Entropy **2009**, *11*, 222-237; doi:10.3390/e11020222

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Maximum Entropy on Compact Groups

Peter Harremoës

Centrum Wiskunde & Informatica, Science Park 123, 1098 GB Amsterdam, Noord-Holland, The Netherlands. E-mail: P.Harremoes@cwi.nl

Received: 30 December 2008 / Accepted: 31 March 2009 / Published: 1 April 2009

Abstract: In a compact group the Haar probability measure plays the role of uniform distribution. The entropy and rate distortion theory for this uniform distribution is studied. New results and simplified proofs on convergence of convolutions on compact groups are presented and they can be formulated as entropy increases to its maximum. Information theoretic techniques and Markov chains play a crucial role. The convergence results are also formulated via rate distortion functions. The rate of convergence is shown to be exponential.

Keywords: Compact group; Convolution; Haar measure; Information divergence; Maximum entropy; Rate distortion function; Rate of convergence; Symmetry.

Classification: MSC 94A34,60B15

1. Introduction

It is a well-known and celebrated result that the uniform distribution on a finite set can be characterized as having maximal entropy. Jaynes used this idea as a foundation of statistical mechanics [1], and the Maximum Entropy Principle has become a popular principle for statistical inference [2–8]. Often it is used as a method to get prior distributions. On a finite set, for any distributions P we have $H(P) = H(U) - D(P\|U)$ where H is the Shannon entropy, D is information divergence, and U is the uniform distribution. Thus, maximizing $H(P)$ is equivalent to minimizing $D(P\|U)$. Minimization of information divergence can be justified by the conditional limit theorem by Csiszár [9, Theorem 4]. So if we have a good reason to use the uniform distribution as prior distribution we automatically get a justification of the Maximum Entropy Principle. The conditional limit theorem cannot justify the use of the uniform distribution itself, so we need something else. Here we shall focus on symmetry.

Example 1. *A die has six sides that can be permuted via rotations of the die. We note that not all*

permutations can be realized as rotations and not all rotations will give permutations. Let G be the group of permutations that can be realized as rotations. We shall consider G as the symmetry group of the die and observe that the uniform distribution on the six sides is the only distribution that is invariant under the action of the symmetry group G.

Example 2. *$G = \mathbb{R}/2\pi\mathbb{Z}$ is a commutative group that can be identified with the group $SO(2)$ of rotations in 2 dimensions. This is the simplest example of a group that is compact but not finite.*

For an object with symmetries the symmetry group defines a group action on the object, and any group action on an object defines a symmetry group of the object. A special case of a group action of the group G is left translation of the elements in G. Instead of studying distributions on objects with symmetries, in this paper we shall focus on distributions on the symmetry groups themselves. It is no serious restriction because a distribution on the symmetry group of an object will induce a distribution on the object itself.

Convergence of convolutions of probability measures were studied by Stromberg [10] who proved weak convergence of convolutions of probability measures. An information theoretic approach was introduced by Csiszár [11]. Classical methods involving characteristic functions have been used to give conditions for uniform convergence of the densities of convolutions [12]. See [13] for a review of the subject and further references.

Finally it is shown that convergence in information divergence corresponds to uniform convergence of the rate distortion function and that weak convergence corresponds to pointwise convergence of the rate distortion function. In this paper we shall mainly consider convolutions as Markov chains. This will give us a tool, which allows us to prove convergence of iid. convolutions, and the rate of convergence is proved to be exponential.

The rest of the paper is organized as follows. In Section 2. we establish a number of simple results on distortion functions on compact set. These results will be used in Section 4.. In Section 3. we define the uniform distribution on a compact group as the uniquely determined Haar probability measures. In Section 4. it is shown that the uniform distribution is the maximum entropy distribution on a compact group in the sense that it maximizes the rate distortion function at any positive distortion level. Convergence of convolutions of a distribution to the uniform distribution is established in Section 5. using Markov chain techniques, and the rate of convergence is discussed in Section 6.. The group $SO(2)$ is used as our running example. We finish with a short discussion.

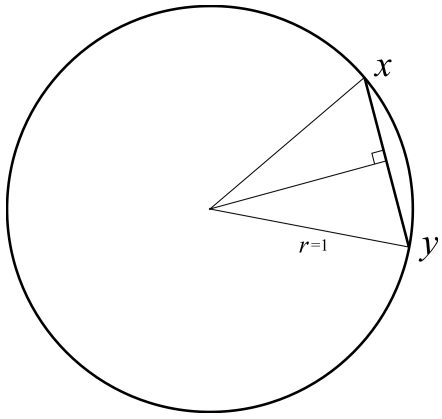
2. Distortion on compact groups

Let G be a compact group where $*$ denotes the composition. The neutral element will be denoted e and the inverse of the element g will be denoted g^{-1}.

We shall start with some general comments on distortion functions on compact sets. Assume that the group both plays the role as source alphabet and reproduction alphabet. A *distortion function* $d : G \times G \to \mathbb{R}$ is given and we will assume that $d(x, y) \geq 0$ with equality if and only if $x = y$. We will also assume that the distortion function is continuous.

Example 3. *As distortion function on $SO(2)$ we use the squared Euclidean distance between the corre-*

Figure 1. Squared Euclidean distance between the rotation angles x and y.



sponding points on the unit circle, i.e.

$$\begin{aligned} d(x,y) &= 4\sin^2\left(\frac{x-y}{2}\right) \\ &= 2-2\cos(x-y). \end{aligned}$$

This is illustrated in Figure 1.

The distortion function might be a metric but even if the distortion function is not a metric, the relation between the distortion function and the topology is the same as if it was a metric. One way of constructing a distortion function on a group is to use the squared Hilbert-Smidt norm in a unitary representation of the group.

Theorem 4. *If C is a compact set and $d : C \times C \to \mathbb{R}$ is a non-negative continuous distortion function such that $d(x,y) = 0$ if and only if $x = y$, then the topology on C is generated by the distortion balls $\{x \in C \mid d(x,y) < r\}$ where $y \in C$ and $r > 0$.*

Proof. We have to prove that a subset $B \subseteq C$ is open if and only if for any $y \in B$ there exists a ball that is a subset of B and contains y. Assume that $B \subset C$ is open and that $y \in B$. Then $\complement B$ compact. Hence, the function $x \to d(x,y)$ has a minimum r on $\complement B$ and r must be positive because $r = d(x,y) = 0$ would imply that $x = y \in B$. Therefore $\{x \in C \mid d(x,y) < r\} \subseteq B$.

Continuity of d implies that the balls $\{x \in C \mid d(x,y) < r\}$ are open. If any point in B is contained in an open ball, then B is a union of open set and open. □

The following theorem may be considered as a kind of uniform continuity of the distortion function or as a substitute for the triangular inequality when d is not a metric.

Lemma 5. *If C is a compact set and $d : C \times C \to \mathbb{R}$ is a non-negative continuous distortion function such that $d(x,y) = 0$ if and only if $x = y$, then there exists a continuous function f_1 satisfying $f_1(0) = 0$ such that*

$$|d(x,y) - d(z,y)| \leq f_1(d(z,y)) \text{ for } x,y,z \in C. \tag{1}$$

Proof. Assume that the theorem does not hold. Then there exists $\epsilon > 0$ and a net $(x_\lambda, y_\lambda, z_\lambda)_{\lambda\in\Lambda}$ such that

$$d(x_\lambda, y_\lambda) - d(z_\lambda, y_\lambda) > \epsilon$$

and $d(z_\lambda, y_\lambda) \to 0$. A net in a compact set has a convergent subnet so without loss of generality we may assume that the net $(x_\lambda, y_\lambda, z_\lambda)_{\lambda\in\Lambda}$ converges to some triple $(x_\infty, y_\infty, z_\infty)$. By continuity of the distortion function we get

$$d(x_\infty, y_\infty) - d(z_\infty, y_\infty) \geq \epsilon$$

and $d(z_\infty, y_\infty) = 0$, which implies $z_\infty = y_\infty$ and we have a contradiction. □

We note that if a distortion function satisfies (1) then it defines a topology in which the distortion balls are open.

In order to define the weak topology on probability distributions we extend the distortion function from $C \times C$ to $M_+^1(C) \times M_+^1(C)$ via

$$d(P, Q) = \inf E[\, d(X, Y)],$$

where X and Y are random variables with values in C and the infimum is taken all joint distributions on (X, Y) such that the marginal distribution of X is P and the marginal distribution of Y is Q. The distortion function is continuous so $(x, y) \to d(x, y)$ has a maximum that we denote $d_{\max}$.

Theorem 6. *If G is a compact set and $d : C \times C \to \mathbb{R}$ is a non-negative continuous distortion function such that $d(x, y) = 0$ if and only if $x = y$, then*

$$|d(P, Q) - d(S, Q)| \leq f_2(d(S, P)) \text{ for } P, Q, S \in M_+^1(C)$$

for some continuous function f_2 satisfying $f_2(0) = 0$.

Proof. According to Lemma 5 there exists a function f_1 satisfying (1). We use that

$$\begin{aligned} E[|d(X, Y) - d(Z, Y)|] &\leq E[f_1(d(Z, X))] \\ &= E[f_1(d(Z, X)) \mid d(Z, X) \leq \delta] \cdot P(d(Z, X) \leq \delta) \\ &+ E[f_1(d(Z, X)) \mid d(Z, X) > \delta] \cdot P(d(Z, X) > \delta) \\ &\leq f_1(\delta) \cdot 1 + f_1(d_{\max}) \cdot \frac{E[d(Z, X)]}{\delta} \\ &\leq f_1(\delta) + f_1(d_{\max}) \cdot \frac{d(S, P)}{\delta}. \end{aligned}$$

This hold for all $\delta > 0$ and in particular for $\delta = (d(S, P))^{1/2}$, which proves the theorem. □

The theorem can be used to construct the *weak topology* on $M_+^1(C)$ with

$$\left\{P \in M_+^1(C) \mid d(P, Q) < r\right\},$$

$P \in M_+^1(C), r > 0$ as open balls that generate the topology. We note without proof that this definition is equivalent with the quite different definition of weak topology that one will find in most textbooks.

For a group G we assume that the distortion function is *right invariant* in the sense that for all $x, y, z \in G$ a distortion function d satisfies

$$d(x * z, y * z) = d(x, y).$$

A right invariant distortion function satisfies $d(x, y) = d(x * y^{-1}, e)$, so right invariant continuous distortion functions of a group can be constructed from non-negative functions with a minimum in e.

3. The Haar measure

We use $*$ to denote convolution of probability measures on G. For $g \in G$ we shall use $g * P$ to denote the g-translation of the measure P or, equivalently, the convolution with a measure concentrated in g. The n-fold convolution of a distribution P with itself will be denoted P^{*n}. For random variables with values in G one can formulate an analog of the central limit theorem. We recall some facts about probability measures on compact groups and their *Haar measures*.

Definition 7. *Let G be a group. A measure P is said to be a* left Haar measure *if $g * P = P$ for any $g \in G$. Similarly, P is said to be a* right Haar measure *if $P * g = P$ for any $g \in G$. A measure is said to be a* Haar measure *if it is both a left Haar measure and a right Haar measure.*

Example 8. *The uniform distribution on $SO(2)$ or $\mathbb{R}/2\pi Z$ has density $1/2\pi$ with respect to the Lebesgue measure on $[0; 2\pi[$. The function*

$$f(x) = 1 + \sum_{n=1}^{\infty} a_n \cos(n(x + \phi_n)) \tag{2}$$

is a density on a probability distribution P on $SO(2)$ if the Fourier coefficients a_n are sufficiently small so that f is non-negative. A sufficient condition for f to be non-negative is that $\sum_{n=1}^{\infty} |a_n| \leq 1$.

Translation by y gives a distribution with density

$$f(x - y) = 1 + \sum_{n=1}^{\infty} a_n \cos(n(x - y + \phi_n)).$$

The distribution P is invariant if and only if f is 1 or, equivalently, all Fourier coefficients $(a_n)_{n \in \mathbb{N}}$ are 0.

A measure P on G is said to have *full support* if the support of P is G, i.e. $P(A) > 0$ for any non-empty open set $A \subseteq G$. The following theorem is well-known [14–16].

Theorem 9. *Let U be a probability measure on the compact group G. Then the following four conditions are equivalent.*

- *U is a left Haar measure.*
- *U is a right Haar measure.*
- *U has full support and is idempotent in the sense that $U * U = U$.*
- *There exists a probability measure P on G with full support such that $P * U = U$.*

- *There exists a probability measure P on G with full support such that $U * P = U$.*

In particular a Haar probability measure is unique.

In [14–16] one can find the proof that any locally compact group has a Haar measure. The unique Haar probability measure on a compact group will be called the *uniform distribution* and denoted U. For probability measures P and Q the *information divergence from P to Q* is defined by

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

We shall often calculate the divergence from a distribution to the uniform distribution U, and introduce the notation

$$D(P) = D(P\|U).$$

For a random variable X with values in G we will sometimes write $D(X\|U)$ instead of $D(P\|U)$ when X has distribution P.

Example 10. *The distribution P with density f given by (2) has*

$$\begin{aligned} D(P) &= \frac{1}{2\pi}\int_0^{2\pi} f(x)\log(f(x))\, dx \\ &\approx \frac{1}{2\pi}\int_0^{2\pi} f(x)(f(x)-1)\, dx \\ &= \frac{1}{2}\sum_{n=1}^{\infty} a_n^2. \end{aligned}$$

Let G be a compact group with uniform distribution U and let F be a closed subgroup of G. Then the subgroup has a Haar probability measure U_F and

$$D(U_F) = \log([G:F]) \tag{3}$$

where $[G:F]$ denotes the index of F in G. In particular $D(U_F)$ is finite if and only if $[G:F]$ is finite.

4. The rate distortion theory

We will develop aspects of the rate distortion theory of a compact group G. Let P be a probability measure on G. We observe that compactness of G implies that a covering of G by distortion balls of radius $\delta > 0$ contains a finite covering. If k is the number of balls in a finite covering then $R_P(\delta) \leq \log(k)$ where R_P is the rate distortion function of the probability measure P. In particular the rate distortion function is upper bounded. The entropy of a probability distribution P is given by $H(P) = R_P(0)$. If the group is finite then the uniform distribution maximizes the Shannon entropy $R_P(0)$ but if the group is not finite then in principle there is no entropy maximizer. As we shall see the uniform distribution still plays the role of entropy maximizer in the sense that the uniform distribution maximize the value $R_P(\delta)$ of the rate distortion function for any positive distortion level $\delta > 0$. The rate distortion function R_P can be studied using its convex conjugate R_P^* given by

$$R_P^*(\beta) = \sup_{\delta} \beta\cdot\delta - R_P(\delta).$$

The rate distortion function is then recovered by the formula

$$R_P(\delta) = \sup_\beta \beta \cdot \delta - R_P^*(\beta).$$

The techniques are pretty standard [17].

Theorem 11. *The rate distortion function of the uniform distribution is given by*

$$R_U^*(\beta) = \log(Z(\beta))$$

where Z is the partition function defined by

$$Z(\beta) = \int_G \exp(\beta \cdot d(g,e))\ dUg.$$

The rate distortion function of an arbitrary distribution P satisfies

$$R_U - D(P\|U) \leq R_P \leq R_U. \tag{4}$$

Proof. First we prove a Shannon type lower bound on the rate distortion function of an arbitrary distribution P on the group. Let X be a random variable with values in G and distribution P, and let $\hat{X}$ be a random variable coupled with X such that the mean distortion $E\left[d\left(X,\hat{X}\right)\right]$ equals δ. Then

$$I\left(X;\hat{X}\right) = D\left(X\|U \mid \hat{X}\right) - D(X\|U) \tag{5}$$

$$= D\left(X * \hat{X}^{-1}\|U \mid \hat{X}\right) - D(X\|U) \tag{6}$$

$$\geq D\left(X * \hat{X}^{-1}\|U\right) - D(X\|U). \tag{7}$$

Now, $E\left[d\left(X,\hat{X}\right)\right] = E\left[d\left(X * \hat{X}^{-1}, e\right)\right]$ and

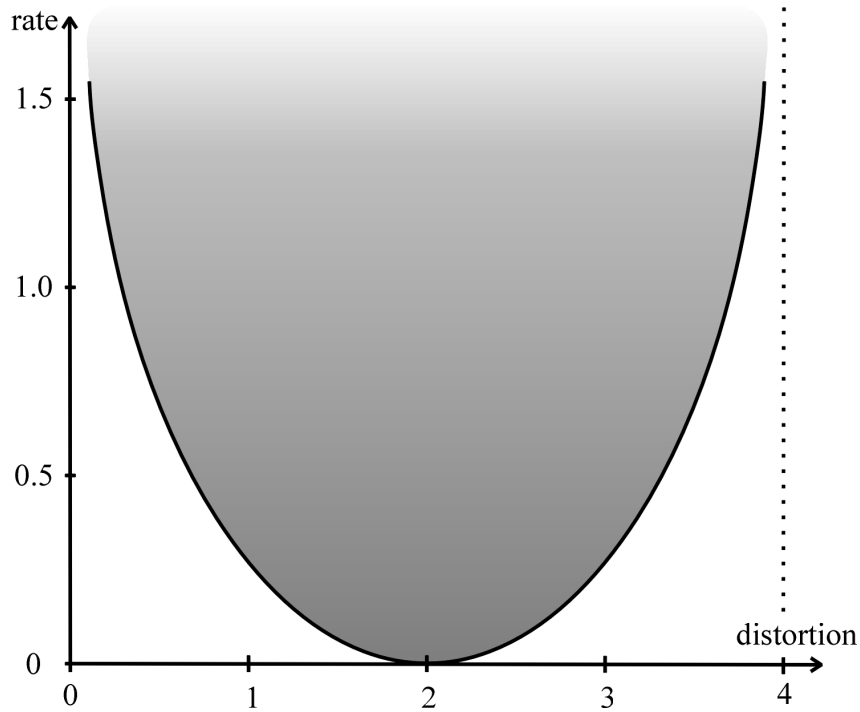
$$D\left(X * \hat{X}^{-1}\|U\right) \geq D(P_\beta\|U)$$

where P_β is the distribution that minimizes divergence under the constraint $E[d(Y,e)] = \delta$ when Y has distribution P_β. The distribution P_β is given by the density

$$\frac{dP_\beta}{dU}(g) = \frac{\exp(\beta \cdot d(g,e))}{Z(\beta)}.$$

where β is determined by the condition $\delta = Z'(\beta)/Z(\beta)$.
If P is uniform then a joint distribution is obtained by choosing $\hat{X}$ uniformly distributed, and choosing Y distributed according to P_β and independent of $\hat{X}$. Then $X = Y * \hat{X}$ is distributed according to $P_\beta * U = U$, and we have equality in (7). Hence the lower bound (7) is achievable for the uniform distribution, which prove the first part of the theorem, and the left inequality in (4).
The joint distribution on $\left(X,\hat{X}\right)$ that achieved the rate distortion function when X has a uniform distribution, defines a Markov kernel $\Psi : X \to \hat{X}$ that is invariant under translations in the group. For any distribution P the joint distribution on $\left(X,\hat{X}\right)$ determined by P and Ψ gives an achievable pair of distortion, and rate that is on the rate distortion curve of the uniform distribution. This proves the right inequality in Equation (4). □

Figure 2. The rate distortion region of the uniform distribution on $SO(2)$ is shaded. The rate distortion function is the lower bounding curve. In the figure the rate is measured in nats. The critical distortion d_{crit} equals 2, and the dashed line indicates $d_{\max} = 4$.



Example 12. *For the group $SO(2)$ the rate distortion function can be parametrized using the modified Bessel functions $I_j, j \in \mathbb{N}_0$. The partition function is given by*

$$\begin{aligned} Z(\beta) &= \int_G \exp(\beta \cdot d(g,e))\ dUg \\ &= \frac{1}{2\pi}\int_0^{2\pi} \exp(\beta \cdot (2 - 2\cos x))\ dx \\ &= \exp(2\beta) \cdot \frac{1}{\pi}\int_0^{\pi} \exp(-2\beta \cdot \cos x)\ dx \\ &= \exp(2\beta) \cdot I_0(-2\beta). \end{aligned}$$

Hence $R_U^*(\beta) = \log(Z(\beta)) = 2\beta + \log(I_0(-2\beta))$. The distortion δ corresponding to β is given by

$$\delta = 2 - 2\frac{I_1(-2\beta)}{I_0(-2\beta)}$$

and the corresponding rate is

$$\begin{aligned} R_U(\delta) &= \beta \cdot \delta - (2\beta + \log(I_0(-2\beta))) \\ &= -\beta \cdot 2\frac{I_1(-2\beta)}{I_0(-2\beta)} - \log(I_0(-2\beta)). \end{aligned}$$

These joint values of distortion and rate can be plotted with β as parameter as illustrated in Figure 2.

The minimal rate of the uniform distribution is achieved when X and $\hat{X}$ are independent. In this case the distortion is $E\left[d\left(X,\hat{X}\right)\right] = \int_G d(x,e)\ dPx$. This distortion level will be called the critical distortion and will be denoted d_{crit}. On the interval $]0;d_{crit}]$ the rate distortion function is decreasing and the distortion rate function is the inverse R_P^{-1} of the rate distortion function R_P on this interval. The distortion rate function satisfies:

Theorem 13. *The distortion rate function of an arbitrary distribution P satisfies*

$$R_U^{-1}(\delta) - f_2(d(P,U)) \leq R_P^{-1}(\delta) \leq R_U^{-1}(\delta) \text{ for } \delta \leq d_{crit} \tag{8}$$

for some increasing continuous function f_2 satisfying $f_2(0) = 0$.

Proof. The right hand side follows because R_U is decreasing in the interval $[0;d_{crit}]$. Let X be a random variable with distribution P and let Y be a random variable coupled with X. Let Z be a random variable coupled with X such that $E[d(X,Z)] = d(P,U)$. The couplings between X and Y, and between X and Z can be extended to a joint distribution on X, Y and Z such that Y and Z are independent given X. For this joint distribution we have

$$I(Z;Y) \leq I(X,Y)$$

and

$$|E[d(Z,Y)] - E[d(X,Y)]| \leq f_2(d(P,U)).$$

We have to prove that

$$E[d(X,Y)] \geq R_U^{-1}(I(X,Y)) - f_2(d(P,U))$$

but $I(Z;Y) \leq I(X,Y)$ so it is sufficient to prove that

$$E[d(X,Y)] \geq R_U^{-1}(I(Z,Y)) - f_2(d(P,U))$$

and this follows because $E[d(Z,Y)] \geq R_U^{-1}(I(Z,Y))$. □

5. Convergence of convolutions

We shall prove that under certain conditions the n-fold convolutions P^{*n} converge to the uniform distribution.

Example 14. *The function*

$$f(x) = 1 + \sum_{n=1}^{\infty} a_n \cos(n(x+\phi_n))$$

is a density on a probability distribution P on G if the Fourier coefficients a_n are sufficiently small. If

(a_n) and (b_n) are Fourier coefficients of P and Q then the convolution has density

$$\begin{aligned}
\frac{1}{2\pi}\int_0^{2\pi}\left(1+\sum_{n=1}^{\infty}a_n\cos n\left(x-y+\phi_n\right)\right)\left(1+\sum_{n=1}^{\infty}b_n\cos n\left(y+\psi_n\right)\right)\,dy \\
= 1+\frac{1}{2\pi}\sum_{n=1}^{\infty}\int_0^{2\pi}a_nb_n\cos n\left(x-y+\phi_n\right)\cos n\left(y+\psi_n\right)\,dy \\
= 1+\frac{1}{2\pi}\sum_{n=1}^{\infty}\int_0^{2\pi}a_nb_n\cos\left(n\left(x+\phi_n+\psi_n\right)-ny\right)\cos\left(ny\right)\,dy \\
= 1+\frac{1}{2\pi}\sum_{n=1}^{\infty}\int_0^{2\pi}a_nb_n\left(\begin{array}{c}\cos n\left(x+\phi_n+\psi_n\right)\cos\left(ny\right)\\ +\sin\left(n\left(x+\phi_n+\psi_n\right)\right)\sin\left(ny\right)\end{array}\right)\cos\left(ny\right)\,dy \\
= 1+\sum_{n=1}^{\infty}\frac{a_nb_n\cos\left(n\left(x+\phi_n+\psi_n\right)\right)}{2\pi}\int_0^{2\pi}\cos^2\left(ny\right)\,dy \\
= 1+\sum_{n=1}^{\infty}\frac{a_nb_n\cos\left(n\left(x+\phi_n+\psi_n\right)\right)}{2}.
\end{aligned}$$

Therefore the n-fold convolution has density

$$1+\sum_{k=1}^{\infty}\frac{a_k^n\cos\left(k\left(x+n\phi_k\right)\right)}{2^{n-1}}=1+\sum_{k=1}^{\infty}\left(\frac{a_k}{2}\right)^n 2\cos\left(k\left(x+n\phi_k\right)\right).$$

Therefore each of the Fourier coefficients is exponentially decreasing.

Clearly, if P is uniform on a proper subgroup then convergence to the uniform distribution on the whole group does not hold. In several papers on this topic [13, 18, and references therein] it is claimed and "proved" that if convergence does not hold then the support of P is contained in the coset of a proper normal subgroup. The proofs therefore contain errors that seem to have been copied from paper to paper. To avoid this problem and make this paper more self-contained we shall reformulate and reprove some already known theorems.

In the theory of finite Markov chains is well-known that there exists an invariant probability measure. Certain Markov chains exhibits periodic behavior where a certain distribution is repeated after a number of transitions. All distributions in such a cycle will lie at a fixed distance from any (fixed) measure, where the distance is given by information divergence or total variation (or any other Csiszár f-divergence). It is also well-known that finite Markov chains without periodic behavior are convergent. In general a Markov chain will converge to a "cyclic" behavior as stated in the following theorem [19].

Theorem 15. *Let Φ be a transition operator on a state space A with an invariant probability measure Q_{in}. If $D\left(S\parallel Q\right)<\infty$ then there exists a probability measure P^* such that $D\left(\Phi^nS\parallel\Phi^nQ\right)\to 0$ and $D\left(\Phi^nQ\parallel Q_{in}\right)$ is constant.*

We shall also use the following proposition that has a purely computational proof [20].

Proposition 16. *Let $P_x, x\in X$ be distributions and let Q be a probability distribution on X. Then*

$$\int D\left(P_x\parallel Q\right)\,dQx=D\left(\int P_xdQx\parallel Q\right)+\int D\left(P_x\parallel\int P_t\,dQt\right)\,dQx.$$

We denote the set of probability measures on G by $M_+^1(G)$.

Theorem 17. *Let P be a distribution on a compact group G and assume that the support of P is not contained in any nontrivial coset of a subgroup of G. If $D(S\|U)$ is finite then $D(P^{*n} * S\|U) \to 0$ for $n \to \infty$.*

Proof. Let $\Psi : G \to M_+^1(G)$ denote the Markov kernel $\Psi(g) = P * g$. Then $P^{*n} * S = \Psi^n(P * S)$. Thus there exists a probability measure Q on G such that $D(\Psi^n(P)\|\Psi^n(Q)) \to 0$ for $n \to \infty$ and such that $D(\Psi^n(Q))$ is constant. We shall prove that $Q = U$.

First we note that

$$\begin{aligned} D(Q) &= D(P * Q) \\ &= \int_G (D(g * Q) - D(g * Q\|P * Q))\ dPg \\ &= D(Q) - \int_G D(g * Q\|P * Q)\ dPg\ . \end{aligned}$$

Therefore $g * Q = P * Q$ for P almost every $g \in G$. Thus there exists at least one $g_0 \in G$ such that $g_0 * Q = P * Q$. Then $Q = \tilde{P} * Q$ where $\tilde{P} = g_0^{-1} * P$.
Let $\tilde{\Psi} : G \to M_+^1(G)$ denote the Markov kernel $g \to \tilde{P} * g$. Put

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{P}^{*i} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\Psi}^{i-1}\left(\tilde{P}\right).$$

According to [19] this ergodic mean will converge to a distribution T such that $\tilde{\Psi}(T) = T$ so that $T * \tilde{P} = T$. Hence we also have that $T * T = T$, i.e. T is idempotent and therefore supported by a subgroup of G. We know that $\tilde{P}$ is not contained in any nontrivial subgroup of G so the support of T must be G. We also get $Q = T * Q$, which together with Theorem 9 implies that $Q = U$. □

By choosing $S = P$ we get the following corollary.

Corollary 18. *Let P be a probability measure on the compact group G with Haar probability measure U. Assume that the support of P is not contained in any coset of a proper subgroup of G and $D(P\|U)$ is finite. Then $D(P^{*n}\|U) \to 0$ for $n \to \infty$.*

Corollary 18 together with Theorem 11 implies the following result.

Corollary 19. *Let P be a probability measure on the compact group G with Haar probability measure U. Assume that the support of P is not contained in any coset of a proper subgroup of G and $D(P\|U)$ is finite. Then the rate distortion function of P^{*n} converges uniformly to the rate distortion function of the uniform distribution.*

We also get weak versions of these results.

Corollary 20. *Let P be a probability measure on the compact group G with Haar probability measure U. Assume that the support of P is not contained in any coset of a proper subgroup of G. Then P^{*n} converges to U in the weak topology, i.e. $d(P^{*n}, U) \to 0$ for $n \to \infty$.*

Proof. If we take $S = P_\beta$ then $D(P_\beta)$ is finite and $D(P^{*n} * P_\beta \| U) \to 0$ for $n \to \infty$. We have

$$\begin{aligned} d(P^{*n} * P_\beta, U) &\leq d_{\max} \|P^{*n} * P_\beta - U\| \\ &\leq d_{\max} (2D(P^{*n} * P_\beta \| U))^{1/2} \end{aligned}$$

implying that $d(P^{*n} * P_\beta, U) \to 0$ for $n \to \infty$. Now

$$\begin{aligned} |d(P^{*n}, U) - d(P^{*n} * P_\beta, U)| &\leq f_2(d(P^{*n} * P_\beta, P^{*n})) \\ &\leq f_2(d(P_\beta, e)). \end{aligned}$$

Therefore $\lim_{n\to\infty} \sup d(P^{*n}, U) \leq f_2(d(P_\beta, e))$ for all β, which implies that

$$\lim_{n\to\infty} \sup d(P^{*n}, U) = 0.$$

□

Corollary 21. *Let P be a probability measure on the compact group G with Haar probability measure U. Assume that the support of P is not contained in any coset of a proper subgroup of G and $D(P\|U)$ is finite. Then $R_{P^{*n}}$ converges to R_U pointwise on the interval $]0; d_{\max}[$ for $n \to \infty$.*

Proof. Corollary 20 together with Theorem 13 implies uniform convergence of the distortion rate function for distortion less than d_{crit}. This implies pointwise convergence of the rate distortion function on $]0; d_{crit}[$ because rate distortion functions are convex functions. The same argument works in the interval $]d_{crit}; d_{\max}[$. Pointwise convergence in d_{crit} must also hold because of continuity. □

6. Rate of convergence

Normally the rate of convergence will be exponential. If the density is lower bounded this is well-known. We bring a simplified proof of this.

Lemma 22. *Let P be a probability distribution on the compact group G with Haar probability measure U. If $dP/dU \geq c > 0$ and $D(P)$ is finite, then*

$$D(P^n) \leq (1-c)^{n-1} D(P).$$

Proof. First we write

$$P = (1-c) \cdot S + c \cdot U$$

where S denotes the probability measure

$$S = \frac{P - cU}{1-c}.$$

For any distribution Q on G we have

$$\begin{aligned} D(Q * P) &= D((1-c) \cdot Q * S + c \cdot Q * U) \\ &\leq (1-c) \cdot D(Q * S) + c \cdot D(Q * U) \\ &\leq (1-c) \cdot D(Q) + c \cdot D(U) \\ &= (1-c) \cdot D(Q). \end{aligned}$$

Here we have used convexity of divergence. □

If a distribution P has support in a proper subgroup F then

$$\begin{aligned} D(P) &\geq D(U_F) \\ &= \log([G:F]) \\ &\geq \log(2) = 1 \text{ bit.} \end{aligned}$$

Therefore $D(P) < 1$ bit implies that P cannot be supported by a proper subgroup, but it implies more.

Proposition 23. *If P is a distribution on the compact group G and $D(P) < 1$ bit then $\frac{d(P*P)}{dU}$ is lower bounded by a positive constant.*

Proof. The condition $D(P) < 1$ bit implies that $U\left\{\frac{dP}{dU} > 0\right\} > 1/2$. Hence there exists $\varepsilon > 0$ such that $U\left\{\frac{dP}{dU} > \varepsilon\right\} > 1/2$. We have

$$\begin{aligned} \frac{d(P*P)}{dU}(y) &= \int_G \frac{dP}{dU}(x) \cdot \frac{dP}{dU}(y-x)\ dUx \\ &\geq \int_{\left\{\frac{dP}{dU} > \varepsilon\right\}} \varepsilon \cdot \frac{dP}{dU}(y-x)\ dUx \\ &\geq \varepsilon \cdot \int_{\left\{\frac{dP}{dU}(x) > \varepsilon\right\} \cap \left\{\frac{dP}{dU}(y-x) > \varepsilon\right\}} \varepsilon\ dUx \\ &= \varepsilon^2 \cdot U\left(\left\{\frac{dP}{dU}(x) > \varepsilon\right\} \cap \left\{\frac{dP}{dU}(y-x) > \varepsilon\right\}\right). \end{aligned}$$

Using the inclusion-exclusion inequalities we get

$$\begin{aligned} &U\left(\left\{\frac{dP}{dU}(x) > \varepsilon\right\} \cap \left\{\frac{dP}{dU}(y-x) > \varepsilon\right\}\right) \\ &\quad = U\left\{\frac{dP}{dU}(x) > \varepsilon\right\} + U\left\{\frac{dP}{dU}(y-x) > \varepsilon\right\} - U\left(\left\{\frac{dP}{dU}(x) > \varepsilon\right\} \cup \left\{\frac{dP}{dU}(y-x) > \varepsilon\right\}\right) \\ &\qquad \geq 2 \cdot U\left\{\frac{dP}{dU}(x) > \varepsilon\right\} - 1. \end{aligned}$$

Hence

$$\frac{d(P*P)}{dU}(y) \geq 2\varepsilon^2\left(U\left\{\frac{dP}{dU}(x) > \varepsilon\right\} - 1/2\right)$$

for all $y \in G$. □

Combining Theorem 17, Lemma 22, and Proposition 23 we get the following result.

Theorem 24. *Let P be a probability measure on a compact group G with Haar probability measure U. If the support of P is not contained in any coset of a proper subgroup of G and $D(P\|U)$ is finite then the rate of convergence of $D(P^{*n}\|U)$ to zero is exponential.*

As a corollary we get the following result that was first proved by Kloss [21] for total variation.

Corollary 25. *Let P be a probability measure on the compact group G with Haar probability measure U. If the support of P is not contained in any coset of a proper subgroup of G and $D(P\|U)$ is finite then P^{*n} converges to U in variation and the rate of convergence is exponential.*

Proof. This follows directly from Pinsker's inequality [22, 23]

$$\frac{1}{2}\left\|P^{*n}-U\right\|^{2}\leq D\left(P^{*n}\|U\right).$$ □

Corollary 26. *Let P be a probability measure on the compact group G with Haar probability measure U. If the support of P is not contained in any coset of a proper subgroup of G and $D\left(P\|U\right)$ is finite, then the density*

$$\frac{dP^{*n}}{dU}$$

converges to 1 point wise almost surely for n tending to infinity.

Proof. The variation norm can be written as

$$\left\|P^{*n}-U\right\|=\int_{G}\left|\frac{dP^{*n}}{dU}-1\right|\,dU.$$

Thus

$$U\left(\left|\frac{dP^{*n}}{dU}-1\right|\geq\varepsilon\right)\leq\frac{\left\|P^{*n}-U\right\|}{\varepsilon}.$$

The result follows by the exponential rate of convergence of P^{*n} to U in total variation combined with the Borel-Cantelli Lemma. □

7. Discussion

In this paper we have assumed the existence of the Haar measure by referring to the literature. With the Haar measure we have then proved convergence of convolutions using Markov chain techniques. The Markov chain approach can also be used to prove the existence of the Haar measure by simply referring to the fact that a homogenous Markov chain on a compact set has an invariant distribution. The problem about this approach is that the proof that a Markov chain on a compact set has an invariant distribution is not easier than the proof of the existence of the Haar measure and is less known.

We have shown that the Haar probability measure maximizes the rate distortion function at any distortion level. The normal proofs of the existence of the Haar measure use a kind of covering argument that is very close to the techniques found in rate distortion technique. There is a chance that one can get an information theoretic proof of the existence of the Haar measure. It seems obvious to use concavity arguments as one would do for Shannon entropy but, as proved by Ahlswede [24], the rate distortion function at a given distortion level is not a concave function of the underlying distribution, so some more refined technique is needed.

As noted in the introduction for any algebraic structure A the group $Aut\left(A\right)$ can be considered as symmetry group, it it has a compact subgroup for which the results of this paper applies. It would be interesting to extend the information theoretic approach to the algebraic object A itself, but in general there is no known equivalent to the Haar measure for other algebraic structures. Algebraic structures are used extensively in channel coding theory and cryptography so although the theory may become more involved extensions of the result presented in this paper are definitely worthwhile.

Acknowledgement

The author want to thank Ioannis Kontoyiannis for stimulating discussions.

References and Notes

1. Jaynes, E. T. Information Theory and Statistical Mechanics, I and II. *Physical Reviews* **1957**, *106 and 108*, 620–630 and 171–190.
2. Topsøe, F. Game Theoretical Equilibrium, Maximum Entropy and Minimum Information Discrimination. In *Maximum Entropy and Bayesian Methods*; Mohammad-Djafari, A.; Demoments, G., Eds. Kluwer Academic Publishers: Dordrecht, Boston, London, 1993, pp. 15–23.
3. Jaynes, E. T. Clearing up mysteries – The original goal. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed. Kluwer: Dordrecht, 1989.
4. Kapur, J. N. *Maximum Entropy Models in Science and Engineering*, revised Ed. Wiley: New York, 1993.
5. Grünwald, P. D.; Dawid, A. P. Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory. *Annals of Mathematical Statistics* **2004**, *32*, 1367–1433.
6. Topsøe, F. Information Theoretical Optimization Techniques. *Kybernetika* **1979**, *15*, 8 – 27.
7. Harremoës, P.; Topsøe, F. Maximum Entropy Fundamentals. *Entropy* **2001**, *3*, 191–226.
8. Jaynes, E. T. *Probability Theory - The Logic of Science*. Cambridge University Press: Cambridge, 2003.
9. Csiszár, I. Sanov Property, Generalized I-Projection and a Conditional Limit Theorem. *Ann. Probab.* **1984**, *12*, 768–793.
10. Stromberg, K. Probabilities on compact groups. *Trans. Amer. Math. Soc.* **1960**, *94*, 295–309.
11. Csiszár, I. A note on limiting distributions on topological groups. *Magyar Tud. Akad. Math. Kutaló INt. Kolzl.* **1964**, *9*, 595–598.
12. Schlosman, S. Limit theorems of probability theory for compact groups. *Theory Probab. Appl.* **1980**, *25*, 604–609.
13. Johnson, O. *Information Theory and Central Limit Theorem*. Imperial Collage Press: London, 2004.
14. Haar, A. Der Massbegriff in der Theorie der kontinuierlichen Gruppen. *Ann. Math.* **1933**, *34*.
15. Halmos, P. *Measure Theory*. D. van Nostrand and Co., 1950.
16. Conway, J. *A Course in Functional Analysis*. Springer-Verlag: New York, 1990.
17. Vogel, P. H. A. On the Rate Distortion Function of Sources with Incomplete Statistics. *IEEE Trans. Inform. Theory* **1992**, *38*, 131–136.
18. Johnson, O. T.; Suhov, Y. M. Entropy and convergence on compact groups. *J. Theoret. Probab.* **2000**, *13*, 843–857.
19. Harremoës, P.; Holst, K. K. Convergence of Markov Chains in Information Divergence. *Journal of Theoretical Probability* **2009**, *22*, 186–202.
20. Topsøe, F. An Information Theoretical Identity and a problem involving Capacity. *Studia Scientiarum Mathematicarum Hungarica* **1967**, *2*, 291–292.
21. Kloss, B. Probability distributions on bicompact topological groups. *Theory Probab. Appl.* **1959**,

4, 237–270.

22. Csiszár, I. Information-type measures of difference of probability distributions andindirect observations. *Studia Sci. Math. Hungar.* **1967**, *2*, 299–318.
23. Fedotov, A.; Harremoës, P.; Topsøe, F. Refinements of Pinsker's Inequality. *IEEE Trans. Inform. Theory* **2003**, *49*, 1491–1498.
24. Ahlswede, R. F. Extremal Properties of Rate-Distortion Functions. *IEEE. Trans. Inform. Theory* **1990**, *36*, 166–171.