that $\sum p_i D(P_i\|\sum p_i P_i) \leq H(p_1, p_2, \dots, p_n)$. Writing $H(p_1, p_2, \dots, p_n) = \sum p_i \log\left(\frac{1}{p_i}\right)$ we see that this follows since

$$D\left(P_i\|\sum p_i P_i\right) \leq \log\left(\frac{1}{p_i}\right).$$

□

For a set of probability measures C put $D(C\|Q) = \inf_{P \in C} D(P\|Q)$. A sequence $(P_n)_{n \in \mathbb{N}} \subseteq C$ is said to be asymptotically optimal if $D(P_n\|Q) \rightarrow D(C\|Q)$ for $n \rightarrow \infty$. The next theorem was formulated and proved by I. Csiszár [Csi75] and F. Topsøe [Top79].

Theorem 1. *Let C be a convex set of probability measures and let Q be a probability measure such that $D(C\|Q) < \infty$. Then there exists a unique distribution $\Pi_{C \leftarrow Q}$ such that $P_n \rightarrow \Pi_{C \leftarrow Q}$ in information for any asymptotically optimal sequence $(P_n)_{n \in \mathbb{N}}$. Furthermore, for every $P \in C$, we have*

$$\text{(Pythagorean Inequality)} \quad D(P\|Q) \geq D(P\|\Pi_{C \leftarrow Q}) + D(C\|Q).$$

The probability measure $\Pi_{C \leftarrow Q}$ is called the generalized information projection of Q onto C and is illustrated on Figure 1. The notation for the generalized information projection is taken from [CM03].

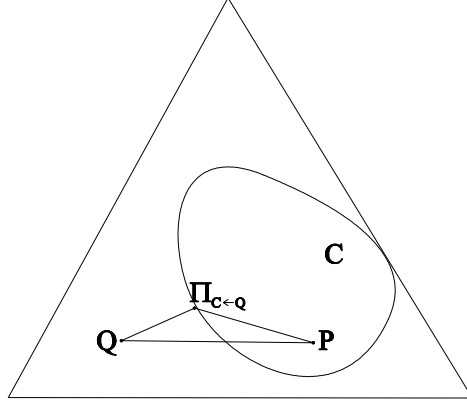


FIGURE 1. The Pythagorean inequality.

2. INCREASING AND DECREASING INFORMATION

In general information divergence is lower semi-continuous, but it is continuous under special conditions. Increasing and decreasing σ -algebras are important examples where continuity holds.

Let A be a set with a σ -algebra \mathbb{G} . The set of probability measures on (A, \mathbb{G}) is denoted $M_+^1(A, \mathbb{G})$ or sometimes $M_+^1(A)$ for short. Let \mathbb{F} be any sub-algebra of \mathbb{G} . Let $C_{\mathbb{F}}$ be the set

$$\{R \in M_+^1(A, \mathbb{G}) \mid R(B) = P(B) \text{ for all } B \in \mathbb{F}\}.$$

For all $R \in C_{\mathbb{F}}$ we have $R|_{\mathbb{F}} = P|_{\mathbb{F}}$ where $|_{\mathbb{F}}$ denotes the restriction of a measure to a subalgebra. Then

$$\begin{aligned} D(R\|Q) &= D(R|_{\mathbb{F}}\|Q|_{\mathbb{F}}) + D(R\|Q \mid \mathbb{F}) \\ &\geq D(P|_{\mathbb{F}}\|Q|_{\mathbb{F}}) \end{aligned}$$

with equality if and only if $D(R\|Q \mid \mathbb{F}) = 0$. This is the so-called *data reduction inequality*. Therefore

$$D(P|_{\mathbb{F}}\|Q|_{\mathbb{F}}) = D(C_{\mathbb{F}}\|Q).$$

We also observe that

$$\frac{d\Pi_{C_{\mathbb{F}}\leftarrow Q}}{dQ} = \frac{dP|_{\mathbb{F}}}{dQ|_{\mathbb{F}}} \in L^1(A, \mathbb{F}, Q).$$

Here we shall see how the divergence $D(C\|Q)$ and the projection $\Pi_{C\leftarrow Q}$ changes when the set C is changed.

Lemma 1. *Let C_1 be a subset of C_2 . Then*

$$D(C_1\|Q) \geq D(\Pi_{C_1\leftarrow Q}\|\Pi_{C_2\leftarrow Q}) + D(C_2\|Q).$$

Proof. Then, obviously $D(C_1\|Q) \geq D(C_2\|Q)$. Assume that $D(C_1\|Q) < \infty$. Let P_n be an asymptotically optimal sequence in C_1 . Then

$$D(P_n\|Q) \geq D(P_n\|\Pi_{C_2\leftarrow Q}) + D(C_2\|Q).$$

By lower semi continuity

$$D(C_1\|Q) \geq D(\Pi_{C_1\leftarrow Q}\|\Pi_{C_2\leftarrow Q}) + D(C_2\|Q).$$

□

Lemma 2. *Let $(C_\pi)_{\pi \in \Pi}$ be a decreasing net of convex I -closed sets such that $(D(C_\pi \| Q))_{\pi \in \Pi}$ is uniformly bounded. Let C_∞ denote the intersection $\bigcap_{\pi \in \Pi} C_\pi$.*

Then $D(C_\infty \| Q)$ is finite and

$$D(C_\pi \| Q) \rightarrow D(C_\infty \| Q).$$

Then $\Pi_{C_\pi \leftarrow Q}$ converge to $\Pi_{C_\infty \leftarrow Q}$ and

$$D(\Pi_{C_\pi \leftarrow Q} \| Q) \rightarrow D(\Pi_{C_\infty \leftarrow Q} \| Q)$$

and $D(C_\infty \| Q) = D(\Pi_{C_\infty \leftarrow Q} \| Q)$.

Proof. If $C_{\pi_2} \subseteq C_{\pi_1}$ Lemma 1 implies that

$$D(C_{\pi_2} \| Q) \geq D(\Pi_{C_{\pi_2} \leftarrow Q} \| \Pi_{C_{\pi_1} \leftarrow Q}) + D(C_{\pi_1} \| Q).$$

Then $D(\Pi_{C_{\pi_2} \leftarrow Q} \| \Pi_{C_{\pi_1} \leftarrow Q})$ converges to 0 because the net $D(C_\pi \| Q)$ is uniformly bounded. Hence $\Pi_{C_\pi \leftarrow Q}$ is a Cauchy net in total variation and converges to a distribution Q_∞ . The set C_π is closed and therefore $Q_\infty \in C_\pi$ and Q_∞ is also an element in the intersection C_∞ , and in particular $D(C_\infty \| Q) \leq D(Q_\infty \| Q)$. By lower semi continuity we have

$$\begin{aligned} D(Q_\infty \| Q) &\leq \liminf D(\Pi_{C_\pi \leftarrow Q} \| Q) \\ &= \liminf D(C_\pi \| Q) \\ &\leq D(C_\infty \| Q). \end{aligned}$$

□

Lemma 3. *Let $(C_\pi)_{\pi \in \Pi}$ be an increasing net of convex sets and assume*

$$D(C_\pi \| Q) < \infty$$

for all π . Let C_∞ denote the union $\bigcup_{\pi \in \Pi} C_\pi$. Then

$$D(\Pi_{C_\pi \leftarrow Q} \| \Pi_{C_\infty \leftarrow Q}) \rightarrow 0.$$

Proof. We have $C_\pi \subseteq C_\infty$ and according to Lemma 1

$$D(C_{\pi_n} \| Q) \geq D(\Pi_{C_\pi \leftarrow Q} \| \Pi_{C_\infty \leftarrow Q}) + D(C_\infty \| Q).$$

Now, let P_n be asymptotically optimal in C_∞ . Then there exist C_{π_n} such that $P_n \in C_{\pi_n}$ and

$$D(C_{\pi_n} \| Q) \leq D(P_n \| Q) \rightarrow D(C_\infty \| Q) \text{ for } n \rightarrow \infty.$$

Therefore $D(\Pi_{C_\pi \leftarrow Q} \| Q) \rightarrow D(C_\infty \| Q)$ and $D(\Pi_{C_\pi \leftarrow Q} \| \Pi_{C_\infty \leftarrow Q}) \rightarrow 0$. □

Theorems related to increasing sequences of σ -algebras have been proved in [Dob59], [Dob60], [Pin60] and [Bar00]. The proof presented here is new.

Theorem 2. *Let $\mathbb{F}_1 \subseteq \mathbb{F}_2 \subseteq \dots$ be an increasing sequence of sub- σ -algebras on a measurable set A . Let \mathbb{F}_∞ denote the σ -algebra generated by the union $\bigcup_{n=1}^{\infty} \mathbb{F}_n$. Let*

P and Q be probability measures on A . Then

$$D(P_{|\mathbb{F}_n} \| Q_{|\mathbb{F}_n}) \nearrow D(P_{|\mathbb{F}_\infty} \| Q_{|\mathbb{F}_\infty}) \text{ for } n \rightarrow \infty.$$

Proof. The sequence $D(P_{|\mathbb{F}_n} \| Q_{|\mathbb{F}_n})$ is increasing. Put

$$D_\infty = \lim_{n \rightarrow \infty} D(P_{|\mathbb{F}_n} \| Q_{|\mathbb{F}_n}).$$

The inequality

$$D_\infty \leq D(P_{|\mathbb{F}_\infty} \| Q_{|\mathbb{F}_\infty})$$

follows from the data reduction inequality. If $D_\infty = \infty$ then the convergence is obvious. Assume that $D_\infty < \infty$.

Now, $C_{\mathbb{F}_n}$ is decreasing in n and the intersection equals \mathbb{F}_∞ . Further the sets $C_{\mathbb{F}_n}$ are convex and closed, and therefore Lemma 2 applies:

$$\begin{aligned} D(P_{|\mathbb{F}_n} \| Q_{|\mathbb{F}_n}) &= D(C_{\mathbb{F}_n} \| Q) \\ &\rightarrow D(C_{\mathbb{F}_\infty} \| Q) = D(P_{|\mathbb{F}_\infty} \| Q_{|\mathbb{F}_\infty}) \end{aligned}$$

for $n \rightarrow \infty$. □

The following theorem is due to Barron [Bar00]. He used so-called absolute divergence in the proof. Here only the usual information divergence and lower semi continuity is used.

Theorem 3. *Let $\mathbb{F}_1 \supseteq \mathbb{F}_2 \supseteq \dots$ be an decreasing sequence of sub- σ -algebras of a measurable space A . Let \mathbb{F}_∞ denote the intersection $\bigcap_{n=1}^{\infty} \mathbb{F}_n$. Let P and Q be probability measures on A . If there exists m such that $D(P_{|\mathbb{F}_m} \| Q_{|\mathbb{F}_m}) < \infty$, then*

$$D(P_{|\mathbb{F}_n} \| Q_{|\mathbb{F}_n}) \searrow D(P_{|\mathbb{F}_\infty} \| Q_{|\mathbb{F}_\infty}) \text{ for } n \rightarrow \infty.$$

Proof. The sequence $D(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n})$ is decreasing. Without loss of generality we may assume that $D(P \| Q) < \infty$.

We have to show that

$$D(C_{\mathbb{F}_n} \| Q) \rightarrow D(C_{\mathbb{F}_\infty} \| Q).$$

For any $P \in C_{\mathcal{F}_n}$ we have

$$D(P \| Q) = D(P \| \Pi_{C_{\mathbb{F}_\infty} \leftarrow Q}) + D(C_{\mathbb{F}_\infty} \| Q),$$

implying that

$$\begin{aligned} D(C_{\mathbb{F}_n} \| Q) &= D(\Pi_{C_{\mathbb{F}_n} \leftarrow Q} \| Q) \\ &= D(\Pi_{C_{\mathbb{F}_n} \leftarrow Q} \| \Pi_{C_{\mathbb{F}_\infty} \leftarrow Q}) + D(C_{\mathbb{F}_\infty} \| Q). \end{aligned}$$

Therefore it is sufficient to prove that

$$D(\Pi_{C_{\mathbb{F}_n} \leftarrow Q} \| \Pi_{C_{\mathbb{F}_\infty} \leftarrow Q}) \rightarrow 0$$

for $n \rightarrow \infty$.

The sequence $C_{\mathbb{F}_n}$ is increasing. Define $C_\infty = \bigcup C_{\mathbb{F}_n}$ and

$$\rho_n = \frac{d(\Pi_{C_{\mathbb{F}_n} \leftarrow Q})}{dQ}$$

and

$$\rho_\infty = \frac{d(\Pi_{C_\infty \leftarrow Q})}{dQ}.$$

Now

$$D(\Pi_{C_{\mathbb{F}_n} \leftarrow Q} \| \Pi_{C_\infty \leftarrow Q}) \rightarrow 0$$

and by Pinsker's inequality

$$\int |\rho_n - \rho_\infty| dQ = \|\Pi_{C_{\mathbb{F}_n} \leftarrow Q} - \Pi_{C_\infty \leftarrow Q}\| \rightarrow 0$$

for $n \rightarrow \infty$. Thus $\rho_n \rightarrow \rho_\infty$ in $L^1(\Omega, Q)$. For $m \leq n$ we have $\rho_n \in L^1(\Omega, \mathbb{F}_m, Q)$ and therefore $\rho_\infty \in L^1(\Omega, \mathbb{F}_m, Q)$ implying that $\rho_\infty \in L^1(\Omega, \mathbb{F}_\infty, Q)$. For $B \in \mathbb{F}_\infty$ we also have $\Pi_{C_\infty \leftarrow Q}(B) = \lim_{n \rightarrow \infty} \Pi_{C_{\mathbb{F}_n} \leftarrow Q}(B) = P(B)$. That implies that $\Pi_{C_\infty \leftarrow Q} = \Pi_{C_{\mathbb{F}_\infty} \leftarrow Q}$.

According to the previous theorem

$$D(\Pi_{C_{\mathbb{F}_n} \leftarrow Q} \|\Pi_{C_{\mathbb{F}_\infty} \leftarrow Q}) \rightarrow 0 \text{ for } n \rightarrow \infty,$$

and we just have to remark that

$$D(\Pi_{C_{\mathbb{F}_n} \leftarrow Q} \|\Pi_{C_{\mathbb{F}_\infty} \leftarrow Q}) = D(\Pi_{C_{\mathbb{F}_n} \leftarrow Q} \|\Pi_{C_{\mathbb{F}_\infty} \leftarrow Q}).$$

□

3. MARKOV CHAINS

Let Φ_x be a Markov kernel $A \rightarrow M_+^1(B)$. Then Φ shall denote the *Markov operator* $M_+^1(A) \rightarrow M_+^1(B)$ given by

$$\Phi(P)(G) = \int_A \Phi_x(G) dPx.$$

For a Markov operator $\Phi : M_+^1(A) \rightarrow M_+^1(B)$ and a distribution $P \in M_+^1(A)$ we get a distribution \tilde{P} on $A \times B$ given by

$$\tilde{P}(G \times H) = \int_G \Phi_x(H) dPx.$$

Then P is the marginal distribution of \tilde{P} on A . If X is the projection $A \times B \rightarrow B$ then

$$D(P\|Q) = D(\Phi(P) \|\Phi(Q)) + D(\tilde{P} \|\tilde{Q} \mid X).$$

Using the positivity of conditional divergence we get

$$(3.1) \quad D(\Phi(P) \|\Phi(Q)) \leq D(P\|Q).$$

This inequality is the *noisy data processing inequality*, and the Markov kernel is considered as a noisy data processing. The data processing inequality holds with equality if and only if $D(\tilde{P} \|\tilde{Q} \mid X) = 0$, i.e. $\tilde{Q}(\cdot \mid X = x)$ almost surely with respect to $\Phi(P)$. If A and B are discrete this means that

$$\frac{\Phi_a(x) \cdot P(a)}{\Phi(P)(x)} = \frac{\Phi_a(x) \cdot Q(a)}{\Phi(Q)(x)}.$$

If all $\Phi_a(x) > 0$ then this is equivalent to

$$\frac{P(a)}{Q(a)} = \frac{\Phi(P)(x)}{\Phi(Q)(x)}$$

and $\frac{P(a)}{Q(a)}$ is independent of a . Therefore $\frac{P(a)}{Q(a)}$ is constant and this constant must be 1 and $P = Q$.

Let (Ω, \mathbb{F}, Q) be a probability space and let (X_1, X_2, \dots) be a Markov chain with state space A . Let \mathbb{G}_n be the σ -algebra generated by X_1, X_2, \dots, X_n and let \mathbb{G}_∞ be the σ -algebra generated by the union $\bigcup_{n=1}^{\infty} \mathbb{G}_n$. The Markov property is that

$$Q(X_n \in B \mid \mathbb{G}_m) = Q(X_n \in B \mid X_m)$$

for all $n \geq m \geq 1$ and all $B \in \mathbb{B}$. The probability distribution $Q|_{\mathbb{G}_1}$ is the initial distribution.

There is a close connection between the divergence of probability distributions on the state space of a Markov chain and the divergence of the corresponding Markov chains. Let P be a probability distribution on (Ω, \mathbb{B}) and assume that (X_1, X_2, \dots) is also a Markov chain with respect to P and with the same Markov kernel Φ . Then

$$\begin{aligned} D(P|_{\mathbb{G}_\infty} \parallel Q|_{\mathbb{G}_\infty}) &= \sup_n D(P|_{\mathbb{G}_n} \parallel Q|_{\mathbb{G}_n}) \\ &= \sup_n \sum_{i=0}^{n-1} D(P|_{\mathbb{G}_{i+1}} \parallel Q|_{\mathbb{G}_{i+1}} \mid X_1, X_2, \dots, X_i) \\ &= \sup_n D(P|_{\mathbb{G}_1} \parallel Q|_{\mathbb{G}_1}) + \sum_{i=1}^{n-1} D(P|_{\mathbb{G}_{i+1}} \parallel Q|_{\mathbb{G}_{i+1}} \mid X_1, X_2, \dots, X_i) \\ &= D(P|_{\mathbb{G}_1} \parallel Q|_{\mathbb{G}_1}). \end{aligned}$$

4. CONVERGENCE OF MARKOV CHAINS

Let X_1, X_2, \dots be a homogeneous Markov chain with Markov operator Φ , and let \mathbb{F}_n be the σ -algebra generated by $X_n, X_{n+1}, X_{n+2}, \dots$. The intersection $\mathbb{F}_\infty = \bigcap_{n=1}^{\infty} \mathbb{F}_n$ is the tail σ -algebra. A probability measure P is said to be invariant if $\Phi P = P$.

Theorem 4. *Let Φ be a transition operator on a state space A with an invariant probability measure Q . If $D(P \parallel Q) < \infty$ then there exists a probability measure P^* such that $D(\Phi^n P \parallel \Phi^n P^*) \rightarrow 0$ and $D(\Phi^n P^* \parallel Q)$ is constant. The density $\frac{d\Phi^n P}{d\Phi^n P^*}$ converges to 1 pointwise almost surely, i.e. the probability with respect to P^* that*

$$\frac{d\Phi^n P}{d\Phi^n P^*}(X_n) \rightarrow 1 \text{ for } n \rightarrow \infty$$

for an infinite sample path (X_1, X_2, X_3, \dots) is 1.

Proof. According to Theorem 3

$$\begin{aligned} D(\Phi^n P \parallel Q) &= D(P|_{\mathbb{F}_n} \parallel Q|_{\mathbb{F}_n}) \\ &\searrow D(P|_{\mathbb{F}_\infty} \parallel Q|_{\mathbb{F}_\infty}) \text{ for } n \rightarrow \infty. \end{aligned}$$

Let \mathbb{E} be the conditional expectation from the tail- σ -algebra \mathbb{F}_∞ to \mathbb{F}_1 equipped with the measure generated by Q . Let P^* be the probability measure on \mathbb{F}_1 given by $\frac{dP^*}{dQ} = \frac{dP|_{\mathbb{F}_\infty}}{dQ|_{\mathbb{F}_\infty}}$. Then

$$D_{\mathbb{F}_\infty}(P \parallel Q) = D(P^* \parallel Q).$$

We also see that

$$D(\Phi^n P^* \parallel Q) = D(P^* \parallel Q),$$

and

$$\begin{aligned} D(\Phi^n P \parallel \Phi^n P^*) &= D\left(P_{|\mathbb{F}_n} \parallel P_{|\mathbb{F}_n}^*\right) \\ &\searrow D\left(P_{|\mathbb{F}_\infty} \parallel P_{|\mathbb{F}_\infty}^*\right) \\ &= 0. \end{aligned}$$

In order to prove almost sure pointwise convergence we note that $\frac{d\Phi^n P}{d\Phi^n P^*}(X_n)$ is a nonnegative martingale with respect to the probability measure on sequences induced by P^* . According to results from [Har05] it implies that

$$\gamma\left(E\left[\sup_{n \geq N} \frac{d\Phi^n P}{d\Phi^n P^*}(X_n)\right]\right) \leq D(\Phi^N P \parallel \Phi^N P^*)$$

and

$$\gamma\left(E\left[\inf_{n \geq N} \frac{d\Phi^n P}{d\Phi^n P^*}(X_n)\right]\right) \leq D(\Phi^N P \parallel \Phi^N P^*)$$

where $\gamma(t) = t - 1 - \log(t)$, $t > 0$. In particular we see that

$$E\left[\sup_{n \geq N} \frac{d\Phi^n P}{d\Phi^n P^*}(X_n)\right] \rightarrow 0 \text{ for } N \rightarrow \infty$$

and

$$E\left[\inf_{n \geq N} \frac{d\Phi^n P}{d\Phi^n P^*}(X_n)\right] \rightarrow 0 \text{ for } N \rightarrow \infty.$$

Therefore

$$\sup_{n \geq N} \frac{d\Phi^n P}{d\Phi^n P^*}(X_n) - \inf_{n \geq N} \frac{d\Phi^n P}{d\Phi^n P^*}(X_n) \rightarrow 0 \text{ for } N \rightarrow \infty$$

and this implies pointwise convergence almost surely. \square

Theorem 4 was used in [Har06] to prove that convolutions of identical probability measures on a compact group converges to the Haar probability measure.

Theorem 5. *Let Φ be a transition operator on a countable state space A with an invariant probability measure Q . Assume that there exists an n such that $\Phi_a^n(b) > 0$ for all $a, b \in A$. If $D(P \parallel Q) < \infty$ then $D(\Phi^n P \parallel Q) \rightarrow 0$.*

Proof. According to Theorem 4 there exists a probability measure P^* such that $D(\Phi^n P \parallel \Phi^n P^*) \rightarrow 0$ and $D(\Phi^n P^* \parallel Q)$ is constant. In particular

$$D(P^* \parallel Q) = D(\Phi P^* \parallel Q).$$

Now $\Phi_a^n(b) > 0$ for all $a, b \in A$ and therefore $P^* = Q$. Thus

$$D(\Phi^n P \parallel Q) \searrow D(P^* \parallel Q) = 0.$$

\square

Let Φ denote the transition operator with the Markov kernel $x \rightarrow \Phi_x$.

Definition 1. *Let Q be a probability measure on (A, \mathbb{F}) . Then Φ is Q -reversible if*

$$(4.1) \quad \int_{F_1} \Phi_x(F_2) dQ(x) = \int_{F_2} \Phi_x(F_1) d\Phi Q(x), \quad F_1, F_2 \in \mathbb{F}.$$

Remark that (4.1) is equivalent with

$$\int_A \int_A \phi(x, y) d\Phi_x(y) dQ(x) = \int_A \int_A \phi(x, y) d\Phi_y d\Phi Q(y)$$

if ϕ is the indicator function of $F_1 \times F_2$. Therefore it also holds for any measurable function on the product space. For an interpretation of the concept of reversibility consider the Markov chain $(X_n)_{n \in \mathbb{N}}$, with probability distribution being the unique invariant probability distribution P , and which is P -reversible. Then

$$\begin{aligned} \hat{P}_Q((X_1, X_2) \in F_1 \times F_2) &= \int_{F_1} \Phi_x(F_2) dQ(x), \\ \hat{P}_Q((X_2, X_1) \in F_1 \times F_2) &= \int_{F_2} \Phi_x(F_1) d\Phi Q(x) = \int_{F_2} \Phi_x(F_1) dQ(x). \end{aligned}$$

Therefore (X_1, X_2) has the same distribution as (X_2, X_1) .

Remark 1. *A reversible Markov chain need not be positive recurrent. The simplest example is a Markov chain on a set with 2 elements where the Markov Φ kernel acts by permutation of the elements. Then Φ^{2n} is the identity and all probability vectors are invariant, but only $(1/2, 1/2)$ is invariant under Φ .*

In [Fri73] Fritz implicitly proved the following result as part of his Theorem 1.

Theorem 6. *Assume that Φ is a Q -reversible transition operator, and Q is invariant and let P a probability measure on A such that $D(P \parallel Q) < \infty$. Then*

$$(4.2) \quad D(P \parallel \Phi^2 P) \leq D(P \parallel Q) - D(\Phi P \parallel Q).$$

Using this result it is easy to prove the following theorem which gives convergence in information where Fritz [Fri73] only got convergence in total variation.

Theorem 7. *If Φ is Q -reversible and P is absolutely continuous with respect to Q where P is a probability measure such that $D(\Phi^n P \parallel \Phi^n Q) < \infty$ eventually then there exists probability measures P_{even} and P_{odd} such that*

$$(4.3) \quad \lim_{m \rightarrow \infty} D(\Phi^{2m} P \parallel P_{\text{even}}) = 0$$

$$(4.4) \quad \lim_{m \rightarrow \infty} D(\Phi^{2m+1} P \parallel P_{\text{odd}}) = 0.$$

The probability measures P_{even} and P_{odd} satisfy $\Phi P_{\text{even}} = P_{\text{odd}}$ and $\Phi P_{\text{odd}} = P_{\text{even}}$.

Proof. Assume that Q is invariant. Theorem 4 implies that there exists a probability measure P^* such that $D(\Phi^n P \parallel \Phi^n P^*) \rightarrow 0$ and $D(\Phi^n P^* \parallel Q)$ is constant. According to inequality (4.2)

$$\begin{aligned} D(P^* \parallel \Phi^2 P^*) &\leq D(P^* \parallel Q) - D(\Phi P^* \parallel Q) \\ &= 0. \end{aligned}$$

Thus $P^* = \Phi^2 P^*$. Put $P_{\text{even}} = P^*$ and $P_{\text{odd}} = \Phi P^*$.

If Q is not invariant one can replace Q by $1/2 \cdot Q + 1/2 \cdot \Phi Q$. □

The methods developed here also have applications to Markov chains with continuous time $t \in [0; \infty[$. If P and Q are probability measures with $D(P \parallel Q) < \infty$ and $(\Phi^t)_{t \in [0; \infty[}$ then Theorem 3 implies that $t \rightsquigarrow D(\Phi^t P \parallel \Phi^t Q)$ is continuous from the left and Theorem 2 implies that $t \rightsquigarrow D(\Phi^t P \parallel \Phi^t Q)$ is continuous from the right.

Example 1. Let X be a random variable with mean zero and variance 1, and let Z be an independent normal random variable with mean zero and variance 1. Let X_t be the random variable $e^{-t}X + (1 - e^{-2t})^{1/2}Z$ for $t \in [0; \infty[$, and put $X_\infty = Z$. The map of (t, X) into the distribution of X_t can be considered as an action of an Ornstein-Uhlenbeck semigroup. Let $D(X)$ denote the divergence from the distribution of X to the distribution of Z . Then our results implies that $t \rightsquigarrow D(X_t)$ is continuous on $[0; \infty]$. This result is needed in the proof of the formula

$$D(X) = \int_0^\infty J(X_t) dt,$$

where $J(X_t)$ is the normalized Fisher information of X [Bar86], [BE85], [ABBN04].

5. ERGODICITY

Let Φ be a Markov operator with invariant measure Q . If $D(\Phi^n P \| Q)$ is finite and constant one may ask what the relation is between the sequence $\Phi^n(P)$ and the invariant measure Q . Ergodic theory tell us to what extent the average value of $\Phi^i(P)$, $i = 1, \dots, n$ converges to Q . Now this question will be explored using information divergence to measure how much one probability measure deviates from another one.

Lemma 4. Let (Ω, \mathbb{F}, P) be a probability space, and let Φ be a Markov operator. If Q is invariant and K is an invariant convex set of distributions on (Ω, \mathbb{B}) with $D(K \| Q) < \infty$, then $\Pi_{K \leftarrow Q}$ is invariant.

Proof. Let $P_n \in K$ be a sequence such that $D(P_n \| Q) \rightarrow D(K \| Q)$. Then Theorem 1 implies that $P_n \rightarrow \Pi_{K \leftarrow Q}$ in information and, by continuity, $\Phi P_n \rightarrow \Phi \Pi_{K \leftarrow Q}$ for $n \rightarrow \infty$. Since Q is invariant $D(\Phi P_n \| Q) = D(\Phi P_n \| \Phi(Q)) \leq D(P_n \| Q)$ and thus $\Phi P_n \rightarrow \Pi_{K \leftarrow Q}$ in information for $n \rightarrow \infty$. The set of probability measures is a Hausdorff space with the information topology [Har07] and therefore the limit points of the sequence $(\Phi P_n)_{n \in \mathbb{N}}$ are equal, i.e. $\Phi \Pi_{K \leftarrow Q} = \Pi_{K \leftarrow Q}$. \square

Theorem 8. Let (Ω, \mathbb{F}, P) be a probability space, and let Φ be a Markov operator. Define $\bar{P}_n = \frac{1}{n} \sum_{i=0}^{n-1} \Phi^i(P)$. Then the following 2 conditions are equivalent:

- (1) The sequence \bar{P}_n converges strongly in information.
- (2) There exists a Φ -invariant distribution Q and a $k > 0$ such that $D(\bar{P}_k \| Q)$ is finite.

Let K be the set $\text{conv} \{ \Phi^i(P) \}$, and L be the set of invariant distributions. If the conditions holds then the limit P^* is invariant and equals the I-projection $\Pi_{K \leftarrow Q}$ and equals the reversed I-projection of P on L . Finally the sequence $nD(\bar{P}_n \| P^*)$ is sub-additive.

Proof. 1 \Rightarrow 2 :

First observe that $D(\bar{P}_n \| P^*)$ is finite if and only if $D(P \| P^*)$ is finite.

To see that P^* is invariant observe that $D(\Phi\bar{P}_n \parallel \Phi P^*) \leq D(\bar{P}_n \parallel P^*) \rightarrow 0$ for $n \rightarrow \infty$. According to Corollary 1

$$\begin{aligned} D(\bar{P}_n \parallel P^*) &= D\left(\frac{1}{n}P + \frac{n-1}{n}\Phi\bar{P}_{n-1} \parallel P^*\right) \\ &\geq \frac{1}{n}D(P \parallel P^*) + \frac{n-1}{n}D(\Phi\bar{P}_{n-1} \parallel P^*) - H\left(\frac{1}{n}, \frac{n-1}{n}\right). \end{aligned}$$

This inequality implies that

$$\limsup_{n \rightarrow \infty} D(\Phi\bar{P}_{n-1} \parallel P^*) \leq \limsup_{n \rightarrow \infty} D(\bar{P}_n \parallel P^*) = 0.$$

Therefore $\Phi\bar{P}_n$ converges to both P^* and ΦP^* and we have $P^* = \Phi P^*$.

2 \Rightarrow 1 :

First observe that $D(\bar{P}_n \parallel Q)$ is finite if and only if $D(P \parallel Q)$ is finite. Further K is invariant.

According to Theorem 1

$$D(\bar{P}_n \parallel Q) \geq D(\bar{P}_n \parallel \Pi_{K \leftarrow Q}) + D(K \parallel Q),$$

and we just have to show that $D(\bar{P}_n \parallel Q) \rightarrow D(K \parallel Q)$ for $n \rightarrow \infty$.

Using the convexity stated in Proposition 1 we get

$$\begin{aligned} (m+n)D(\bar{P}_{m+n} \parallel Q) &= (m+n)D\left(\frac{m}{m+n}\bar{P}_m + \frac{n}{m+n}\Phi^m\bar{P}_n \parallel Q\right) \\ &\leq mD(\bar{P}_m \parallel Q) + nD(\Phi^m\bar{P}_n \parallel Q) \\ &\leq mD(\bar{P}_m \parallel Q) + nD(\bar{P}_n \parallel Q), \end{aligned}$$

so the sequence $nD(\bar{P}_n \parallel Q)$ is sub-additive, and by $D(\bar{P}_n \parallel Q) < \infty$ we have that $D(\bar{P}_n \parallel Q)$ converges.

Let $R = \sum_{i=0}^{n-1} s_i \cdot \Phi^i P$ be an element in K , where $s_i \geq 0$, $\sum s_i = 1$ If $m \geq n-1$ we have

$$\begin{aligned} \frac{1}{m} \sum_{j=0}^{m-1} \Phi^j R &= \frac{1}{m} \sum_{j=0}^{m-1} \Phi^j \left(\sum_{i=0}^{n-1} s_i \cdot \Phi^i P \right) \\ &= \frac{1}{m} \sum_{j=0}^{m-1} \sum_{i=0}^{n-1} s_i \cdot \Phi^{i+j} P \\ &= t\Phi^{n-1}\bar{P}_{m-n+1} + (1-t)S \end{aligned}$$

for $t = \frac{m-n+1}{m}$ and some $S \in K$. Therefore

$$\begin{aligned} D(R \parallel Q) &\geq D(t\Phi^{n-1}\bar{P}_{m-n+1} + (1-t)S \parallel Q) \\ &\geq tD(\Phi^{n-1}\bar{P}_{m-n+1} \parallel Q) + (1-t)D(S \parallel Q) - H(t, 1-t), \end{aligned}$$

and hence

$$D(R \parallel Q) \geq \limsup_{n \rightarrow \infty} D(\Phi^{n-1}\bar{P}_{m-n+1} \parallel Q)$$

and $D(K \parallel Q) = \inf_{i,j} D(\Phi^i \bar{P}_j \parallel Q)$. Further we have

$$\begin{aligned} D(\bar{P}_n \parallel Q) &= D\left(\frac{i}{n}\bar{P}_i + \frac{n-i}{n}\Phi^i \bar{P}_n \parallel Q\right) \\ &\leq \frac{i}{n}D(\bar{P}_i \parallel Q) + \frac{n-i}{n}D(\Phi^i \bar{P}_n \parallel Q), \end{aligned}$$

and therefore $\liminf_{n \rightarrow \infty} D(\bar{P}_n \parallel Q) \leq \liminf_{j \rightarrow \infty} D(\Phi^i \bar{P}_j \parallel Q)$. This shows that $D(K \parallel Q) = \inf D(\bar{P}_n \parallel Q)$.

The inequality

$$D(P \parallel Q) \geq D(P \parallel \Pi_{K \leftarrow Q}) + D(K \parallel Q)$$

holds for all $Q \in L$. By Lemma 4 we have $\Pi_{K \leftarrow Q} \in L$ which shows that $\Pi_{K \leftarrow Q}$ is the reversed I -projection of P . \square

Corollary 2. *Let (Ω, \mathbb{F}, P) be a probability space, and let Φ be a Markov operator. Assume that $\Phi^n P \rightarrow P^*$ for $n \rightarrow \infty$. Then P^* is invariant and equals the I -projection $\Pi_{K \leftarrow Q}$, where K is the set $\text{conv}\{\Phi^i(P)\}$, for any invariant distribution Q .*

Corollary 3. *Let (Ω, \mathbb{F}, Q) be a probability space, and let Φ be a Markov operator, where Q is the only invariant measure under Φ . For any probability measure P define $\bar{P}_n = \frac{1}{n} \sum_{i=0}^{n-1} \Phi^i P$. If $D(P \parallel Q) < \infty$ then \bar{P}_n converges strongly to Q in information.*

The requirement $D(P \parallel Q) < \infty$ in Corollary 3 is rather strong. It implies that P is absolutely continuous with respect to Q . On a topological space it even implies that $\text{supp}(P) \subseteq \text{supp}(Q)$. By invariance of Q the Markov kernel Φ has a restriction to $\text{supp}(Q)$, and $\Phi^i(P)$ is fully determined by this restriction, and in this case no generality is lost by assuming that $\text{supp}(Q) = \Omega$. In this case the Markov chain is ergodic. Let us see how some well-known results from ergodicity theory are corollaries to Theorem 8.

Corollary 4. *A measurable transformation is ergodic if and only if, for all A and B in \mathbb{B} ,*

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} Q(A \cap T^{-k} B) = Q(A)Q(B).$$

Proof. To prove ergodicity of T , assume that A is invariant, and put $B = A$. Then $A \cap T^{-k} B = B$ and therefore $Q(B) = Q(B)^2$, i.e. $Q(A) = 0$ or 1.

Conversely, let T be ergodic. If $Q(A) = 0$ then the theorem is trivial. Otherwise apply Theorem 8 to the distribution $P = Q(\cdot | A)$ and we

$$\begin{aligned} \bar{P}_n(B) &= \frac{1}{n} \sum_{k=0}^{n-1} Q(T^{-k} B | A) \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \frac{Q(A \cap T^{-k} B)}{Q(A)} \rightarrow Q(B) \text{ for } n \rightarrow \infty. \end{aligned}$$

\square

Corollary 5. *Let (Ω, \mathbb{F}, Q) be a probability space with an ergodic transformation T . Then for all sets $A \in \mathbb{F}$ we have*

$$E \left| \frac{1}{n} \sum_{i=0}^{n-1} 1_A(T^i(\omega)) - Q(A) \right| \rightarrow 0 \text{ for } n \rightarrow \infty.$$

Proof. First assume that $Q(A) = 0$. Then for all $i \in \mathbb{N}$ we have

$$Q \{ \omega \in \Omega \mid T^i(\omega) \in A \} = 0,$$

and therefore $Q \{ \omega \in \Omega \mid \exists i < n : T^i(\omega) \in A \} = 0$, and the theorem follows.

Assume that $Q(A) > 0$. The transformation is measure preserving and therefore $Q(T^{-i}A) = Q(A)$ is constant. Define $P = Q(\cdot \mid A)$. Then

$$T^i(P) = T^i(Q(\cdot \mid A)) = Q(\cdot \mid T^{-i}A).$$

Furthermore

$$D(P \parallel Q) = -\log(Q(A)) < \infty$$

and we can use Theorem 8. We have

$$\bar{P}_n = \frac{1}{n} \sum_{i=0}^{n-1} Q(\cdot \mid T^{-i}A),$$

and

$$\frac{d\bar{P}_n}{dQ}(\omega) = \frac{\frac{1}{n} \sum_{i=0}^{n-1} 1_A(T^i(\omega))}{Q(A)}.$$

Using Pinsker's inequality we have $\|\bar{P}_n - Q\| \rightarrow 0$ for $n \rightarrow \infty$, but

$$\begin{aligned} \|\bar{P}_n - Q\| &= E \left\| \frac{d\bar{P}_n}{dQ} - 1 \right\| \\ &= \frac{1}{Q(A)} E \left| \frac{1}{n} \sum_{i=0}^{n-1} 1_A(T^i(\omega)) - Q(A) \right|. \end{aligned}$$

□

6. DISCUSSION

In this paper convergence of Markov chains in information has been discussed. A crucial condition turned out to be that $D(\Phi^n P \parallel Q)$ is finite eventually. This excludes many examples. In order to cover more examples one could use Jensen-Shannon divergence

$$JSD(P, Q) = \frac{1}{2} D \left(P \parallel \frac{P+Q}{2} \right) + \frac{1}{2} D \left(Q \parallel \frac{P+Q}{2} \right)$$

in stead of information divergence. The Jensen-Shannon divergence satisfies the inequalities

$$\frac{1}{8} \|P - Q\|^2 \leq JSD(P, Q) \leq \frac{\log 2}{2} \|P - Q\|,$$

and in particular the Jensen-Shannon divergence is automatically bounded. Thus $JSD(P_n, Q) \rightarrow 0$ for $n \rightarrow \infty$ implies that $P_n \rightarrow Q$ in total variation. Formulation of results involving Jensen-Shannon divergence becomes less elegant, but by this method one can cover all cases of convergence treated by A. Rényi [Rén61], I. Csiszár [Csi63] and Kendall [Ken64].

Several of the results in this paper can easily be extended to a non-commutative setup. Then (Ω, \mathbb{F}) is replaced by a Von Neumann algebra \mathcal{A} , Φ is replaced by a completely positive map, and P, Q are replaced by normal states on \mathcal{A} . See Ohya and Petz [OP93] for definition and discussion of *relative entropy in operator algebras*.

Acknowledgement The authors want to thank Andrew Barron for useful discussions.

REFERENCES

- [ABBN04] S. Artstein, K. Ball, F. Barthe, and A. Naor. Solution of Shannon’s problem on the monotonicity of entropy. *J. Amer. Math. Soc.*, 17:975–982, 2004.
- [Bar86] A. R. Barron. Entropy and the Central Limit Theorem. *Annals Probab. Theory*, 14(1):336 – 342, 1986.
- [Bar00] A. R. Barron. Limits of information, Markov chains, and projections. In *Proceedings 2000 International Symposium on Information Theory*, page 25, 2000.
- [BE85] D. Bakry and M. Emery. *Diffusions hypercontractives*, volume 1123 of *Lect. Notes in Math.*, pages 179–206. Springer, 1985.
- [CM03] I. Csiszár and F. Matús. Information projections revisited. *IEEE Trans. Inform. Theory*, 49(6):1474–1490, June 2003.
- [Csi63] I. Csiszár. Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *Publ. Math. Inst. Hungar. Acad.*, 8:95–108, 1963.
- [Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.
- [Csi75] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.
- [Csi84] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12:768–793, 1984.
- [Dob59] R. L. Dobrusin. General formulation of Shannon’s main theorem in information theory. *Usp. Mat. Nauk.*, 14:3–104, 1959. In Russian. English translation in *Transl. A.M.S.*, Ser. 2, 33, pp. 323–438, 1963.
- [Dob60] R. L. Dobrusin. Passage to the limit under the information and entropy integrals. *Theory Probab. Appl.*, 5:25–32, 1960. English translation.
- [Fri73] J. Fritz. An information-theoretical proof of limit theorems for reversible Markov processes. In *Trans. Sixth Prague Conf. on Inform. Theory, Statist. Decision Functions, Random Processes*, Prague, 1973. Czech. Acad. Science, Academia Publ. Prague, Sept. 1971.
- [Har01] P. Harremoës. Binomial and Poisson distributions as maximum entropy distributions. *IEEE Trans. Inform. Theory*, 47(5):2039–2041, July 2001.
- [Har05] P. Harremoës. Martingales and information divergence. In *Proceedings of 2005 IEEE International Symposium on Information Theory*, pages 164–168, Adelaide, Australia, Sept. 2005. IEEE.
- [Har06] P. Harremoës. Maximum entropy on compact groups. In *Proceedings of International Symposium on Information Theory*, pages 108–112, Seattle, USA, 9-14. July 2006. IEEE.
- [Har07] P. Harremoës. Information topologies with applications. In I. Csiszár, Gyula O. H. Katona, and Gábot Tardos, editors, *Entropy, Search, Complexity*, volume 16 of *Bolyai Society Mathematical Studies*, pages 113–150. János Bolyai Mathematical Society and Springer-Verlag, 2007.
- [Ken64] D. G. Kendall. Information theory and the limit theorem for Markov chains and processes with a countable infinity of states. *Ann. Inst. Stat. Math.*, 15:137–143, 1964.
- [OP93] M. Ohya and D. Petz. *Quantum Entropy and Its Use*. Springer, Berlin Heidelberg New York, 1993.
- [Pin60] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Izv. Akad. Nauk, Moskva, 1960. in Russian.
- [Rén61] A. Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Symp. Math. Statist. and Prob.*, volume 1, pages 547–561, Berkely, 1961. Univ. Calif. Press.

- [Tak87] S. Takano. Convergence of entropy in the central limit theorem. *Yokohama Mathematical Journal*, 35:143–148, 1987.
- [Top74] F. Topsøe. *Informationstheorie, eine Einführung*. Teubner, Stuttgart, 1974.
- [Top79] F. Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8 – 27, 1979.

(P. Harremoës) QUANTUM COMPUTING AND ADVANCED SYSTEMS RESEARCH, CENTRE FOR MATHEMATICS AND COMPUTER SCIENCE (CWI), AMSKRDAM

E-mail address: P.Harremoes@cwi.nl

URL: www.cwi.nl/~ph

(K. K. Holst) DEPARTMENT OF BIOSTATISTICS, UNIVERSITY OF COPENHAGEN

E-mail address: kkho@biostat.ku.dk