Szegö’s theorem and its probabilistic descendants

N.H. Bingham

Mathematics Department, Imperial College London, London SW7 2AZ, UK
e-mail: nick.bingham@btinternet.com; n.bingham@ic.ac.uk

Abstract: The theory of orthogonal polynomials on the unit circle (OPUC) dates back to Szegö’s work of 1915-21, and has been given a great impetus by the recent work of Simon, in particular his survey paper and three recent books; we allude to the title of the third of these, Szegö’s theorem and its descendants, in ours. Simon’s motivation comes from spectral theory and analysis. Another major area of application of OPUC comes from probability, statistics, time series and prediction theory; see for instance the classic book by Grenander and Szegö, Toeplitz forms and their applications. Coming to the subject from this background, our aim here is to complement this recent work by giving some probabilistically motivated results. We also advocate a new definition of long-range dependence.

AMS 2000 subject classifications: Primary 60G10; secondary 60G25.

Keywords and phrases: Stationary process, prediction theory, orthogonal polynomials on the unit circle, partial autocorrelation function, moving average, autoregressive, long-range dependence, Hardy space, cepstrum.

Received July 2011.

Contents

1 Introduction .................................................. 288
2 The Kolmogorov Isomorphism Theorem ..................... 291
3 Verblunsky’s theorem and partial autocorrelation ........ 293
4 Weak conditions: Szegö’s theorem .......................... 295
5 Strong conditions: Baxter’s theorem ....................... 302
6 Strong conditions: The strong Szegö theorem ............. 305
6.1 $\phi$-mixing .............................................. 308
7 Intermediate conditions .................................. 309
7.1 $\rho$-mixing .............................................. 309
7.2 Positive angle: The Helson-Szegö and Helson-Sarason conditions 309
7.3 Pure minimality ........................................ 310
7.4 Rigidity; $(LM)$, $(CND)$, $(IPF)$ .................... 311
8 Remarks ..................................................... 312
Acknowledgements .......................................... 314
References .................................................. 314

287
1. Introduction

The subject of orthogonal polynomials on the real line (OPRL), at least some of which forms part of the standard undergraduate curriculum, has its roots in the mathematics of the 19th century. The name of Gabor Szegő (1895-1985) is probably best remembered nowadays for two things: co-authorship of ‘Pólya and Szegő’ [PoSz] and authorship of ‘Szegő’ [Sz4], his book of 1938, still the standard work on OPRL. Perhaps the key result in OPRL concerns the central role of the three-term recurrence relation ([Sz4], III.3.2: ‘Favard’s theorem’).

Much less well known is the subject of orthogonal polynomials on the unit circle (OPUC), which dates from two papers of Szegő in 1920-21 ([Sz2, Sz3]), and to which the last chapter of [Sz4] is devoted. Again, the key is the appropriate three-term recurrence relation, the Szegő recursion or Durbin-Levinson algorithm (§3). This involves a sequence of coefficients (not two sequences, as with OPRL), the Verblunsky coefficients $\alpha = (\alpha_n)$ (§3), named (there are several other names in use) and systematically exploited in the magisterial two-volume book on OPUC ([Si4, Si5]) by Barry Simon. See also his survey paper [Si3], written from the point of view of analysis and spectral theory, the survey [GolTo], and his recent book [Si9].

Complementary to this is our own viewpoint, which comes from probability and statistics, specifically time series (as does the excellent survey of 1986 by Bloomfield [Bl3]). Here we have a stochastic process (random phenomenon unfolding with time) $X = (X_n)$ with $n$ integer (time discrete, as here, corresponds to compactness of the unit circle by Fourier (or Pontryagin) duality, whence the relevance of OPUC; continuous time is also important, and corresponds to OPRL). To set the stage for this, we begin (§1) with the Kolmogorov Isomorphism Theorem [Kol] of 1941, allowing us to pass between the time domain and frequency domain. This is the appropriate instance for us of the spectral theorem, and allows us to make systematic use of Hilbert-space methods.

We make a simplifying assumption, and restrict attention to the stationary case. The situation is then invariant under the shift $n \mapsto n + 1$, which makes available the powerful mathematical machinery of Beurling’s work on invariant subspaces ([Beu, Nik1]). While this is very convenient mathematically, it is important to realize that this is both a strong restriction and one unlikely to be satisfied exactly in practice. One of the great contributions of the statistician and econometrician Sir Clive Granger (1934-2009) was to demonstrate that statistical/econometric methods appropriate for stationary situations can, when applied indiscriminately to non-stationary situations, lead to misleading conclusions (via the well-known statistical problem of spurious regression). This has profound implications for macroeconomic policy. Governments depend on statisticians and econometricians for advice on interpretation of macroeconomic data. When this advice is misleading and mistaken policy decisions are implemented, avoidable economic losses (in terms of GDP) may result which are large-scale and permanent (cf. Japan’s ‘lost decade’ in the 1990s, or lost two decades, and the global problems of 2007-8 on).
The mathematical machinery needed for OPUC is function theory on the (unit) disc, specifically the theory of Hardy spaces and Beurling’s theorem (factorization into inner and outer functions and Blaschke products). We shall make free use of this, referring for what we need to standard works (we recommend [Du, Ho, Gar, Koo1, Nik1, Nik2]), but giving detailed references. The theory on the disc (whose boundary the circle is compact) corresponds analytically to the theory on the upper half-plane, whose boundary the real line is non-compact (for which see e.g. [DymMcK]). Probabilistically, we work on the disc in discrete time and the half-plane in continuous time. In each case, what dominates is an integrability condition. In discrete time, this is Szegő’s condition (Sz), or non-determinism (ND) – integrability of log \( w \), where \( w \) is the spectral density \( w \) (the density of the absolutely continuous component of the spectral measure \( \mu \)) (§4). In continuous time, this is the logarithmic integral, which gives its name to Koosis’ book [Koo2].

In view of the above, the natural context in which to work is that of complex-valued stochastic processes, rather than real-valued ones, in discrete time. We remind the reader that here the Cauchy-Schwarz inequality tells us that correlation coefficients lie in the unit disc, rather than the interval \([-1, 1]\).

The time-series aspects here go back at least as far as the work of Wiener [Wi1] in 1932 on generalized harmonic analysis, GHA (which, incidentally, contains a good historical account of the origins of spectral methods, e.g. in the work of Sir Arthur Schuster in the 1890s on heliophysics). During World War II, the linear filter (linearity is intimately linked with Gaussianity) was developed independently by Wiener in the USA [Wi2], motivated by problems of automatic fire control for anti-aircraft artillery, and Kolmogorov in Russia (then USSR) [Kol]. This work was developed by the Ukrainian mathematician M. G. Krein over the period 1945-1985 (see e.g. [Dym]), by Wiener in the 1950s ([Wi3], Section IG Prediction and Filtering, including commentaries) and by I. A. Ibragimov (1968 on).

The subject of time series is of great practical importance (e.g. in econometrics), but suffered within statistics by being regarded as ‘for experts only’. This changed with the 1970 book by Box and Jenkins (see [BoxJeRe]), which popularized the subject by presenting a simplified account (including an easy-to-follow model-fitting and model-checking recipe), based on ARMA models (AR for autoregressive, MA for moving average). The ARMA approach is still important; see e.g. Brockwell and Davis [BroDav] for a modern textbook account. The realization that the Verblunsky coefficients \( \alpha \) of OPUC are actually the partial autocorrelation function (PACF) of time series opened the way for the systematic exploitation of OPUC within time series by a number of authors. These include Inoue, in a series of papers from 2000 on (see especially [In3] of 2008), and Inoue and Kasahara from 2004 on (see especially [InKa2] of 2006).

Simon’s work ([Si3, Si4, Si5]) focusses largely on four conditions, two weak (and comparable) and two strong (and non-comparable). Our aim here is to complement the expository account in [Si3] by adding the time-series viewpoint. This necessitates adding (at least) five new conditions. Four of these (comparable) we regard as intermediate, the fifth as strong. In our view, one
needs three levels of strength here, not two. One is reminded of the Goldilocks principle (from the English children’s story: not too hot/hard/high/..., not too cold/soft/low/..., but just right).

We begin in §2 with the Kolmogorov Isomorphism Theorem [Kol] of 1941 (Theorem 1), relating the time domain and frequency domain and introducing the spectral measure $\mu$. We continue in §3 with our key tools: Verblunsky’s theorem (Theorem 2), and the partial autocorrelation function (PACF) $\alpha$. We turn in §4 to weak conditions (Szegö’s condition (Sz), or (ND); Szegö’s theorem (Th. 3); $\alpha \in \ell_2; \sigma > 0$, with $\sigma$ the one-step prediction error). In §5 we look at our first strong condition, Baxter’s condition (B), and Baxter’s theorem (Th. 4: $\alpha \in \ell_1$). The satisfaction or otherwise of Baxter’s condition (B) marks the transition between short- and long-range dependence; in this connection we give Inoue’s theorem, Th. 5. The second strong condition, the strong Szegö condition (sSz), follows in §6 (Th. 6: strong Szegö limit theorem, Ibragimov’s theorem, Golinskii-Ibragimov theorem, Borodin-Okounkov formula; $\alpha \in H^{1/2}$), together with a weakening of (sSz), absolute regularity. We turn in §7 to intermediate conditions: in decreasing order of strength, (i) complete regularity; (ii) positive angle (Helson-Szegö, Helson-Sarason and Sarason theorems); (iii) (pure) minimality (Kolmogorov); (iv) rigidity (Sarason), Levinson-McKean condition (LM), complete non-determinism (CND), intersection of past and future (IPF); see [KasBi] for details. We close in §8 with some remarks.

The (weak) Szegö limit theorem dates from 1915 [Sz1], the strong Szegö limit theorem from 1952 [Sz5]. Simon ([Si4], 11) rightly says how remarkable it is for one person to have made major contributions to the same area 37 years apart. We note that Szegö’s remarkable longevity here is actually exceeded (over the 40 years 1945-1985) by that of the late, great Mark Grigorievich Krein (1907-1989).

What follows is a survey of this area, which contains (at least) eight different layers, of increasing (or decreasing) generality. This is an increase on Simon’s (basic minimum of) four. We hope that no one will be deterred by this increase in dimensionality, and so in apparent complexity. Our aim is the precise opposite: to open up this fascinating area to a broader mathematical public, including the time-series, probabilistic and statistical communities. For this, one needs to open up the ‘grey zone’ between the strong and weak conditions, and examine the third category, of intermediate conditions. We focus on these three levels of generality. This largely reduces the effective dimensionality to three, which we feel simplifies matters. Mathematics should be made as simple as possible, but not simpler (to adapt Einstein’s immortal dictum about physics).

We close by quoting Barry Simon ([Si8], 85): “It’s true that until Euclidean Quantum Field Theory changed my tune, I tended to think of probabilists as a priesthood who translated perfectly simple functional analytic ideas into a strange language that merely confused the uninitiated.” He continues: in his 1974 book on Euclidean Quantum Field Theory, “the dedication says: “To Ed Nelson who taught me how unnatural it is to view probability theory as unnatural” ”.

The (weak) Szegö limit theorem dates from 1915 [Sz1], the strong Szegö limit theorem from 1952 [Sz5]. Simon ([Si4], 11) rightly says how remarkable it is for one person to have made major contributions to the same area 37 years apart. We note that Szegö’s remarkable longevity here is actually exceeded (over the 40 years 1945-1985) by that of the late, great Mark Grigorievich Krein (1907-1989).

What follows is a survey of this area, which contains (at least) eight different layers, of increasing (or decreasing) generality. This is an increase on Simon’s (basic minimum of) four. We hope that no one will be deterred by this increase in dimensionality, and so in apparent complexity. Our aim is the precise opposite: to open up this fascinating area to a broader mathematical public, including the time-series, probabilistic and statistical communities. For this, one needs to open up the ‘grey zone’ between the strong and weak conditions, and examine the third category, of intermediate conditions. We focus on these three levels of generality. This largely reduces the effective dimensionality to three, which we feel simplifies matters. Mathematics should be made as simple as possible, but not simpler (to adapt Einstein’s immortal dictum about physics).

We close by quoting Barry Simon ([Si8], 85): “It’s true that until Euclidean Quantum Field Theory changed my tune, I tended to think of probabilists as a priesthood who translated perfectly simple functional analytic ideas into a strange language that merely confused the uninitiated.” He continues: in his 1974 book on Euclidean Quantum Field Theory, “the dedication says: “To Ed Nelson who taught me how unnatural it is to view probability theory as unnatural” ”.
2. The Kolmogorov Isomorphism Theorem

Let \( X = (X_n : n \in \mathbb{Z}) \) be a discrete-time, zero-mean, (wide-sense) stationary stochastic process, with autocovariance function \( \gamma = (\gamma_n) \),

\[
\gamma_n = E[X_n X_0]
\]

(the variance is constant by stationarity, so we may take it as 1, and then the autocovariance reduces to the autocorrelation).

Let \( \mathcal{H} \) be the Hilbert space spanned by \( X = (X_n) \) in the \( L_2 \)-space of the underlying probability space, with inner product \( \langle X, Y \rangle := E[X\overline{Y}] \) and norm \( \|X\| := [E(|X|^2)]^{1/2} \). Write \( T \) for the unit circle, the boundary of the unit disc \( D \), parametrised by \( z = e^{i\theta} \); unspecified integrals are over \( T \).

**Theorem 1** (Kolmogorov Isomorphism Theorem). There is a process \( Y \) on \( T \) with orthogonal increments and a probability measure \( \mu \) on \( T \) with

(i) \[
X_n = \int e^{in\theta} dY(\theta);
\]

(ii) \[
E[dY(\theta)^2] = d\mu(\theta).
\]

(iii) The autocorrelation function \( \gamma \) then has the spectral representation

\[
\gamma_n = \int e^{-in\theta} d\mu(\theta).
\]

(iv) One has the Kolmogorov isomorphism between \( \mathcal{H} \) (the time domain) and \( L_2(\mu) \) (the frequency domain) given by

\[
X_t \leftrightarrow e^{it}, \quad (KIT)
\]

for integer \( t \) (as time is discrete).

**Proof.** Parts (i), (ii) are the Cramér representation of 1942 ([Cra], [Do] X.4; Cramér and Leadbetter [CraLea] §7.5). Part (iii), due originally to Herglotz in 1911, follows from (i) and (ii) ([Do] X.4, [BroDav] §4.3). Part (iv) is due to Kolmogorov in 1941 [Kol]. All this rests on Stone’s theorem of 1932, giving the spectral representation of groups of unitary transformations of linear operators on Hilbert space; see [Do] 636-7 for a historical account and references (including work of Khintchine in 1934 in continuous time), Sz.-Nagy [SzN], [DunSch] X.5 for background on spectral theory. 

**Stochastic integration**

The reader will observe the link between the Kolmogorov Isomorphism Theorem and (ii), and its later counterpart from 1944, the Itô Isomorphism Theorem and \((dB_t)^2 = dt\) in stochastic calculus. Indeed, Kallenberg ([Kal], Ch. 13 and Notes to Ch. 13) and Janson ([Jan], 7.4) treat this material from the point of
view of stochastic integration. This viewpoint reveals the equivalence of (ii) and (iii) [Kal]. If we are willing to restrict ourselves to Gaussian processes (which does not restrict the first or second moments), one can work with either the process \( Y \) above with orthogonal increments, or a complex Gaussian stochastic measure \( Z \) with

\[
E[Z(A)\overline{Z(B)}] = \mu(A \cap B)
\]

(the counterpart of (ii) above in this framework). This viewpoint further reveals that the Kolmogorov Isomorphism Theorem is related to the series expansion of Brownian motion (or more general Gaussian processes) with respect to a complete orthonormal system in Hilbert space, with random coefficients. This idea goes back to the Paley-Wiener-Zygmund construction of Brownian motion in 1933 [PaWiZy]; later developments include the Karhunen-Loève expansion ([Kar], [Loe] Ch. XI §37), Kahane’s work on random series of functions [Kah], etc.

To avoid trivialities, we suppose in what follows that \( \mu \) is non-trivial – has infinite support.

Since for integer \( t \) the \( e^{it\theta} \) span polynomials in \( e^{i\theta} \), prediction theory for stationary processes reduces to approximation by polynomials. This is the classical approach to the main result of the subject, Szegő’s theorem (§4 below); see e.g. [GrSz], Ch. 3, [Ach], Addenda, B.

We write

\[
d\mu(\theta) = w(\theta)d\theta/2\pi + d\mu_s(\theta),
\]

so \( w \) is the spectral density (w.r.t. normalized Lebesgue measure) and \( \mu_s \) is the singular part of \( \mu \).

By stationarity,

\[
E[X_m\overline{X}_n] = \gamma_{|m-n|}.
\]

The Toeplitz matrix for \( X \), or \( \mu \), or \( \gamma \), is

\[
\Gamma := (\gamma_{ij}), \quad \text{where} \quad \gamma_{ij} := \gamma_{|i-j|}.
\]

It is positive definite.

For \( n \in N \), write \( \mathcal{H}_{[-n,-1]} \) for the subspace of \( \mathcal{H} \) spanned by \( \{X_{-n}, \ldots, X_{-1}\} \) (the finite past at time 0 of length \( n \)), \( P_{[-n,-1]} \) for projection onto \( \mathcal{H}_{[-n,-1]} \) (thus \( P_{[-n,-1]}X_0 \) is the best linear predictor of \( X_0 \) based on the finite past), \( P_{[-n,-1]}^\perp := I - P_{[-n,-1]} \) for the orthogonal projection (thus \( P_{[-n,-1]}^\perp X_0 := X_0 - P_{[-n,-1]}X_0 \) is the prediction error). We use a similar notation for prediction based on the infinite past. Thus \( \mathcal{H}_{(-\infty,-1]} \) is the closed linear span (cls) of \( X_k \), \( k \leq -1 \), \( P_{(-\infty,-1]} \) is the corresponding projection, and similarly for other time-intervals. Write

\[
\mathcal{H}_n := \mathcal{H}_{(-\infty,n]}
\]

for (the subspace generated by) the past up to time \( n \),

\[
\mathcal{H}_{-\infty} := \bigcap_{n=-\infty}^\infty \mathcal{H}_n
\]

for their intersection, the remote past.
3. Verblunsky’s theorem and partial autocorrelation

With \( \text{corr}(Y, Z) := \frac{E[YZ]}{\sqrt{E[Y^2]E[Z^2]}} \) for \( Y, Z \) zero-mean and not a.s. 0, write

\[ \alpha_n := \text{corr}(X_n - P_{[1,n-1]} X_n, X_0 - P_{[1,n-1]} X_0) \]

for the correlation between the residuals at times 0, \( n \) resulting from (linear) regression on the intermediate values \( X_1, \ldots, X_{n-1} \). The sequence

\[ \alpha = (\alpha_n)_{n=1}^\infty \]

is called the partial autocorrelation function (PACF). It is also called the sequence of Verblunsky coefficients, for reasons which will emerge below.

**Theorem 2 (Verblunsky’s Theorem).** There is a bijection between the sequences \( \alpha = (\alpha_n) \) with each \( \alpha_n \in D \) and the probability measures \( \mu \) on \( T \).

This result dates from Verblunsky in 1936 [V2], in connection with OPUC. It was re-discovered long afterwards by Barndorff-Nielsen and Schou [BarN-S] in 1973 and Ramsey [Ram] in 1974, both in connection with parametrization of time-series models in statistics. The Verblunsky bijection has the great advantage to statisticians of giving an unrestricted parametrization: the only restrictions on the \( \alpha_n \) are the obvious ones resulting from their being correlations – \( |\alpha_n| \leq 1 \), or as \( \mu \) is non-trivial, \( |\alpha_n| < 1 \). By contrast, \( \gamma = (\gamma_n) \) gives a restricted parametrization, in that the possible values of \( \gamma_n \) are restricted by the inequalities of positive-definiteness (principal minors of the Toeplitz matrix \( \Gamma \) are positive). This partly motivates the detailed study of the PACF in, e.g., [In1, In2, In3, InKa1, InKa2]. For general statistical background on partial autocorrelation, see e.g. [KenSt], Ch. 27 (Vol. 2), §46.26-28 (Vol. 3).

As we mentioned in §1, the basic result for OPUC corresponding to Favard’s theorem for OPRL is Szegő’s recurrence (or recursion): given a probability measure \( \mu \) on \( T \), let \( \Phi_n \) be the monic orthogonal polynomials they generate (by Gram-Schmidt orthogonalization). For every polynomial \( Q_n \) of degree \( n \), write

\[ Q_n^*(z) := z^n Q_n(1/z) \]

for the reversed polynomial. Then the Szegő recursion is

\[ \Phi_{n+1}(z) = z\Phi_n(z) - \bar{\alpha}_{n+1} \Phi_n^*(z), \]

where the parameters \( \alpha_n \) lie in \( D \):

\[ |\alpha_n| < 1, \]

and are the Verblunsky coefficients (also known variously as the Szegő, Schur, Geronimus and reflection coefficients; see [Si4], §1.1). The double use of the name Verblunsky coefficients and the notation \( \alpha = (\alpha_n) \) for the PACF and the coefficients is justified: the two coincide. Indeed, the Szegő recursion is known in the time-series literature as the Durbin-Levinson algorithm; see e.g. [BroDav],
The term Verblunsky coefficient is from Simon \cite{Si4}, to which we refer repeatedly. We stress that Simon writes $\alpha_n$ for our $\alpha_{n+1}$, and so has $n = 0, 1, \ldots$ where we have $n = 1, 2, \ldots$. Our notational convention is already established in the time-series literature (see e.g. \cite{BroDav}, §§3.4, 5.2), and is more convenient in our context of the PACF, where $n = 1, 2, \ldots$ has the direct interpretation as a time-lag between past and future (cf. \cite{Si4}, (1.5.15), p. 56-57). See \cite{Si4}, §1.5 and (for two proofs of Verblunsky’s theorem) §1.7, 3.1, and \cite{McLZ} for a recent application of the unrestricted PACF parametrization.

One may partially summarize the distributional aspects of Theorems 1 and 2 by the bijections

$$\alpha \leftrightarrow \mu \leftrightarrow \gamma.$$

The Durbin-Levinson algorithm

Write

$$\hat{X}_{n+1} := \phi_{n1}X_n + \ldots + \phi_{nn}X_1$$

for the best linear predictor of $X_{n+1}$ given $X_n, \ldots, X_1$,

$$v_n := E[(X_{n+1} - \hat{X}_{n+1})^2] = E[(X_{n+1} - P_{[1,n]}X_{n+1})^2]$$

for the mean-square error in the prediction of $X_{n+1}$ based on $X_1, \ldots, X_n$,

$$\phi_n := (\phi_{n1}, \ldots, \phi_{nn})^T$$

(fpc) for the vector of finite-predictor coefficients. The Durbin-Levinson algorithm (\cite{Lev, Dur}; \cite{BroDav} §5.2, \cite{Pou} §7.2) gives the $\phi_{n+1}, v_{n+1}$ recursively, in terms of quantities known at time $n$, as follows:

(i) The first component of $\phi_{n+1}$ is given by

$$\phi_{n+1,n+1} = [\gamma_{n+1} - \sum_{j=1}^{n} \phi_{nj} \gamma_{n-j}] / v_n.$$

The $\phi_{nn}$ are the Verblunsky coefficients $\alpha_n$:

$$\phi_{nn} = \alpha_n.$$

(ii) The remaining components are given by

$$\begin{pmatrix} \phi_{n+1,1} \\ \vdots \\ \phi_{n+1,n} \end{pmatrix} = \begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{nn} \end{pmatrix} - \begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{nn} \end{pmatrix} \begin{pmatrix} \phi_{n1} \\ \vdots \\ \phi_{nn} \end{pmatrix} - \alpha_{n+1} \begin{pmatrix} \phi_{nn} \\ \phi_{n1} \end{pmatrix}.$$

(iii) The prediction errors are given recursively by

$$v_0 = 1, \quad v_{n+1} = v_n[1 - |\phi_{n+1,n+1}|^2] = v_n[1 - |\alpha_{n+1}|^2].$$
In particular, \( v_n > 0 \) and we have from (ii) that
\[
\phi_{nj} - \phi_{n+1,j} = \alpha_{n+1} \phi_{n,n+1-j}. \quad (DL)
\]
Since by (iii)
\[
v_n = \prod_{j=1}^{n} [1 - |\alpha_n|^2],
\]
the \( n \)-step prediction error variance \( v_n \to \sigma^2 > 0 \) iff the infinite product converges, that is, \( \alpha \in \ell_2 \), an important condition that we will meet in §4 below in connection with Szegő’s condition.

Note. 1. The Durbin-Levinson algorithm is related to the Yule-Walker equations of time-series analysis (see e.g. [BroDav], §8.1), but avoids the need there for matrix inversion.

2. The computational complexity of the Durbin-Levinson algorithm grows quadratically, rather than cubically as one might expect; see e.g. Golub and van Loan [GolvL], §4.7. Its good numerical properties result from efficient use of the Toeplitz character of the matrix \( \Gamma \) (or equivalently, of Szegő recursion).

3. See [KatSeTe] for a recent approach to the Durbin-Levinson algorithm, and [Deg] for the multivariate case.

**Stochastic versus non-stochastic**

This paper studies prediction theory for stationary stochastic processes. As an extreme example (in which no prediction is possible), take the ‘free’ case, in which the \( X_n \) are independent (and identically distributed). Then \( d\mu(\theta) = d\theta/2\pi \), \( \gamma_n = \delta_{0n}, \alpha_n \equiv 0 \), \( \Phi_n(z) = z^n \) ([Si4], Ex. 1.6.1).

In contrast to this is the situation where \( X = (X_n) \) is non-stochastic – deterministic, but (typically) chaotic. This case often arises in non-linear time-series analysis and dynamical systems; for a monograph treatment, see Kantz and Schreiber [KanSch].

**4. Weak conditions: Szegő’s theorem**

**Rakhmanov’s Theorem**

One naturally expects that the influence of the distant past decays with increasing lapse of time. So one wants to know when
\[
\alpha_n \to 0 \quad (n \to \infty).
\]
By Rakhmanov’s theorem ([Rak]; [Si5] Ch 9, and Notes to §9.1, [MatNeTo]), this happens if the density \( w \) of the absolutely continuous component \( \mu_a \) is positive on a set of full measure:
\[
|\{ \theta : w(\theta) > 0 \}| = 1
\]
(using normalized Lebesgue measure – or \( 2\pi \) using Lebesgue measure).
Non-determinism and the Wold decomposition

Write $\sigma^2$ for the one-step mean-square prediction error:

$$\sigma^2 := E[(X_0 - P_{(-\infty,-1]}X_0)^2];$$

by stationarity, this is the $\sigma^2 = \lim_{n \to \infty} v_n$ above. Call $X$ non-deterministic (ND) if $\sigma > 0$, deterministic if $\sigma = 0$. (This usage is suggested by the usual one of non-randomness being zero-variance, though here a non-deterministic process may be random, but independent of time, so the stochastic process reduces to a random variable; a deterministic process may be a trigonometric polynomial with random coefficients, so the only randomness is “at $-\infty$”.) The Wold decomposition (von Neumann [vN] in 1929, Wold [Wo] in 1938; see e.g. Doob [Do], XII.4, Hannan [Ha1], Ch. III) expresses a process $X$ as the sum of a non-deterministic process $U$ (the ‘good’ part) and a deterministic process $V$ (the ‘bad’ part):

$$X_n = U_n + V_n;$$

the process $U$ is a moving average,

$$U_n = \sum_{j=-\infty}^{n} m_{n-j} \xi_j = \sum_{k=0}^{\infty} m_k \xi_{n-k},$$

with the $\xi_j$ zero-mean and uncorrelated, with each other and with $V$; $E[\xi_n] = 0$, $\text{var}(\xi_n) = E[\xi_n^2] = \sigma^2$. Thus when $\sigma = 0$ the $\xi_n$ are 0, $U$ is missing and the process is deterministic. When $\sigma > 0$, the spectral measures of $U_n$, $V_n$ are $\mu_{ac}$ and $\mu_s$, the absolutely continuous and singular components of $\mu$ (again, the ‘good’ and ‘bad’ parts). Think of $\xi_n$ as the ‘innovation’ at time $n$ – the new random input, a measure of the unpredictability of the present from the past. This is only present when $\sigma > 0$; when $\sigma = 0$, the present is determined by the past – even by the remote past.

The Wold decomposition arises in operator theory ([vN]; Sz.-Nagy and Foias in 1970 [SzNF], Ch. 1, Rosenblum and Rovnyak in 1985 [RoRo], §1.3, [FrBh] Ch. 1, [Nik2]), as a decomposition into the completely non-unitary (cnu – ‘good’) and unitary (‘bad’) parts.

Szegő’s Theorem

**Theorem 3** (Szegő’s Theorem). (i) $\sigma > 0$ iff $\log w \in L_1$, that is,

$$\int -\log w(\theta) d\theta > -\infty. \quad (Sz)$$

(ii) $\sigma > 0$ iff $\alpha \in \ell_2$.

(iii) $\sigma^2 = \prod_{1}^{\infty} (1 - |\alpha_n|^2)$,

so $\sigma > 0$ iff the product converges, i.e. iff

$$\sum |\alpha_n|^2 < \infty : \quad \alpha \in \ell_2;$$
(iv) $\sigma^2$ is the geometric mean $G(\mu)$ of $\mu$:

$$\sigma^2 = \exp\left(\frac{1}{2\pi} \int \log w(\theta) d\theta\right) = G(\mu) > 0. \tag{K}$$

Proof. Parts (i), (ii) are due to Szegö [Sz2], [Sz3] in 1920-21, with $\mu$ absolutely continuous, and to Verblunsky [V2] in 1936 for general $\mu$. See [Si4] Ch. 2, [Si9] Ch. 2. Parts (iii) and (iv) are due to Kolmogorov in 1941 [Kol]. Thus (K) is called Kolmogorov’s formula. The alternative name for Szegö’s condition (Sz) is the non-determinism condition (ND), above.

We now restrict attention to processes for which Szegö’s condition holds; indeed, we shall move below to stronger conditions.

The original motivation of Szegö, and later Verblunsky, was approximation theory, specifically approximation by polynomials. The Kolmogorov Isomorphism Theorem allows us to pass between finite sections of the past and polynomials; denseness of polynomials allows prediction with zero error (a ‘bad’ situation – determinism), which happens iff (Sz) does not hold. There is a detailed account of the (rather involved) history here in [Si4] §2.3. Other classic contributions include work of Krein in 1945, Levinson in 1947 [Lev], Wiener in 1949 [Wi2]; cf. Achieser ([Ach], Appendix B) in 1956, Grenander and Szegö [GrSz] in 1958. See also [BroDav] §5.8 (where un-normalized Lebesgue measure is used, so there is an extra factor of $2\pi$ on the right of (K)), [Roz] §II.5 from the point of view of time series, [Si4] for OPUC.

Pure non-determinism, (PND)

When the remote past is trivial, $H_{-\infty} := \bigcap_{n=-\infty}^{\infty} H_n = \{0\} \tag{PND}$

there is no deterministic component in the Wold decomposition, and no singular component in the spectral measure. The process is then called purely non-deterministic. Thus

$$PND = \text{(ND)} + (\mu_s = 0) = (Sz) + (\mu_s = 0) = (\sigma > 0) + (\mu_s = 0) \tag{PND}$$

(usage differs here: the term ‘regular’ is used for (PND) in [IbRo], IV.1, but for (ND) in [Do], XII.2).

The Szegö function and Hardy spaces

Szegö’s theorem is the key result in the whole area, and to explore it further we need the Szegö function ($h$, below). For this, we need the language and viewpoint of the theory of Hardy spaces, and some of its standard results; several good textbook accounts are cited in §1. For $0 < p < \infty$, the Hardy space $H_p$ is the class of analytic functions $f$ on $D$ for which

$$\sup_{r<1} \left(\frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^p d\theta\right)^{1/p} < \infty. \tag{H_p}$$
As well as in time series and prediction, as here, Hardy spaces are crucial for martingale theory (see e.g. [Bin1] and the references there). For an entertaining insight into Hardy spaces in probability, see Diaconis [Dia].

For non-deterministic processes, define the Szegö function $h$ by

$$h(z) := \exp\left(\frac{1}{4\pi} \int \frac{e^{i\theta} + z}{e^{i\theta} - z} \log w(\theta) d\theta\right) \quad (z \in D) \quad (OF)$$

(note that in [In1, In2, In3], [InKa1, InKa2], [Roz] II.5 an extra factor $\sqrt{2\pi}$ is used on the right), or equivalently

$$H(z) := h^2(z) = \exp\left(\frac{1}{2\pi} \int \frac{e^{i\theta} + z}{e^{i\theta} - z} \log w(\theta) d\theta\right) \quad (z \in D).$$

Because $\log w \in L^1$ by (Sz), $H$ is an outer function for $H_1$ (whence the name (OF) above); see Duren [Du], §2.4. By Beurling’s canonical factorization theorem,

(i) $H \in H_1$, the Hardy space of order 1 ([Du], §2.4), or as $H = h^2$, $h \in H_2$.

(ii) The radial limit

$$H(e^{i\theta}) := \lim_{r \uparrow 1} H(re^{i\theta})$$

exists a.e., and

$$|H(e^{i\theta})| = |h(e^{i\theta})|^2 = w(\theta)$$

(thus $h$ may be regarded as an ‘analytic square root’ of $w$: indeed, Nikolskii uses the notation $\tilde{h} = [w^{-1/2}]$ – [Nik2], Vol. 1, 380). See also Hoffman [Ho], Ch. 3-5, Rudin [Ru], Ch. 17, Helson [He], Ch. 4.

Kolmogorov’s formula now reads

$$\sigma^2 = m_0^2 = h(0)^2 = G(\mu) = \exp\left(\frac{1}{2\pi} \int \log w(\theta) d\theta\right). \quad (K)$$

When $\sigma > 0$, the Maclaurin coefficients $m = (m_n)$ of the Szegö function $h(z)$ are the moving-average coefficients of the Wold decomposition (recall that the moving-average component does not appear when $\sigma = 0$); see Inoue [In3] and below. When $\sigma > 0$, $m \in \ell_2$ is equivalent to convergence in mean square of the moving-average sum $\sum_{j=0}^{\infty} m_{n-j} \xi_j$ in the Wold decomposition. This is standard theory for orthogonal expansions; see e.g. [Do], IV.4. Note that a function being in $H_2$ and its Maclaurin coefficients being in $\ell_2$ are equivalent by general Hardy-space theory; see e.g. [Ru], 17.10 (see also Th. 17.17 for factorization), [Du] §1.4, 2.4, [Z2], VII.7.

Simon [Si4], §2.8 – ‘Lots of equivalences’ – gives Szegö’s theorem in two parts. One ([Si4] Th. 2.7.14) gives twelve equivalences, the other ([Si4], Th. 2.7.15) gives fifteen; the selection of material is motivated by spectral theory [Si5].

Theorem 3 above extends these lists of equivalences, and treats the material from the point of view of probability theory. (It does not, however, give a condition on
the autocorrelation \( \gamma = (\gamma_n) \) equivalent to \((Sz)\); this is one of the outstanding problems of the area.)

The contrast here with Verblunsky’s theorem is striking. In general, one has unrestricted parametrization: all values \( |\alpha_n| \) are possible, for all \( n \). But under Szegö’s condition, one has \( \alpha \in \ell_2 \), and in particular \( \alpha_n \to 0 \), as in Rakhmanov’s theorem. Thus non-deterministic processes fill out only a tiny part of the \( \alpha \)-parameter space \( D^\infty \). One may regard this as showing that the remote past, trivial under \((Sz)\), has a rich structure in general, as follows:

**Szegö’s alternative (or dichotomy)**

One either has

\[
\log w \in L_1 \text{ and } \mathcal{H}_{-\infty} \neq \mathcal{H}_{-n} \neq \mathcal{H}
\]

or

\[
\log w \notin L_1 \text{ and } \mathcal{H}_{-\infty} = \mathcal{H}_{-n} = \mathcal{H}.
\]

In the former case, \( \alpha \) occupies a tiny part \( \ell_2 \) of \( D^\infty \), and the remote past \( \mathcal{H}_{-\infty} \) is identified with \( L_2(\mu_s) \). This is trivial iff \( \mu_s = 0 \); cf. \((PND)\). In the second case, \( \alpha \) occupies all of \( D^\infty \), and the remote past is the whole space.

Mathematically, Szegö’s dichotomy reflects the Beurling inner-outer factorization above (invariant subspaces correspond to inner functions, by the Beurling-Lax-Halmos theorem – see e.g. [FrBh] §3.1; these are both ‘bad’). It may help the reader’s intuition to mention a physical analogy. Some systems (typically, liquids and gases) are ‘loose’ – left alone, they will thermalize, and tend to an equilibrium in which the details of the past history are forgotten. By contrast, some systems (typically, solids) are ‘tight’: for example, in tempered steel, the thermal history is locked in permanently by the tempering process. Long memory is also important in economics and econometrics; for background here, see e.g. [Rob], [TeKi].

**Note.** 1. Our \( h \) is the Szegö function \( D \) of Simon [Si4], (2.4.2), and \(-1/h \) (see below) its negative reciprocal \(-\Delta \) [Si4], (2.2.92):

\[
h = D, \quad -1/h = -\Delta
\]

(we use both notations to facilitate comparison between [In1, In2, In3], [InKa1, InKa2], which use \( h \), to within the factor \( \sqrt{2\pi} \) mentioned above, and [Si4], our reference on OPUC, which uses \( D \)).

2. Both \( h \) and \(-1/h \) are analytic and non-vanishing in \( D \). See [Si4], Th. 2.2.14 (for \(-1/h \), or \( \Delta \)), Th. 2.4.1 (for \( h \), or \( D \)).

3. That \((Sz)\) implies \( h = D \) is in the unit ball of \( H_2 \) is in [Si4], Th. 2.4.1.

4. See de Branges and Rovnyak [dBR] for general properties of such square-summable power series.

5. Our autocorrelation \( \gamma \) is Simon’s \( c \) (he calls our \( \gamma_n \), or his \( c_n \), the moments of \( \mu \): [Si4], (1.1.20)). Our moving-average coefficients \( m = (m_n) \) have no counterpart in [Si4], and nor do the autoregressive coefficients \( r = (r_n) \) or minimality (see below for these). We will also need the Fourier coefficients of \( \log w \), known for reasons explained below as the *cepstrum*, which we write as
$L = (L_n)$ (‘L for logarithm’: Simon’s $\hat{L}_n$ [Si4], 6.1.13)), and a sequence $b = (b_n)$, the phase coefficients (Fourier coefficients of $\hat{h}/h$).

6. Lund et al. [LuZhKi] give several properties – monotonicity, convexity etc. – which one of $m$, $\gamma$ has iff the other has.

*MA*(∞) and *AR*(∞)

The power series expansion

$$h(z) = \sum_{n=0}^{\infty} m_n z^n \quad (z \in D)$$

generates the *MA*(∞) coefficients $m = (m_n)$ in the Wold decomposition. That of

$$-1/h(z) = \sum_{n=0}^{\infty} r_n z^n \quad (z \in D)$$

generates the *AR*(∞) coefficients $r = (r_n)$ in the (infinite-order) autoregression

$$\sum_{j=-\infty}^{n} r_{n-j} X_j + \xi_n = 0 \quad (n \in Z). \quad (AR)$$

See [InKa2] §2, [In3] for background.

One may thus extend the above list of bijections as follows:

Under (Sz), $\alpha, \mu, \gamma \leftrightarrow m = (m_n) \leftrightarrow h, -1/h \leftrightarrow r = (r_n)$.  

**Finite and infinite predictor coefficients**

We met the $n$-vector $\phi_n$ of finite-predictor coefficients in (fpc) of §3; we can extend it to an infinite vector, still denoted $\phi_n$, by adding zeros. The corresponding vector $\phi := (\phi_1, \phi_2, \ldots)$ of infinite-predictor coefficients gives the infinite predictor

$$P_{(-\infty,-1]} X_0 = \sum_{j=1}^{\infty} \phi_j X_{-j} \quad (ipc)$$

([InKa2], (1.4)). One would expect convergence of finite-predictor to infinite-predictor coefficients; under Szegö’s condition, one has such convergence in $\ell_2$ iff (PND), i.e., $\mu_s = 0$:

$$\phi_n \to \phi \quad \text{in} \quad \ell_2 \iff (PND)$$

(Pourahmadi [Pou], Th. 7.14).

**The Szegö limit theorem**

With $G(\mu)$ as above, write $T_n$ (or $T_n(\gamma)$, or $T_n(\mu)$) for the $n \times n$ Toeplitz matrix $\Gamma^{(n)}$ with elements

$$\Gamma^{(n)}_{ij} := c_{j-i}$$
obtained by truncation of the Toeplitz matrix \( \Gamma \) (cf. [BotSi2]). Szegő’s limit theorem states that, under \((Sz)\), its determinant satisfies
\[
\frac{1}{n} \log \det T_n \rightarrow G(\mu) \quad (n \rightarrow \infty)
\]
(note that \((Sz)\) is needed for the right to be defined). A stronger statement – Szegő’s strong limit theorem – holds; we defer this till §6.

The Szegő limit theorem is used in the Whittle estimator of time-series analysis; see e.g. Whittle [Wh], Hannan [Ha2].

**Phase coefficients**

When the Szegő condition \((Sz)\) holds, the Szegő function \( h(z) = \sum_{n=0}^{\infty} m_n z^n \) is defined. We can then define the phase function \( \bar{h}/h \), so called because it has unit modulus and depends only on the phase or argument of \( h \) (Peller [Pel], §8.5). Its Fourier coefficients \( b_n \) are called the phase coefficients. They are given in terms of \( m = (m_n) \) and \( r = (r_n) \) by
\[
b_n := \sum_{k=0}^{\infty} m_k r_{n+k} \quad (n = 0, 1, 2, \ldots). \tag{b}
\]

The role of the phase coefficients is developed in [BinIK]. They are important in connection with rigidity (§7.4 below), and Hankel operators [Pel].

**Rajchman measures**

In the Gaussian case, mixing in the sense of ergodic theory holds iff \( \gamma_n \rightarrow 0 \quad (n \rightarrow \infty) \) ([CorFoSi], §14.2, Th. 2). Measures for which this condition holds are called Rajchman measures (they were studied by A. Rajchman in the 1920s). Here the continuous singular part \( \mu_{cs} \) of \( \mu \) is decisive; for a characterization of Rajchman measures, see Lyons ([Ly1] – [Ly3] and the appendix to [KahSa]).

**ARMA\((p, q)\)**

The Box-Jenkins ARMA\((p, q)\) methodology ([BoxJeRe], [BroDav]: autoregressive of order \( p \), moving average of order \( q \)) applies to stationary time series where the roots of the relevant polynomials lie in the unit disk (see e.g. [BroDav] §3.1). The limiting case, of unit roots, involves non-stationarity, and so the statistical dangers of spurious regression (§1); cf. Robinson [Rob], p.2. We shall meet other instances of unit-root phenomena later (§7.2).

**Szegő’s theorem and the Gibbs Variational Principle**

We point out that Verblunsky [V2] proved the Gibbs Variational Principle, one of the cornerstones of nineteenth-century statistical mechanics, for the Szegő integral:
\[
\inf_{\theta} \left[ \int e^\theta \, d\mu / \exp \left( \int g \, d\theta / 2\pi \right) \right] = \exp \left[ \int \log w(\theta) \, d\theta / 2\pi \right].
\]
For details, see e.g. Simon [Si9] §§2.2, 10.6, [Si10], Ch. 16, 17. For background on the Gibbs Variational Principle, see e.g. Simon [Si1], III.4, Georgii [Geo], 15.4, Ellis [Ell], III.8.

5. Strong conditions: Baxter’s theorem

The next result ([Bax1, Bax2, Bax3]; [Si4], Ch. 5) gives the first of our strong conditions.

**Theorem 4** (Baxter’s theorem). The following are equivalent:

(i) the Verblunsky coefficients (or PACF) are summable,

\[ \alpha \in \ell_1; \quad (B) \]

(ii) the autocorrelations are summable, \( \gamma \in \ell_1 \), and \( \mu \) is absolutely continuous with continuous positive density:

\[ \min_\theta w(\theta) > 0. \]

Of course, \( (\gamma_n) \) summable gives, as the \( \gamma_n \) are the Fourier coefficients of \( \mu \), that \( \mu \) is absolutely continuous with continuous density \( w \); thus \( w > 0 \) iff \( \inf w = \min w > 0 \). We extend this list of equivalences, and bring out its probabilistic significance, in Theorem 5 below on \( \ell_1 \) (this is substantially Theorem 4.1 of [In3]). We call \( \alpha \in \ell_1 \) (or any of the other equivalences in Theorem 4) Baxter’s condition (whence \( (B) \) above). Since \( \ell_1 \subset \ell_2 \), Baxter’s condition \( (B) \) (‘strong’) implies Szegö’s condition \( (S_z) \) (‘weak’).

**Theorem 5** (Inoue). For a stationary process \( X \), the following are equivalent:

(i) Baxter’s condition \( (B) \) holds: \( \alpha \in \ell_1 \).

(ii) \( \gamma \in \ell_1 \), \( \mu_s = 0 \) and the spectral density \( w \) is continuous and positive.

(iii) \( (PND) \) (that is, \( (S_z)/(ND) + \mu_s = 0 \)) holds, and the moving-average and autoregressive coefficients are summable:

\[ m \in \ell_1, \quad r \in \ell_1. \]

(iv) \( m \in \ell_1 \), \( \mu_s = 0 \) and the spectral density \( w \) is continuous and positive.

(v) \( r \in \ell_1 \), \( \mu_s = 0 \) and the spectral density \( w \) is continuous and positive.

**Proof.** (i) \( \Leftrightarrow \) (ii). This is Baxter’s theorem, as above.

(iii) \( \Rightarrow \) (iv), (v). By \( (PND), (S_z) \) holds, so the non-tangential limit

\[ h(e^{i\theta}) = \lim_{r \uparrow 1} h(re^{i\theta}) = \sum_{n=0}^{\infty} m_n e^{in\theta} \]

exists a.e. But as \( m \in \ell_1 \), \( h(e^{i\theta}) \) is continuous, so this holds everywhere. Since

\[ w(\theta) = |h(e^{i\theta})|^2 = |D(e^{i\theta})|^2 = \sum_{n=0}^{\infty} |m_n e^{in\theta}|^2, \]
w is continuous. Letting \( r \uparrow 1 \) in
\[
h(z)(-1/h(z)) = \left( \sum_{n=0}^{\infty} m_n r^n e^{in\theta} \right) \left( \sum_{n=0}^{\infty} r_n r^n e^{in\theta} \right) = -1
\]
gives similarly
\[
\sum_{n=0}^{\infty} m_n e^{in\theta} \left( \sum_{n=0}^{\infty} r_n e^{in\theta} \right) = -1.
\]
So \( h(e^{i\theta}) \) has no zeros, so neither does \( w \). That is, (iv), (v) hold.

(iv) \( \Rightarrow \) (iii). As \( w \) is positive and continuous, \( w \) is bounded away from 0 and \( \infty \). So \( 1/w \) is also. So
\[
1/w(\theta) = |1/h(e^{i\theta})|^2 = |\Delta(e^{i\theta})|^2 = \left| \sum_{n=0}^{\infty} r_n e^{in\theta} \right|^2,
\]
where \( \Delta = 1/D \). (See [Si4], Th. 2.2.14, 2.7.15: the condition \( \lambda_\infty(.) > 0 \) there is \((Sz)__\), so holds here.) By Wiener’s theorem, the reciprocal of a non-vanishing absolutely convergent Fourier series is an absolutely convergent Fourier series (see e.g. [Ru], Th. 18.21). So from \( m \in \ell_1 \) we obtain \( r \in \ell_1 \), whence (iii) (cf. [Berk], p.493).

(v) \( \Rightarrow \) (iii). This follows as above, by Wiener’s theorem again.

(iv) \( \Rightarrow \) (ii). From the MA\( (\infty) \) representation,
\[
\gamma_n = \sum_{k=0}^{\infty} m_{|n|+k} m_k \quad (n \in \mathbb{Z}) \quad \text{(conv)}
\]
([InKa2], (2.21)). So as \( \ell_1 \) is closed under convolution, \( m \in \ell_1 \) implies \( \gamma \in \ell_1 \), indeed with
\[
\|\gamma\|_1 \leq \|m\|_1^2,
\]
giving (ii).

(ii) \( \Rightarrow \) (v). We have
\[
\phi_j = c_0 r_j = \sigma r_j
\]
with \( \phi_j \) the infinite-predictor coefficients ([InKa2], (3.1)). Then \( r \in \ell_1 \) follows by the Wiener-Lévy theorem, as in Baxter [Bax3], 139-140.

Note. 1. Under Baxter’s condition, both \( |h| \) and \( |1/h| \) (or \( |D| \) and \( |\Delta| = |1/D| \)) are continuous and positive on the unit circle. As \( h, 1/h \) are analytic in the disk, and so attain their maximum modulus on the circle by the maximum principle,
\[
\inf_D |h(.)| > 0, \quad \inf_D |1/h(.)| > 0
\]
(and similarly for \( D(.), \Delta \): [Si4], (5.2.3), (5.2.4)).

2. The hard part of Baxter’s theorem is (ii) \( \Rightarrow \) (i), as Simon points out ([Si4], 314).
3. Simon [Si4], Th. 5.2.2 gives twelve equivalences in his final form of Baxter’s theorem. (He does not, however, deal explicitly with \( m \) and \( r \).)

4. Simon also gives a more general form, in terms of Beurling weights, \( \nu \). The relevant Banach algebras contain the Wiener algebra used above as the special case \( \nu = 1 \).

5. The approach of [Si4], §5.1 is via truncated Toeplitz matrices and their inverses. The method derives, through Baxter’s work, from the Wiener-Hopf technique. This point of view is developed at length in [BotSi1, BotSi2]. Baxter’s motivation was approximation to infinite-past predictors by finite-past predictors.

**Long-range dependence**

In various physical models, the property of long-range dependence (LRD) is important, particularly in connection with phase transitions (see e.g. [Si1], Ch. II, [Gri1], Ch. 9, [Gri2], Ch. 5), to which we return below. This is a spatial property, but applies also in time rather than space, when the term used is long memory. A good survey of long-memory processes was given by Cox [Cox] in 1984, and a monograph treatment by Beran [Ber] in 1994. For more recent work, see [DouOpTa, Rob], [Gao] Ch. 6, [TeKi, GiKoSu].

Baxter’s theorem is relevant to the definition of LRD recently proposed independently by Debowski [Deb] and Inoue [In3]: long-range dependence, or long memory, is non-summability of the PACF:

\[ X \text{ has LRD iff } \alpha \notin \ell_1. \] \hfill (DI)

While the broad concept of long memory, or LRD, has long been widely accepted, authors differed over the precise definition. There were two leading candidates:

(i) LRD is non-summability of covariances, \( \gamma \notin \ell_1 \).

(ii) LRD is covariance decaying like a power: \( \gamma_n \sim c/n^{1-2d} \) as \( n \to \infty \), for some parameter \( d \in (0, 1/2) \) (\( d \) for differencing – see below) and constant \( c \in (0, \infty) \) (and so \( \sum \gamma_n = \infty \)).

**Note.** 1. In place of (ii), one may require \( w(\theta) \sim C/\theta^{2d} \) as \( \theta \downarrow 0 \), for some constant \( C \in (0, \infty) \). The constants here may be replaced by slowly varying functions. See e.g. [BinGT] §4.10 for relations between regular variation of Fourier series and Fourier coefficients.

2. One often encounters, instead of \( d \in (0, 1/2) \), a parameter \( H = d + \frac{1}{2} \in (1/2, 1) \). This \( H \) is the Hurst parameter, named after the classic studies by the hydrologist Hurst of water flows in the Nile; see [Ber], Ch. 2.

3. For \( d \in (0, \frac{1}{2}) \), \( \ell(.) \) slowly varying, the following class of prototypical long-memory examples is considered in [InKa2], §2.3 (see also [In1], Th. 5.1):

\[
\begin{align*}
\gamma_n & \sim \ell(n)^2 B(d, 1-2d)/n^{1-2d}, \\
m_n & \sim \ell(n)/n^{1-d},
\end{align*}
\]
\[ r_n \sim \frac{d \sin(\pi d)}{\pi} \frac{1}{\ell(n)^{1+d}}. \]

See the sources cited for inter-relationships between these.

4. In [InKa2], Example 2.6, the class of FARIMA\((p,d,q)\) processes is considered (obtained from an ARIMA\((p,q)\) process by fractional differencing of order \(d\) – see [Hos, BroDav, KokTa]). For \(d \in (0,1/2)\) these have long memory; for \(d = 0\) they reduce to the familiar ARMA\((p,q)\) processes.

Li ([Li], §3.4) has recently given a related but different definition of long memory; we return to this in §5 below.

6. **Strong conditions: The strong Szegő theorem**

The work of this section may be motivated by work from two areas of physics.

1. **The cepstrum**

During the Cold War, the problem of determining the signature of the underground explosion in a nuclear weapon test, and distinguishing it from that of an earthquake, was very important, and was studied by the American statistician J. W. Tukey and collaborators. Write \(L = (L_n)\), where the \(L_n\) are the Fourier coefficients of log \(w\), the log spectral density:

\[ L_n := \int \log w(\theta)e^{i\theta}d\theta/2\pi. \]

Thus \(\exp(L_0)\) is the geometric mean \(G(\mu)\). The sequence \(L\) is called the cepstrum. \(L_n\) the cepstral coefficients (Simon’s notation here is \(\hat{L}_n\); [Si4], (2.1.14), (6.1.11)); see e.g. [OpSc], Ch. 12. The terminology dates from work of Bogert, Healy and Tukey of 1963 on echo detection [BogHeTu]; see McCullagh [McC], Brillinger [Bri] (the term is chosen to suggest both echo and spectrum, by reversing the first half of the word spectrum; it is accordingly pronounced with the c hard, like a k).

2. **The strong Szegő limit theorem**

This (which gives the weak form on taking logarithms) states (in its present form, due to Ibragimov) that

\[ \det \frac{T_n}{G(\mu)^n} \to E(\mu) := \exp\left\{ \sum_{k} \infty kL_k^2 \right\} \quad (n \to \infty) \]

(of course the sum here must converge; it turns out that this form is best-possible; the result is valid whenever it makes sense ([Si4], 337). The motivation was Onsager’s work in the two-dimensional Ising model, and in particular Onsager’s formula, giving the existence of a critical temperature \(T_c\) and the decay of the magnetization as the temperature \(T \uparrow T_c\); see [BotSi2] §5.1, [Si1] II.6, [McCW]. The mechanism was a question by Onsager (c. 1950) to his Yale colleague Kakutani, who asked Szegő ([Si4], 331).
Write $H^{1/2}$ for the subspace of $\ell_2$ of sequences $a = (a_n)$ with

$$\|a\|^2 := \sum_n (1 + |n|)|a_n|^2 < \infty \quad (H^{1/2})$$

(the function of the ‘1’ on the right is to give a norm; without it, $\|\cdot\|$ vanishes on the constant functions). This is a Sobolev space ([Si4], 329, 337; it is also a Besov space, whence the alternative notation $B_2^{1/2}$; see e.g. Peller [Pel], Appendix 2.6 and §7.13). This is the space that plays the role here of $\ell_2$ in §4 and $\ell_1$ in §5. Note first that, although $\ell_1$ and $H^{1/2}$ are close in that a sequence $(a_n)$ of powers belongs to both or neither, neither contains the other (consider $a_n = 1/(n \log n)$, $a_n = 1/\sqrt{n}$ if $n = 2^k$, 0 otherwise).

**Theorem 6 (Strong Szegö Theorem).**

(i) If $(PND)$ holds (i.e. $(Sz) = (ND)$ holds and $\mu_s = 0$), then

$$E(\mu) = \prod_{j=1}^{\infty} (1 - |\alpha_j|^2)^{-1} = \exp \left( \sum_{n=1}^{\infty} nL_n^2 \right)$$

(all three may be infinite), with the infinite product converging iff the strong Szegö condition

$$\alpha \in H^{1/2}, \quad sSz$$

holds.

(ii) $(sSz)$ holds iff

$$L \in H^{1/2} \quad (sSz')$$

holds.

(iii) Under $(Sz)$, finiteness of any (all three) of the expressions in (i) forces $\mu_s = 0$.

**Proof.** Part (i) is due to Ibragimov ([Si4], Th. 6.1.1), and (ii) is immediate from this. Part (iii) is due to Golinski and Ibragimov ([Si4], Th. 6.1.2; cf. [Si2]).

Part of Ibragimov’s theorem was recently obtained independently by Li [Li], under the term reflectrum identity (so called because it links the Verblunsky or reflection coefficients with the cepstrum), based on information theory – mutual information between past and future. Earlier, Li and Xie [LiXi] had shown the following:

(i) a process with given autocorrelations $\gamma_0, \ldots, \gamma_p$ with minimal information between past and future must be an autoregressive model $AR(p)$ of order $p$;

(ii) a process with given cepstral coefficients $L_0, \ldots, L_p$ with minimal information between past and future must be a Bloomfield model $BL(p)$ of order $p$ ([Bl1, Bl2]), that is, one with spectral density $w(\theta) = \exp\{L_0 + 2 \sum_{k=1}^{p} L_k \cos k\theta\}$.

Another approach to the strong Szegö limit theorem, due to Kac [Kac], uses the conditions

$$\inf w(.) > 0, \quad \gamma = (\gamma_n) \in \ell_1, \quad \gamma \in H^{1/2}$$
(recall that $\ell_1$ and $H^{1/2}$ are not comparable). This proof, from 1954, is linked to probability theory – Spitzer’s identity of 1956, and hence to fluctuation theory for random walks, for which see e.g. [Ch], Ch. 8.

The Borodin-Okounkov formula

This turns the strong Szegő limit theorem above from analysis to algebra by identifying the quotient on the left there as a determinant which visibly tends to 1 as $n \to \infty$ [BorOk]; see [Si4] §6.2. (It was published in 2000, having been previously obtained by Geronimo and Case [GerCa] in 1979; see [Si4] 337, 344, [Bot] for background here.) In terms of operator theory and in Widom’s notation [Bot], the result is

$$\frac{\det T_n(a)}{G(a)^n} = \frac{\det(I - Q_n H(b) H(\tilde{c}) Q_n)}{\det(I - H(b) H(\tilde{c}))},$$

for $a$ a sufficiently smooth function without zeros on the unit circle and with winding number 0. Then $a$ has a Wiener-Hopf factorization $a = a_- a_+; b := a_- a_+^{-1}; c := a_-^{-1} a_+; H(b), H(\tilde{c})$ are the Hankel matrices $H(b) = (b_{j+k})_{j,k=0}^\infty, H(\tilde{c}) = (c_{-j-k})_{j,k=0}^\infty$, and $Q_n$ is the orthogonal projection of $\ell^2(\{n, n+1, \ldots\})$ onto $\ell^2(\{n, n+1, \ldots\})$. By Widom’s formula,

$$1/\det(I - H(b) H(\tilde{c})) = \exp\{\sum_1^\infty k L_k^2\} =: E(a)$$

(see e.g. [Si4], Th. 6.2.13), and $Q_n H(b) H(\tilde{c}) Q_n \to 0$ in the trace norm, whence

$$\det T_n(a)/G(a)^n \to E(a),$$

the strong Szegő limit theorem. See [Si4], Ch. 6, [Si6, BasW, BotW] (in [Si4] §6.2 the result is given in OPUC terms; here $b, c$ are the phase function $h/\tilde{h}$ and its inverse).

$(B + sS\bar{z})$

We may have both of the strong conditions $(B)$ and $(sS\bar{z})$ (as happens in Kac’s method [Kac], for instance). Matters then simplify, since the spectral density $w$ is now continuous and positive. So $w$ is bounded away from 0 and $\infty$, so log $w$ is bounded. Write

$$\omega^2(\delta, h) := \sup_{|\theta| \leq \delta} \left( \int |h(\lambda + \theta) - h(\lambda)|^2 d\lambda \right)^{1/2}$$

for the $L_2$ modulus of continuity. Applying [IbRo], IV.A, Lemma 7 to log $w$,

$$L \in H^{1/2} \iff \sum_1^\infty \omega^2(1/k, \log w) < \infty,$$
and applying it to $w$,

$$
\gamma \in H^{1/2} \iff \sum_{k=1}^{\infty} \omega^2(1/k, w) < \infty.
$$

Thus under (B), $L \in H^{1/2}$ and $\gamma \in H^{1/2}$ become equivalent. This last condition is Li’s proposed definition of long-range dependence:

$$
\text{LRD} \iff \gamma \notin H^{1/2}
$$

([Li], §3.4; compare the Debowksi-Inoue definition (DI) above, that LRD iff $\alpha \notin \ell_1$).

We are now in $W \cap H^{1/2}$, the intersection of $H^{1/2}$ with the Wiener algebra $W$ (of absolutely convergent Fourier series) relevant to Baxter’s theorem as in §5. As there, we can take inverses, since the Szegö function is non-zero on the circle (cf. [BotSi2], §5.1). One can thus extend Theorem 4 to this situation, including the cepstral condition $L \in H^{1/2}$ (Li [Li], Th. 1 part 3, showed that $L \in H^{1/2}$ and $\gamma \in H^{1/2}$ are equivalent if $w$ is continuous and positive).

$L_\infty + (sSz)$

The bounded functions in $H^{1/2}$ form an algebra, the Krein algebra $K$, a Banach algebra under convolution; see Krein [Kr], Böttcher and Silbermann [BotSi1] Ch. 10, [BotSi2] Ch. 5, [Si4], 344, [BotKaSi]. The Krein algebra may be used as a partial substitute for the Wiener algebra $W$ used to treat Baxter’s theorem in §5 ($W \cap H^{1/2}$ is also an algebra: [BotSi2], §5.1).

6.1. $\phi$-mixing

Weak dependence may be studied by a hierarchy of mixing conditions; for background, see e.g. Bradley [Bra1, Bra2, Bra3], Bloomfield [B3], Ibragimov and Linik [IbLi], Ch. 17, Cornfeld et al. [CorFoSi]), and in the Gaussian case Ibragimov and Rozanov [IbRo], Peller [Pel]. We need two sequences of mixing coefficients:

$$
\phi(n) := E \sup\{|P(A|F_{-\infty}) - P(A)| : A \in \mathcal{F}_{\infty}\};
$$

$$
\rho(n) := \rho(\mathcal{F}^{-\infty}, \mathcal{F}_n),
$$

where

$$
\rho(A, B) := \sup\{|E(f|A) - Ef\|_2 : f : f \in L_2(A)\}.
$$

The process is called $\phi$-mixing if $\phi(n) \to 0$ as $n \to \infty$, $\rho$-mixing if $\rho(n) \to 0$. (The reader is warned that some authors use other letters here – e.g. [IbRo] uses $\beta$ for our $\phi$; we follow Bradley.)

We quote [Bra1] that $\phi$-mixing implies $\rho$-mixing. We regard the first as a strong condition, so include it here, but the second and its several weaker relatives as intermediate conditions, which we deal with in §7 below.
The spectral characterization for $\phi$-mixing is
\[
\mu_s = 0, \quad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),
\]
where $P$ is a polynomial with its roots on the unit circle and the cepstrum $L^* = (L^*_n)$ of $w^*$ satisfies the strong Szegő condition ($sSz$) ([IbRo] IV.4, p. 129). This is weaker than ($sSz$). In the Gaussian case, $\phi$-mixing (also known as absolute regularity) can also be characterized in operator-theoretic terms: $\phi(n)$ can be identified as $\sqrt{\text{tr}(B_n)}$, where $B_n$ are compact operators with finite trace, so $\phi$-mixing is $\text{tr}(B_n) \to 0$ ([IbRo], IV.2 Th. 4, IV.3 Th. 6).

7. Intermediate conditions

We turn now to four intermediate conditions, in decreasing order of strength.

7.1. $\rho$-mixing

The spectral characterization of $\rho$-mixing (also known as complete regularity – see §6.1 above) is
\[
\mu_s = 0, \quad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),
\]
where $P$ is a polynomial with its roots on the unit circle and
\[
\log w^* = u + \tilde{v},
\]
with $u, v$ real and continuous (Sarason [Sa2]; Helson and Sarason [HeSa]). An alternative spectral characterization is
\[
\mu_s = 0, \quad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),
\]
where $P$ is a polynomial with its roots on the unit circle and for all $\epsilon > 0$,
\[
\log w^* = r_\epsilon + u_\epsilon + \tilde{v}_\epsilon,
\]
where $r_\epsilon$ is continuous, $u_\epsilon, v_\epsilon$ are real and bounded, and $\|u_\epsilon\| + \|v_\epsilon\| < \epsilon$ ([IbRo], V.2 Th. 3; we note here that inserting such a polynomial factor preserves complete regularity, merely changing $\rho$ – [IbRo] V.1, Th. 1).

7.2. Positive angle: The Helson-Szegő and Helson-Sarason conditions

We turn now to a weaker condition. For subspaces $A, B$ of $\mathcal{H}$, the angle between $A$ and $B$ is defined as
\[
\cos^{-1} \sup \{ |(a, b)| : a \in A, b \in B, \|a\|, \|b\| = 1 \}.
\]
Then $A, B$ are at a positive angle iff this supremum is $< 1$. One says that the process $X$ satisfies the positive angle condition, $(PA)$, if for some time lapse $k$
the past $\text{cls}(X_m : m < 0)$ and the future $\text{cls}(X_{k+m} : m \geq 0)$ are at a positive angle, i.e. $\rho(0) = \ldots \rho(k-1) = 1$, $\rho(k) < 1$, which we write as $PA(k)$ (Helson and Szegö [HeSz], $k = 1$; Helson and Sarason [HeSa], $k > 1$). The spectral characterization of this is

$$\mu_s = 0, \quad w(\theta) = |P(e^{i\theta})|^2 w^*(\theta),$$

where $P$ is a polynomial of degree $k - 1$ with its roots on the unit circle and

$$\log w^* = u + \tilde{v},$$

where $u, v$ are real and bounded and $\|v\| < \pi/2$ ([IbRo] V.2, Th. 3, Th. 4). (The role of $\pi/2$ here stems from Zygmund’s theorem of 1929, that if $u$ is bounded and $\|u\| < \pi/2$, $\exp\{\tilde{u}\} \in L_1$; see e.g. [Z1], [Z2] VII, (2.11), [Tor], V.3: cf. [Pel] §3.2.) Thus $\rho$-mixing implies ($PA$) (i.e. $PA(k)$ for some $k$).

The case $PA(k)$ for $k > 1$ is a unit-root phenomenon (cf. the note on ARMA$(p, q)$ in §4). We may (with some loss of information) reduce to the case $PA(1)$ by sampling only at every $k$th time point (cf. [Pel], §§8.5, 12.8). We shall do this for convenience in what follows.

It turns out that the Helson-Szegö condition ($PA(1)$) coincides with Muckenhoupt’s condition $A_2$ in analysis:

$$\sup_I \left( \frac{1}{|I|} \int_I w(\theta) d\theta \right) \left( \frac{1}{|I|} \int_I \frac{1}{w(\theta)} d\theta \right) < \infty,$$

where $|.|$ is Lebesgue measure and the supremum is taken over all subintervals $I$ of the unit circle $T$. See e.g. Hunt, Muckenhoupt and Wheeden [HuMuWe]. With the above reduction of $PA$ to $PA(1)$, we then have $\rho$-mixing implies $PA(1)$ (= $A_2$).

### 7.3. Pure minimality

Consider now the interpolation problem, of finding the best linear interpolation of a missing value, $X_0$ say, from the others. Write

$$H'_n := \text{cls}\{X_m : m \neq n\}$$

for the closed linear span of the values at times other than $n$. Call $X$ minimal if

$$X_n \notin H'_n,$$

purely minimal if

$$\bigcap_n H'_n = \{0\}.$$

The spectral condition for minimality is (Kolmogorov in 1941, [Kol] §10)

$$1/w \in L_1,$$

(min)
and for pure minimality, $\mu_s = 0$ also (Makagon-Weron in 1976, [MakWe]; Sarason in 1978, [Sa1]; [Pou], Th. 8.10):

$$1/w \in L_1, \quad \mu_s = 0.$$  \hspace{1cm} \text{(purmin)}

Of course $(A_2)$ implies $1/w \in L_1$, so the Helson-Szegö condition $(PA(1))$ (or Muckenhoupt condition $(A_2)$) implies pure minimality. (From $\log x < x - 1$ for $x > 1$, $(min)$ implies $(Sz)$: both restrict the small values of $w \geq 0$, and in particular force $w > 0$ a.e.) For background on the implication from the Helson-Szegö condition $PA(1)$ to $(A_2)$, see e.g. Garnett [Gar], Notes to Ch. VI, Treil and Volberg [TrVo2].

Under minimality, the relationship between the moving-average coefficients $m = (m_n)$ and the autoregressive coefficients $r = (r_n)$ becomes symmetrical, and one has the following complement to Theorem 4:

**Theorem 7** (Inoue). For a stationary process $X$, the following are equivalent:

(i) The process is minimal.

(ii) The autoregressive coefficients $r = (r_n)$ in $(AR)$ satisfy $r \in \ell_2$.

(iii) $1/h \in H_2$.

**Proof.** Since

$$1/h(z) = \exp\left(\frac{1}{4\pi} \int \left(\frac{e^{i\theta} + z}{e^{i\theta} - z}\right) \log(1/w(\theta)) d\theta\right) \quad (z \in D), \quad \text{(OF')}

and $\pm \log w$ are in $L_1$ together, when $1/w \in L_1$ (i.e. the process is minimal) one can handle $1/w, 1/h, m = (m_n)$ as we handled $w, h$ and $r = (r_n)$, giving

$$1/h \in H_2$$

and

$$r = (r_n) \in \ell_2.$$

Conversely, each of these is equivalent to $(min)$; [In1], Prop. 4.2. \qed

### 7.4. Rigidity; (LM), (CND), (IPF)

**Rigidity; the Levinson-McKean condition**

Call $g \in H^1$ rigid if is determined by its phase or argument:

$$f \in H^1 \quad (f \not\equiv 0), \quad f/|f| = g/|g| \quad \text{a.e.} \quad \Rightarrow \quad f = cg$$

for some positive constant $c$.

This terminology is due to Sarason [Sa1, Sa2]; the alternative terminology, due to Nakazi, is *strongly outer* [Nak1, Nak2]. One could instead say that such a function is determined by its phase. The idea originates with de Leeuw and Rudin [dLR] and Levinson and McKean [LevMcK]. In view of this, we call the condition that $\mu$ be absolutely continuous with spectral density $w = |h|^2$ with $h^2$ rigid, or determined by its phase, the *Levinson-McKean condition*, $(LM)$. 

Complete non-determinism; intersection of past and future
In [InKa2], the following two conditions are discussed:

(i) complete non-determinism,
\[ \mathcal{H}_{(-\infty,-1]} \cap \mathcal{H}_{[0,\infty)} = \{0\} \]  
(CND)

(for background on this, see [BlJeHa, JeBl, JeBlBa]),

(ii) the intersection of past and future property,
\[ \mathcal{H}_{(-\infty,-1]} \cap \mathcal{H}_{[-n,\infty)} = \mathcal{H}_{[-n,-1]} \]  
\[ (n = 1, 2, \ldots) \]  
(IPF)

These are shown to be equivalent in [InKa2]. In [KasBi], it is shown that both
are equivalent to the Levinson-McKean condition, or rigidity:
\[ (LM) \Leftrightarrow (IPF) \Leftrightarrow (CND) \]

These are weaker than pure minimality ([Bl3], §7, [KasBi]). But since (CND)
was already known to be equivalent to (PND) + (IPF), they are stronger than
(PND). This takes us from the weakest of the four intermediate conditions of
this section to the stronger of the weak conditions of §4.

8. Remarks

1. \(VMO \subset BMO\)

The spectral characterizations given above were mainly obtained before the
work of Fefferman [Fe] in 1971, Fefferman and Stein [FeSt] in 1972 (see Garnett
[Gar], Ch. VI for a textbook account): in particular, they predate the Fefferman-
Stein decomposition of a function of bounded mean oscillation, \(f \in BMO\), as
\[ f = u + \tilde{v}, \quad u, v \in L_\infty. \]

This has a complement due to Sarason [Sa3], where \(f\) here is in \(VMO\) iff \(u, v\) are continuous. Sarason also gives ([Sa3], Th. 2) a characterization of his class of functions of vanishing mean oscillation \(VMO\) within \(BMO\) related to
Muckenhoupt’s condition \((A_2)\).

While both components \(u, v\) are needed here, and may be large in norm, it
is important to note that the burden of being large in norm may be born by a
continuous function, leaving \(u\) and \(\tilde{v}\) together to be small in \((L_\infty)\) norm (in
particular, less than \(\pi/2\)). This is the Ibragimov-Rozanov result ([IbRo], V.2
Th. 3), used to show that absolute regularity \((\phi\text{-mixing}, \S 6.1)\) implies complete
regularity \((\rho\text{-mixing}, \S 7.1)\).

2. \(H^{1/2} \subset VMO\)

The class \(H^{1/2}\) is contained densely within \(VMO\) (Prop. A2, Boutet de
Monvel-Berthier et al. [BouGePu]). For \(H^{1/2}\), one has a version of the Fefferman-
Stein decomposition for BMO:
\[ f \in H^{1/2} \iff f = u + \tilde{v}, \quad u, v \in H^{1/2} \cap L_\infty \]
([Pel] §7.13).
3. **Winding number and index**

   The class $H^{1/2}$ occurs in recent work on topological degree and winding number; see Brezis [Bre], Bourgain and Kozma [BouKo]. The winding number also occurs in operator theory as an index in applications of Banach-algebra methods and the Gelfand transform; see e.g. [Si4], Ch. 5 (cf. Tsirelson [Ts]).

4. **Conformal mapping**

   The class $H^{1/2}$ also occurs in work of Zygmund on conformal mapping ([Z2], VII.10).

5. **Rapid decay and continuability**

   Even stronger than the strong conditions considered here in §§5, 6 is assuming that the Verblunsky coefficients are rapidly decreasing. This is connected to analytic continuability of the Szegő function beyond the unit disk; see [Si7].

6. **Scattering theory**

   The implication from the strong Szegő (or Golinskii-Ibragimov) condition to the Helson-Szegő/Helson-Sarason condition (PA) has a recent analogue in scattering theory (Golinskii et al. [GolKPY], under ‘(GI) implies (HS)’).

7. **Wavelets**

   Traditionally, the subject of time series seemed to consist of two non-intercommunicating parts, ‘time domain’ and ‘frequency domain’ (known to be equivalent to each other via the Kolmogorov Isomorphism Theorem of §2). The subject seemed to suffer from schizophrenia (see e.g. [BriKri] and [HaKR]) – though the constant relevance of the spectral or frequency side to questions involving time directly is well illustrated in the apt title ‘Past and future’ of the paper by Helson and Sarason [HeSa] (cf. [Pel] §8.6). This unfortunate schism has been healed by the introduction of wavelet methods (see e.g. the standard work Meyer [Me], Meyer and Coifman [MeCo], and in OPUC, Treil and Volberg [TrVo1]). The practical importance of this may be seen in the digitization of the FBI's finger-print data-bank (without which the US criminal justice system would long ago have collapsed). Dealing with time and frequency together is also crucial in other areas, e.g. in the high-quality reproduction of classical music.

8. **Higher dimensions: matrix OPUC (MOPUC)**

   We present the theory here in one dimension for simplicity, reserving the case of higher dimensions for a sequel [Bin2]. We note here that in higher dimensions the measure $\mu$ and the Verblunsky coefficients $\alpha_n$ become matrix-valued (matrix OPUC, or MOPUC), so one loses commutativity. The multidimensional case is needed for portfolio theory in mathematical finance, where one holds a (preferably balanced) portfolio of risky assets rather than one; see e.g. [BinFrKi].

9. **Non-commutativity**

   Much of the theory presented here has a non-commutative analogue in operator theory; see Blecher and Labuschagne [BlLa], Bekjan and Xu [BekXu] and the references cited there.
10. Non-stationarity

As mentioned in §1, the question of whether or not the process is stationary is vitally important, and stationarity is a strong assumption. The basic Kolmogorov Isomorphism Theorem can be extended beyond the stationary case in various ways, e.g. to harmonisable processes (see e.g. [Rao]). For background, and applications to filtering theory, see e.g. [Kak]; for filtering theory, we refer to e.g. [BaiCr].

11. Continuous time

The Szegő condition (Sz) for the unit circle (regarded as the boundary of the unit disc) corresponds to the condition

\[ \int_{-\infty}^{\infty} \log |f(x)| \frac{dx}{1 + x^2} > -\infty \]

for the real line (regarded as the boundary of the upper half-plane). This follows from the Möbius function \( w = (z - i)/(z + i) \) mapping the half-plane conformally onto the disc; see e.g. [Du], 189-190, [Nik2] Ch. A6. The consequences of this condition are explored at length in Koosis’ monograph on the ‘logarithmic integral’, [Koo2]. Passing from the disc to the half-plane corresponds probabilistically to passing from discrete to continuous time (and analytically to passing from Fourier series to Fourier integrals). The probabilistic theory is considered at length in Dym and McKean [DymMcK].

12. Gaussianity and linearity

We have mentioned the close links between Gaussianity and linearity in §1. For background on Gaussian Hilbert spaces and Fock space, see Janson [Jan], Peller [Pel]; for extensions to §§6.1, 7 in the Gaussian case, see [IbRo, Pel], [Bra1] §5. To return to the undergraduate level of our opening paragraph: for an account of Gaussianity, linearity and regression, see e.g. Williams [Wil], Ch. 8, or [BinFr].

Acknowledgements

This work arises out of collaboration with Akihiko Inoue of Hiroshima University and Yukio Kasahara of Hokkaido University. It is a pleasure to thank them both. It is also a pleasure to thank the Mathematics Departments of both universities for their hospitality, and a Japanese Government grant for financial support. I am very grateful to the referee for a thorough and constructive report, which led to many improvements.

References

Szegő’s theorem and its probabilistic descendants


[Bra1] Bradley, R. C., Basic properties of strong mixing conditions. Pages 165-192 in [EbTa], MR0899990


Szegő’s theorem and its probabilistic descendants


Szegő’s theorem and its probabilistic descendants


Szegő’s theorem and its probabilistic descendants


[LevMcK] Levinson, N. and McKean, H. P., Weighted trigonometrical approximation on $\mathbb{R}^1$ with application to the germ field of a stationary Gaussian noise. Acta Math. 112 (1964), 99-143. MR0163111


[LiXi] Li, L. M. and Xie, Z., Model selection and order determination for time series by information between the past and the future. J. Time Series Analysis 17 (1996), 65-84. MR1380897


Nabil H. Bingham


[Sa1] Sarason, D., *Function theory on the unit circle*. Virginia Polytechnic Institute and State University, Blacksburg VA, 1979. MR0521811


Szegö, G., Beiträge zur Theorie der Toeplitzschen Formen, II. Math. Z. 9 (1921), 167-190. MR1544462


Szegö, G., On certain Hermitian forms associated with the Fourier series of a positive function. Festschrift Marcel Riesz 222-238, Lund, 1952. MR0051961


