Information Theory for Angular Data

Peter Harremoës  
Centrum Wiskunde & Informatica (CWI)  
Amsterdam, The Netherlands  
Email: P.Harremoes@cwi.nl

Abstract—In this paper we study some aspect of information theory for data with values on a circle or on a sphere. We develop rate distortion theory with squared Euclidian distance as distortion function. Various inequalities are proved including a new Pinsker type inequality. A rate distortion test of uniformity is developed. Pointwise convergence of a sequence of information projections is conjectured and some evidence is provided.

I. INTRODUCTION

Information theory for data with values in \( \{0, 1\} \) or in \( \mathbb{R} \) is well developed. For both these structures one can calculate the rate distortion function and we have an entropy power inequality. Here we shall develop aspects of information theory on angular data. By angular data we mean data with values on a circle or more generally on a sphere. For a textbook on the standard approach to the statistical analysis of angular data we refer to [1]. See [2] for some previous quite different results relating harmonic analysis and information theory.

The spheres are denoted \( s_j \) where \( j \) is the dimension of the sphere as a manifold. One important property of \( s_j \) is it has a natural embedding in \( \mathbb{R}^{j+1} \) (center in origin and radius 1). As usual we shall use squared Euclidian distance as distortion function in \( \mathbb{R}^2 \) and \( \mathbb{R}^3 \) and this is also the distortion function we shall use for \( s_1 \) and \( s_2 \) via their natural embedding in \( \mathbb{R}^2 \) and \( \mathbb{R}^3 \). With such an embedding \( s_j \) has \( SO(j+1) \) as symmetry group. Some of the results derived in this article are related to results on their symmetry group. We refer to [3] for some recent results about information theory on compact groups. Note that the sphere \( s_1 \) is special in that it can be identified with its symmetry group \( SO(2) \). The unique distributions on \( s_j \) invariant under rotations will be called the uniform distributions and be denoted by \( U \).

In this note \( D(\cdot||\cdot) \) shall denote information divergence. Sometimes we shall use \( H(P) \) as an abbreviation of \(-D(P||U)\).

II. DIVERGENCE AND DISTORTION

For angular data the von Mises distributions play a role similar to the role of the Gaussian distribution of data in a vector space. The von Mises distributions form an exponential family based on the uniform distribution on \( s_j \subseteq \mathbb{R}^{j+1} \) that can be parametrized by the mean values that are interior points in the ball with \( s_j \) as boundary. Let \( \beta \) denote the canonical parameter of a von Mises distribution \( P_\beta \). The density with respect to the uniform distribution is

\[
\frac{dP_\beta}{dU} = \frac{\exp(\beta \cdot x)}{Z(\beta)}
\]

where \( Z(\beta) = \int \exp(\beta \cdot x) \ dU \) denotes the partition function. Then

\[
D(P_\beta_1 || P_\beta_2) = \int \log \left( \frac{\exp(\beta_1 \cdot x)}{Z(\beta_1)} \right) \frac{dP_{\beta_1}}{dU} \ dU
= (\beta_1 - \beta_2) \cdot \int x \frac{dP_{\beta_1}}{dU} \ dU - \log \left( \frac{Z(\beta_1)}{Z(\beta_2)} \right)
= (\beta_1 - \beta_2) \cdot \mu_{\beta_1} - \log \left( \frac{Z(\beta_1)}{Z(\beta_2)} \right).
\]

Now assume that \( \|\beta_1\| = \|\beta_2\| \). Then the last term vanish and

\[
D(P_\beta_1 || P_\beta_2) = (\beta_1 - \beta_2) \cdot \mu_{\beta_1} = \|\beta_1\| \cdot \mu_{\beta_1} ||1 - \cos(\theta)||
\]

where \( \theta = \angle(\beta_1, \beta_2) \). We see that the information divergence is proportional to where \( \theta \) the angle between the directions. We will define the distortion between angle \( \theta_1 \) and \( \theta_2 \) to be \( d(\theta_1, \theta_2) = \sqrt{2(1 - \cos(\theta_1 - \theta_2))} \). If \( \theta_1 \) and \( \theta_2 \) are identified with unit vectors \( x_1 \) and \( x_2 \) then

\[
\|x_1 - x_2\|^2 = \left( 2 \sin \left( \frac{\angle(x_1, x_2)}{2} \right) \right)^2
= 2(1 - \cos(\angle(x_1, x_2)))
= d(\theta_1, \theta_2).
\]

If a number of observation \( x_1, x_2, \ldots, x_n \) in \( s_j \) are made then the maximum likelihood distribution in the exponential family of von Mises distributions is the distribution with mean equal to the sample average \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \). As usual \( \bar{x} \) is the

\[
F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2} t^2} \ dt
\]

Fig. 1. Distortion between two angles is their squared Euclidean distance.
vector that minimizes the total squared Euclidean distance
\[ \sum \| x_i - x \|^2 \] over all \( x \in \mathbb{R}^{d+1} \). The angle of direction of the maximum likelihood von Mises distribution is \( \bar{x} / \| \bar{x} \| \). Now
\[ \sum \| x_i - x \|^2 = \sum \| x_i - \bar{x} \|^2 + n \| x - \bar{x} \|^2 \]
so the direction of the maximum likelihood von Mises distribution is the minimizer of (1) over all \( x \in s_j \). We see that using squared Euclidean distance as distortion measure implies that the maximum likelihood direction of a sample is the direction that minimizes distortion.

For angles \( \phi, \psi \in s_1 \) squared Euclidean distance also appear if \( \phi \) and \( \psi \) as phase translations of oscillations. This follows by
\[
\frac{1}{2\pi} \int_0^{2\pi} (\sin(x + \phi) - \sin(x + \psi))^2 \, dx = 1 - \cos(\phi - \psi).
\]
Here the squared difference \( (\sin(x + \phi) - \sin(x + \psi))^2 \) has a natural interpretation as the energy of the interfering oscillations. Therefore the squared Euclidean distance has a natural interpretation as a time average in interference energy.

An obvious alternative to the use of squared Euclidean distance is to use the absolute value of angle between two angles as distortion function. This would be more in accordance with the view of a sphere as a Riemannian manifold where we measure distance between points as the length of the shortest curve between them. The elementary inequalities
\[
\frac{4\theta^2}{\pi^2} \leq 2 - 2 \cos(\theta) \leq 2\theta, \quad \theta \in [0, \pi]
\]
can be used to translate inequalities for squared Euclidean distance into inequalities for distance along the manifold.

III. DIVERGENCE AND WASSERSTEIN DISTANCE

Pinsker’s inequality [8] relates information divergence and variational distance by
\[
\frac{1}{2} \| P - Q \|^2_{\text{tot}} \leq D(P \| Q).
\]
This inequality applies to angular data, but total variation is the transportation distance metric corresponding to Hamming distance, so we want a similar inequality for squared Euclidean distance. Let \( P \) and \( Q \) denote probability distributions on \( \mathbb{R}^{d+1} \). The squared Wasserstein distance between \( P \) and \( Q \) is given by
\[
W_2^2(P, Q) = \inf \| X - Y \|^2_{\text{tot}} \text{ where the infimum is taken over all joint distributions of } X \text{ and } Y \text{ with values in } s_j \text{ such that } P \text{ is the distribution of } X \text{ and } Q \text{ is the distribution of } Y. \text{ We shall give a bound on Wasserstein distance in terms of divergence.}
\]
Let \( Z \) denote a random variable with values in \( \mathbb{R}^{d+1} \) and with a standard Gaussian distribution \( G \). Then \( Z / \|Z\|_2 \) is uniform \( U \) on \( s_j \). Let \( X_0 \) denote a random variable with distribution \( P \) on \( s_j \). Let \( \tilde{P} \) denote the distribution of \( X = \|Z\|_2 \cdot X_0 \). Next using an inequality that holds for the Gaussian distribution [9, Eq. 22.0.6] we get
\[
-H(P) = D(\tilde{P} \| G) \geq \frac{1}{2} W_2^2(\tilde{P}, G) \geq \frac{1}{2} E \left[ \| X - Y \|^2_{\text{tot}} \right],
\]
where \( Y \) is random variable with marginal distribution \( G \) and jointly distributed with \( X \) such that the Wasserstein distance is achieved. Now \( \frac{Z}{\|Z\|_2} \) has a uniform distribution on \( s_j \) and \( \frac{X}{\|X\|_2} \) has distribution \( P \) on \( s_j \). We state the following lemma without proof.

Lemma 1: If \( \|X\|_2 \) and \( \|Y\|_2 \) have the same distribution then
\[
E \left[ \| X - Y \|^2_{\text{tot}} \right] \geq E \left[ \frac{\|Y\|_2}{2} \right] E \left[ \left\| \frac{X}{\|X\|_2} - \frac{Y}{\|Y\|_2} \right\|^2 \right].
\]
Combining Lemma 1 with Inequality 3a we get
\[
-H(P) \geq \frac{E \left[ \|Y\|_2 \right]}{4} E \left[ \left\| \frac{X}{\|X\|_2} - \frac{Y}{\|Y\|_2} \right\|^2 \right] = \frac{j + 1}{4} E \left[ \left\| \frac{X}{\|X\|_2} - \frac{Y}{\|Y\|_2} \right\|^2 \right].
\]
This leads us to the following theorem.

Theorem 2: For a probability measure \( P \) on \( s_j \) the following inequality holds.
\[
\frac{j + 1}{4} W_2^2(P, U) \leq -H(P).
\]
We do not know whether this inequality is tight or whether a better constant than \( (j + 1)/4 \) can be found. For sequences of data we have the following result that can be used to prove a measure concentration inequality related to the work of K. Marton [10], [11].

Proposition 3: Let \( P \) denote a probability distribution on \( s_j^n \). Then
\[
\frac{j + 1}{4} W_2^2(P, U^n) \leq D(P \| U^n).
\]
Proof: Let $P$ be a probability distribution on $s_j^n$. For $i = 0, 1, ..., n$ we define random variables $X^i$ and $Y^i$ with values in $A^i$. For $i = 0$ the construction is trivial. Assume that the random variables $X^i$ and $Y^i$ have been constructed. Then for $X^i = x^i$ there exist random variables $\tilde{X}^i$ and $\tilde{Y}^i$ with values in $A$ such that
\[
\left\| \tilde{X}^i - \tilde{Y}^i \mid X^i = x^i \right\|^2_2 \leq 4D \left( P_{s_j^{i+1}} \mid Q_{s_j^{i+1}} \mid X^i = x^i \right).
\]
By applying Jensen’s inequality once more we get
\[
E \left[ \left\| X - Y \right\|^2_2 \right] = \sum_{i=1}^n E \left[ \left\| \tilde{X}^i - \tilde{Y}^i \mid X^i \right\|^2_2 \right]
\leq \sum_{i=1}^n 4D \left( P_{s_j^{i+1}} \mid U_{s_j^{i+1}} \mid X^i \right) \frac{j+1}{j+1}
\leq \frac{4}{j+1} \sum_{i=1}^n \frac{D \left( P_{s_j^{i+1}} \mid U_{s_j^{i+1}} \mid X^i \right)}{j+1}
= \frac{4D \left( P \mid U \right)}{j+1}.
\]

IV. DIVERGENCE AND THE RAYLEIGH STATISTIC

In the exponential family of von Mises distributions the sample average is the sufficient statistic when the data are identified with vectors in the ball with boundary $s_j \subseteq \mathbb{R}^{j+1}$. Hence if we want to test the uniformity of a distribution against alternatives in the exponential family of von Mises distributions the sample average the most efficient is to use the statistic that is a function of the sample average. Because of the build-in symmetry one will normally use the norm of the sample average to test uniformity. This is called the Rayleigh statistic. The Rayleigh test rejects uniformity if the Rayleigh statistics exceeds some threshold value. Although the Rayleigh test is designed for von Mises distributions as alternatives to uniform distribution it is often used as a general test for uniformity [1], [12], [4].

Let $X$ be a random vector with distribution $P$ and $Y$ a random vector with distribution $U$. Then
\[
W^2 (P, U) \geq E \left[ \left\| X - Y \right\|^2_2 \right]
= E \left[ \left\| X - Y \right\|^2_2 + \left\| X - Y \right\|^2_2 \right]
\geq \left\| E \left[ X \right] \right\|^2_2
\frac{j+1}{4} \left\| E \left[ X \right] \right\|^2_2 \leq -H (P).
\]
Similar inequalities have been proved for Gaussian and Poisson distributions but in these cases the underlying space is not compact which may be the reason why a simple derivation from a bound on Wasserstein distance has not been found in these cases. Inequality 4 is of direct relevance to the Rayleigh test. If $T$ denotes the norm of the empirical mean then
\[
-\frac{1}{n} \log \Pr \left( T \geq T_0 \right) \geq \frac{j+1}{4} T_0^2.
\]
For a distribution $P$ on $s_1$ that we shall now identify $[0; 2\pi]$ the mean value of the random vector $X$ that is
\[
\int_{0}^{2\pi} \left( \cos t \sin t \right) dP \left( t \right) = \left( a_1 b_1 \right)
\]
where $a_j$ and $b_j$ denotes the $j$’th Fourier coefficients of $P$. Thus $\left\| E \left[ X \right] \right\|^2_2 = a_1^2 + b_1^2$ and $a_j^2 + b_j^2 \leq -H (P)$. This holds for all probability measures on $[0; 2\pi]$. Let $f_j$ denote the transformation $x \rightarrow jx \mod 2\pi$. Then $D (P \mid U) \geq D (f_j (P) \mid f_j (U))$ implying that
\[
-H (P) \geq \frac{a_j^2 + b_j^2}{2}
\]
because the first Fourier coefficients of $f_j (P)$ equals the $j$’th Fourier coefficients of $P$.

It is also possible to upper bound divergence by Fourier coefficients. The Rényi divergence of order $\alpha$ is given by
\[
D_\alpha (P \mid Q) = \frac{\log \int \left( \frac{dP}{dQ} \right)^\alpha dQ}{\alpha - 1}.
\]

Then, according to Parceval’s identity we have
\[
D (P \mid U) \leq D_2 (P \mid U) = \log \int \left( \frac{dP}{dU} \right)^2 dU
= \log \frac{1}{2\pi} \int_{0}^{2\pi} \left( \frac{dP}{dU} (t) \right)^2 dt
= \log \left( 1 + \sum_{j=1}^{\infty} \left( a_j^2 + b_j^2 \right) \right) \leq \sum_{j=1}^{\infty} \left( a_j^2 + b_j^2 \right).
\]
Note that in general
\[
-H (P) \geq \frac{\sum_{j=1}^{\infty} \left( a_j^2 + b_j^2 \right)}{2}
\]
because $\log (1 + x) \leq x/2$ and $\sum_{j=1}^{\infty} \left( a_j^2 + b_j^2 \right)$ may take arbitrary large values whereas $a_j^2 + b_j^2$ is upper bounded by 1 for probability measures.

This type of inequality is of particular interest if we consider a noise model where the density of $P = P_t$ changes over time according to the heat (diffusion) equation. Let $f (x, t)$ be the density as function of place and time. Then
\[
\frac{\partial f}{\partial t} = a \frac{\partial^2 f}{\partial x^2}.
\]
If $f (x, 0) = \frac{1}{2\pi} + \sum_{j=1}^{\infty} \left( a_j \cos jx + b_j \sin jx \right)$ then
\[
f (x, t) = \frac{1}{2\pi} + \sum_{j=1}^{\infty} \exp \left( -a_j^2 t \right) \left( a_j \cos jx + b_j \sin jx \right).
\]
We see that the entropy converges exponentially to 0 and that the rate of convergence is determined by the first non-trivial Fourier coefficient in the Fourier series of $P$. 

183
V. Calculation of the Rate Distortion Function

We shall calculate the rate distortion function for the uniform distribution when \( s_j \) is used both as source and reconstruction alphabet with squared Euclidean distance as distortion function.

Compactness of \( s_j \) implies that the rate distortion function is finite and can be studied using its convex conjugate.

Let \( P \) be a probability measure on \( s_j \). Let \( X \) be a random variable with values in \( s_j \) and distribution \( P \), and let \( \hat{X} \) be a random variable coupled with \( X \) such that \( E \left[ \left\| X - \hat{X} \right\|_2^2 \right] = d_0 \). Introduce \( e_1 = (1,0) \in s_1 \) or \( e_1 = (1,0,0) \in s_2 \). For each \( \hat{X} \) we choose a rotation \( g_X \in O(j+1) \) such that \( g_X \left( \hat{X} \right) = e_1 \). Then

\[
I \left( X, \hat{X} \right) = H \left( X \right) - H \left( X \mid \hat{X} \right) \\
\geq H \left( X \right) - H \left( g_X \left( X \right) \right) = D \left( g_X \left( X \right) \parallel U \right) - D \left( X \parallel U \right).
\]

Now,

\[
E \left[ \left\| X - \hat{X} \right\|_2^2 \right] = E \left[ \left\| g_X \left( X \right) - e_1 \right\|_2^2 \right]
\]

and

\[
-H \left( g_X \left( X \right) \right) \geq -H \left( Y \right)
\]

under the condition \( E \left[ \left\| Y - e_1 \right\|_2^2 \right] = d_0 \). The minimum of the rate divergence is achieved for a random variable with a von Mises distribution \( P_\beta \) [1] given by the density

\[
\frac{dP_\beta}{dU_y} (y) = \frac{\exp \left( \beta \cdot \left\| y - e_1 \right\|_2^2 \right)}{Z (\beta)},
\]

where \( Z (\beta) \) is the partition function given by

\[
Z (\beta) = \int_{s_j} \exp \left( \beta \cdot \left\| y - e_1 \right\|_2^2 \right) dU_y.
\]

For both \( s_1 \) and \( s_2 \) one gets quite simple expressions for the partition function. The case \( s_1 \) was discussed in [3]. For \( s_2 \) we get

\[
Z (\beta) = \int_{s_j} \exp \left( \beta \cdot \left( 2 - y_1 \right) \right) dU_y
\]

\[
= \exp \left( 2\beta \right) \frac{1}{2} \int_{-1}^{1} \exp \left( -\beta y \right) dy
\]

\[
= \exp \left( 2\beta \right) \frac{\exp \left( \beta \right) - \exp \left( -\beta \right)}{2\beta}.
\]

An upper bound on the rate distortion function of the distribution \( P \) is obtained as follows. Let \( X \) be a random variable with distribution \( P \). We will specify the distribution of \( \hat{X} \) given \( X \). Assume that \( X = x \). Then the conditional distribution of \( \hat{X} \) given \( X = x \) is given by \( g^{-1}_x \left( P_\beta \right) \). The mean distortion is then \( d_0 \) for all values of \( x \). The mutual information is

\[
I \left( X, \hat{X} \right) = H \left( \hat{X} \right) - H \left( \hat{X} \mid X \right)
\]

\[
= H \left( \hat{X} \right) - H \left( P_\beta \right) \leq -H \left( P_\beta \right).
\]

We see that \( H \left( P \right) - H \left( P_\beta \right) \leq R_P \left( d_0 \right) \leq -H \left( P_\beta \right) \). This holds for all \( P \) and in particular for \( P = U \). Therefore \( R_U \left( d_0 \right) = -H \left( P_\beta \right) \) and we get

\[
R_U - D \left( P \parallel U \right) \leq R_P \leq R_U.
\]

In particular the uniform distribution on a sphere is the maximum entropy distribution in the sense that for any distortion level it maximizes the value of the rate distortion function. By standard arguments we get that the convex conjugate \( R^*_U \left( \beta \right) \) of the rate distortion function equals \( \log ( Z (\beta) ) \).

VI. Rate Distortion Tests

The results on the rate distortion function can be used to test uniformity of a distribution using a so-called rate distortion test as described in [13]. The idea is to smooth both the empirical distribution and the uniform distribution by a von Mises distribution because this smoothing correspond to the optimal compression in the rate distortion sense. After smoothing we compare the smoothed empirical distribution with the uniform distribution by the information divergence from the smoothed empirical distribution to the uniform distribution. The distortion level is decreased slowly as the sample size is increased. The Hodge and Lehman efficiency of the rate distortion test was proved in [13]. The Bahadur efficiency for the rate distortion test of uniformity on \( s_j \) that can be identified with \( SO \left( 2 \right) \), was proved in [13]. Here we shall show the Bahadur efficiency of the rate distortion test on \( s_j \) for \( j \in \mathbb{N} \).

We shall analyze this question in the case of testing uniformity of angular data because this is of particular simplicity because angles can be identified with elements of \( SO \left( 2 \right) \).

Theorem 4: Let \( d_n \) denote a decreasing sequence of positive distortion values. Let \( \Psi \left( d_n \right) \) denote the associated Markov kernels that correspond to optimal compression at that distortion level. Assume that \( P \) generates data. Then

\[
\lim \inf D \left( \Psi \left( d_n \right) \left( \text{Emp}_n \left( \omega \right) \right) \parallel U \right) \geq D \left( P \parallel U \right)
\]
almost surely.

Proof: See [13].

The theorem implies that for any \( K < D(P\|U) \) we have \( D(\Psi_{d_n}(Emp_n(\omega))\|U) \geq K \) eventually almost surely so if \( P \) is the distribution of the alternative hypothesis then and the power of the test is kept fixed, then the acceptance regions of alternative \( P \) in the rate distortion test must have the form \( D(\Psi_{d_n}(Emp_n(\omega))\|U) \geq K_n \) for \( K_n \to D(P\|U) \).

In order to determine the Bahadur efficiency we have to find \( \frac{d}{dU} \) under the null hypothesis that data are generated by a uniform distribution. We do this by covering \( s_j \) by \( k_n \) distortion balls of radius \( r_n \) in such a way that \( r_n \) is minimal. We choose \( k_n \) such that \( \frac{1}{k_n \log \frac{1}{k_n}} \to \infty \) for \( n \to \infty \). Let \( F_n \) denote the \( \sigma \)-algebra generated by these intervals. Then

\[
\lim -\frac{1}{n} \Pr \left( \frac{D(\Psi_{d_n}(Emp_n(\omega))\|F_n) \geq K_n}{U|F_n} \right) = D(P\|U).
\]

We are interested in

\[
D(\Psi_{d_n}(Emp_n(\omega))\|U) = D(\Psi_{d_n}(Emp_n(\omega))\|\Psi_{d_n}(U))
\]

and not \( D(\Psi_{d_n}(Emp_n(\omega))\|U|F_n) \) but each subinterval has length \( 2\pi/k_n \) so

\[
\left| \log \frac{d\Psi_{d_n}(Emp_n(\omega))}{d\Psi_{d_n}(U)} - \log \frac{dEmp_n(\omega)}{d\Psi_{d_n}(U)} \right| 
\leq \sup_{|x-x_0| \leq r_n} \left| \log \frac{d\Psi_{d_n}(x)}{d\Psi_{d_n}(x_0)} \right|.
\]

Therefore

\[
D(\Psi_{d_n}(Emp_n(\omega))\|U) \leq D\left( Emp_n(\omega)\|U|F_n \right) + \sup_{|x-x_0| \leq r_n} \left| \log \frac{d\Psi_{d_n}(x)}{d\Psi_{d_n}(x_0)} \right|
\]

where \( x \) and \( x_0 \) denote points in \( s_j \). Therefore

\[
\lim -\frac{1}{n} \Pr \left( D(\Psi_{d_n}(Emp_n(\omega))\|U) \geq K_n \right) \geq K
\]

if the sequence \( d_n \) tends to zero so slowly that \( \sup_{|x-x_0| \leq r_n} \left| \log \frac{d\Psi_{d_n}(x)}{d\Psi_{d_n}(x_0)} \right| \to \infty \) for \( n \to \infty \). This leads us to the following theorem.

**Theorem 5:** The rate distortion test of uniformity of angular data is Bahadur efficient if the distortion level tends to zero sufficiently slowly.

**VII. DISCUSSION**

Let \( P \) be a probability distribution on \( s_1 \) that we shall identify with \([0; 2\pi]\). We shall assume that \( D(P\|U) < \infty \). Let \( C_k \) denote the convex set of distributions \( Q \) on \( s_1 \) with the same Fourier coefficients as \( P \) for orders less than or equal to \( k \). Then \( C_1 \supseteq C_2 \supseteq C_3 \supseteq \cdots \). Let \( P_k \) denote the information projection of \( U \) into \( C_k \). Note in particular that \( P_1 \) is a von Mises distribution. Then \( D(P\|P_k) \to 0 \) for \( k \to \infty \) according to [14]. We are interested in whether convergence of \( P_k \) to \( P \) in information implies almost sure pointwise convergence of the density \( dP_k/dU \) to \( dP/dU \). By information theoretic techniques the almost sure pointwise convergence a system of information projections has been proved in the special case where \( dP_k/dU \) form a martingale if \( D(P\|U) < \infty \) [15].

A sequence of truncated Fourier series \( f_k \) can be interpreted as \( L^2 \) projections of (the density of) \( P \) into \( C_k \) and according to the celebrated theorem by Carleson \( f_k \) converges pointwise almost surely to \( dP/dU \) if \( dP/dU \in L^2([0; 2\pi]) \) [16], [17]. Kolmogorov has proved that almost sure convergence need not hold if the condition on \( P \) is relaxed to \( dP/dU \in L^1([0; 2\pi]) \). Newer results by Sjölin and others implies that almost sure pointwise convergence can be proved under the much weaker condition

\[
\int_0^{2\pi} \frac{dP}{dU} \left( \log \frac{dP}{dU} \right)^{1+\theta} dU < \infty
\]

for some \( \theta > 0 \). These results give a strong indication that almost sure pointwise convergence of \( dP_n/dU \) to 1 can be proved if \( D(P\|U) < \infty \).

**References**


