The nonlinear Fourier transform

Terence Tao

Christoph Thiele

Ya-Ju Tsai

Department of Mathematics, UCLA, Los Angeles, CA 90095
E-mail address: tao@math.ucla.edu

Department of Mathematics, UCLA, Los Angeles, CA 90095
E-mail address: thiele@math.ucla.edu

Department of Mathematics, UCLA, Los Angeles, CA 90095
E-mail address: yaju.tsai@gmail.com

1991 *Mathematics Subject Classification.* Primary ???

The author is supported by ???.
The author is supported by ???.
The author is supported by ???.

Contents

CHAPTER 1

The nonlinear Fourier transform

1.1. Introduction

These are lecture notes for a short course presented at the IAS Park City Summer School in July 2003 by the second author. The material of these lectures has been developed in cooperation by both authors.

The aim of the course was to give an introduction to nonlinear Fourier analysis from a harmonic analyst's point of view. Indeed, even the choice of the name for the subject reflects the harmonic analyst's taste, since the subject goes by many names such as for example scattering theory, orthogonal polynomials, operator theory, logarithmic integrals, continued fractions, integrable systems, Riemann Hilbert problems, stationary Gaussian processes, bounded holomorphic functions, etc.

We present only one basic model for the nonlinear Fourier transform among a large family of generalizations of our model. The focus then is to study analogues of classical questions in harmonic analysis about the linear Fourier transform in the setting of the nonlinear Fourier transform. These questions concern for example the definition of the Fourier transform in classical function spaces, continuity properties, invertibility properties, and a priori estimates. There is an abundance of analytical questions one can ask about the nonlinear Fourier transform, and we only scratch the surface of the subject.

The second half of the lecture series is devoted to showing how the nonlinear Fourier transform appears naturally in several fields of mathematics. We only present a few of the many applications that are suggested by the above (incomplete) list of names for the subject.

There is a vast literature on the subject of this course, in part generated by research groups with few cross-references to each other. Unfortunately we are not sufficiently expert to turn these lecture notes into anything near a survey of the existing literature. In the bibliography, we present only a small number of fairly randomly chosen entrance points to the vast literature.

We would like to thank the Park City Math Institute, its staff, and the conference organizers for organizing a stimulating and enjoyable summer school. We would like to thank R. Killip and S. Klein for carefully reading earlier versions of the manuscript and making many suggestions to improve the text. Finally, we thank J. Garnett for teaching us bounded analytic functions.

1.2. The nonlinear Fourier transform on l_0, l^1 and l^p

1.3. The nonlinear Fourier transform

In this lecture series, we study a special case of a wide class of nonlinear Fourier transforms which can be formulated at least as general as in the framework of generalized AKNS-ZS systems in the sense of ([**3**]). For simplicity we refer to the special case of a nonlinear Fourier transform in this lecture series as "the nonlinear Fourier transform", but the possibility of a more general setting should be kept in mind.

More precisely, we discuss (briefly) a nonlinear Fourier transform of functions on the real line, and (at length) a nonlinear Fourier series of coefficient sequences, i.e., functions on the integer lattice $\mathbf{Z}$. Fourier series can be regarded as abstract Fourier transform on the circle group $\mathbf{T}$ or dually as abstract Fourier transform on the group $\mathbf{Z}$ of integers, while ordinary Fourier transform is the abstract Fourier transform of the group $\mathbf{R}$ of real numbers. We shall therefore use the word *Fourier transform* for both models which we discuss. Indeed, to the extend that we discuss the general theory here, it is mostly parallel in both models, with the possible exception of the general existence result for an inverse Fourier transform in Lecture 1.10 which the authors have not been able to verify in the model of the nonlinear Fourier transform of functions on the real line.

For a sequence $F = (F_n)$ of complex numbers parameterized by $n \in \mathbf{Z}$, we define the Fourier transform as

$$\widehat{F}(\theta) = \sum_{n\in\mathbf{Z}} F_n e^{-2\pi i\theta n} \tag{1.1}$$

and one has the inversion formula

$$F_n = \int_0^1 \widehat{F}(\theta) e^{2\pi i\theta n}\, d\theta$$

A natural limiting process takes this Fourier transform to the usual Fourier transform of functions on the real line. We have made the choice of signs in the exponents so that this limit process is consistent with the definition of the Fourier transform in [**26**].

We shall pass to a complex variable

$$z = e^{-2\pi i\theta}$$

so that (1.1) becomes

$$\widehat{F}(z) = \sum_{n\in\mathbf{Z}} F_n z^n$$

after identifying 1-periodic functions in θ with functions in $z \in \mathbf{T}$. The choice of sign in the exponent here is the one most convenient for us.

The discrete nonlinear Fourier transform acts on sequences F_n parameterized by the integers, $n \in \mathbf{Z}$, such that each F_n is a complex number in the unit disc $\mathcal{D}$. To begin with we shall assume these sequences are compactly supported. That is, $F_n = 0$ for all but finitely many values of n.

For a complex parameter z consider the following formally infinite recursion:

$$\begin{pmatrix} a_n & b_n \end{pmatrix} = \frac{1}{\sqrt{1-|F_n|^2}} \begin{pmatrix} a_{n-1} & b_{n-1} \end{pmatrix} \begin{pmatrix} 1 & F_n z^n \\ \overline{F_n} z^{-n} & 1 \end{pmatrix}$$

$$a_{-\infty} = 1, \quad b_{-\infty} = 0$$

Here $a_{-\infty} = 1$ and $b_{-\infty} = 0$ is to be interpreted as $a_n = 1$ and $b_n = 0$ for sufficiently small n, which is consistent with the recursion formula since the transfer matrix

$$\frac{1}{\sqrt{1-|F_n|^2}} \begin{pmatrix} 1 & F_n z^n \\ \overline{F_n} z^{-n} & 1 \end{pmatrix} \tag{1.2}$$

is the identity matrix for sufficiently small n by the assumption that F_n is compactly supported.

The nonlinear Fourier transform of the sequence F_n is the pair of functions (a_∞, b_∞) in the parameter $z \in \mathbf{T}$, where a_∞ and b_∞ are equal to a_n and b_n for sufficiently large n. We write

$$\overbrace{F}(z) = (a_\infty(z), b_\infty(z))$$

We will momentarily identify the pair of functions (a_∞, b_∞) with an $SU(1,1)$ valued function on $\mathbf{T}$.

While evidently a_∞ and b_∞ are finite Laurent polynomials in z (rational functions with possible poles only at 0 and ∞), we regard the nonlinear Fourier transform as functions on the unit circle $\mathbf{T}$. Later, when we consider properly infinite sequences F_n, restriction to $\mathbf{T}$ as domain will be a necessity.

Observe that for $z \in \mathbf{T}$ the transfer matrices are all in $SU(1,1)$. Hence we can write equivalently for the above recursion

$$\begin{pmatrix} a_n & b_n \\ \overline{b_n} & \overline{a_n} \end{pmatrix} = \frac{1}{\sqrt{1-|F_n|^2}} \begin{pmatrix} a_{n-1} & b_{n-1} \\ \overline{b_{n-1}} & \overline{a_{n-1}} \end{pmatrix} \begin{pmatrix} 1 & F_n z^n \\ \overline{F_n} z^{-n} & 1 \end{pmatrix}$$

with

$$\begin{pmatrix} a_{-\infty} & b_{-\infty} \\ \overline{b_{-\infty}} & \overline{a_{-\infty}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and all matrices

$$\begin{pmatrix} a_n & b_n \\ \overline{b_n} & \overline{a_n} \end{pmatrix}$$

are in $SU(1,1)$, and in particular $|a_n|^2 = 1 + |b_n|^2$.

Thus the Fourier transform can be regarded as a map

$$l_0(\mathbf{Z}, \mathcal{D}) \to C(\mathbf{T}, SU(1,1))$$

where $l_0(\mathbf{Z}, \mathcal{D})$ are the compactly supported sequences with values in $\mathcal{D}$, and $C(\mathbf{T}, SU(1,1))$ are the continuous functions on $\mathbf{T}$ with values in $SU(1,1)$.

While we shall not do this here, one can naturally define similar nonlinear Fourier transforms for a variety of Lie groups in place of $SU(1,1)$. The group $SU(2)$ leads to an interesting example. We remark that here we define Fourier transforms using Lie groups in a quite different manner from the way it is done in representation theory. There one defines Fourier transforms of complex valued functions on groups, and one remains in the realm of linear function spaces. Here we end up with group valued functions, a much more nonlinear construction.

If E is an open set in the Riemann sphere, define E^* to be the set reflected across the unit circle, i.e,

$$E^* = \{z : \overline{z}^{-1} \in E\}$$

The operation $*$ is the identity map on $E \cap \mathbf{T}$.

If c is a function on E, define

$$c^*(z) = \overline{c(\overline{z}^{-1})}$$

as a function on E^*. This operation preserves analyticity. On the circle T, this operation coincides with complex conjugation:

$$c^*(z) = \overline{c(z)}$$

for all $z \in \mathbf{T} \cap E$.

We then observe the recursion

$$\begin{pmatrix} a_n & b_n \\ b_n^* & a_n^* \end{pmatrix} = \frac{1}{\sqrt{1-|F_n|^2}} \begin{pmatrix} a_{n-1} & b_{n-1} \\ b_{n-1}^* & a_{n-1}^* \end{pmatrix} \begin{pmatrix} 1 & F_n z^n \\ \overline{F_n} z^{-n} & 1 \end{pmatrix} \tag{1.3}$$

with

$$\begin{pmatrix} a_{-\infty} & b_{-\infty} \\ b_{-\infty}^* & a_{-\infty}^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

All entries in these matrices are meromorphic functions on the entire Riemann sphere. Namely, these recursions hold on $\mathbf{T}$ and thus hold on the entire sphere by meromorphic continuation of a_n, b_n, a_n^*, b_n^*.

Observe that the matrix

$$\begin{pmatrix} a_n(z) & b_n(z) \\ b_n^*(z) & a_n^*(z) \end{pmatrix}$$

is not necessarily in $SU(1,1)$ for z outside the circle $\mathbf{T}$. However,

$$a_n a_n^* = 1 + b_n b_n^*$$

continues to hold on the complex plane since it holds on the circle $\mathbf{T}$.

Thinking of the pair (c, d) as the first row of an element of a function which takes values in $SU(1,1)$ on the circle $\mathbf{T}$, we shall use the convention to write

$$(a, b)(c, d) = (ac + bd^*, ad + bc^*)$$

For small values of F_n the nonlinear Fourier transform is approximated the linear inverse Fourier transform. This can be seen by linearizing in F. The factor $(1-|F_n|^2)^{-1/2}$ is quadratic and we disregard it. The remaining formula for a_∞ and b_∞ is polynomial in F and $\overline{F}$. If we only collect the constant and the linear term, we obtain

$$(a_\infty, b_\infty) = (1, \sum_{n\in\mathbf{Z}} F_n z^n)$$

Thus a_∞ is constant equal to 1 in linear approximation and b_∞ is the Fourier transform

$$\sum_{n\in\mathbf{Z}} F_n z^n$$

in linear approximation.

The following lemma summarizes a few algebraic properties of the nonlinear Fourier transform.

Lemma 1.1. *If $F_n = 0$ for $n \neq m$, then*

$$\widehat{(F_n)} = (1-|F_m|^2)^{-1/2}(1, F_m z^m) \tag{1.4}$$

If $\widehat{(F_n)} = (a, b)$, *then we have for the shifted sequence whose* n*-th entry is* F_{n+1}

$$\widehat{(F_{n+1})} = (a, bz^{-1}) \tag{1.5}$$

If the support of F *is entirely to the left of the support of* G*, then*

$$\widehat{(F+G)} = \widehat{F}\,\widehat{G} \tag{1.6}$$

If $|c| = 1$ *then*

$$\widehat{(cF_n)} = (a, cb) \tag{1.7}$$

For the reflected sequence whose n*-th entry is* F_{-n}

$$\widehat{(F_{-n})}(z) = (a^*(z^{-1}), b(z^{-1})) \tag{1.8}$$

Finally, for the complex conjugate of a sequence, we have

$$\widehat{(\overline{F_n})}(z) = (a^*(z^{-1}), b^*(z^{-1})) \tag{1.9}$$

Observe that statements (1.5),(1.7), (1.8) and (1.9) are exactly the behaviour of the linearization $a \sim 1$ and $b \sim \sum F_n z^n$. Statements (1.5) and (1.7) are most easily proved by conjugation with diagonal elements in $SU(1,1)$. Concerning statements (1.8) and (1.9), observe that under the reflection $n \to -n$ or under complex conjugation the Laurent expansion of the diagonal element a turns into the expansion with complex conjugate coefficients.

1.4. The image of finite sequences

Our next concern is the space of functions a_∞, b_∞ obtained as Fourier transforms of finite D-valued sequences F_n.

It is immediately clear that a and b are finite Laurent polynomials. The following lemma describes the degree of these Laurent polynomials. Define the upper degree of a Laurent polynomial to be the largest N such that the N-th coefficient is nonzero, and define the lower degree to be the least N such that the N-th coefficient is nonzero.

Lemma 1.2. *Let* F_n *be a nonzero finite sequence with NLFT* (a, b)*. Let* N_- *be the smallest integer such that* $F_{-N} \neq 0$*. and let* N_+ *be the largest integer such that* $F_{N_+} \neq 0$*, Then* a *is a Laurent polynomial*

$$a = \sum_{n=N_- - N_+}^{0} \check{a}(n) z^k$$

with exact lowest degree $N_- - N_+$ *and exact highest degree* 0*. The constant term of this Laurent polynomial is*

$$\check{a}(0) = \prod_k (1 - |F_k|^2)^{-1/2}$$

Moreover, b *is of the form*

$$b = \sum_{k=N_-}^{N_+} \check{b}(n) z^n$$

with exact highest degree N_+ *and exact lowest degree* N_-*.*

A particular consequence of this lemma is that the order of the highest and lowest nonvanishing coefficients for the Laurent polynomial of b are the same as for the sequence F.

PROOF. The Lemma can be proved by induction on the length $1+N_+-N_-$ of the sequence F_n. If the length is 1, then (a,b) is equal to a transfer matrix (set $N=N_-=N_+$),

$$(a,b)=\Big((1-|F_N|^2)^{-1/2},(1-|F_N|^2)^{-1/2}F_Nz^N\Big)$$

This proves the lemma for sequences of length one.

Now let $l>1$ and assume the theorem is true for lengths less than l and assume F_n has length l. Set $N=N_+$, then we have

$$a=(1-|F_N|^2)^{-1/2}a'+(1-|F_N|)^{-1/2}z^{-N}\overline{F_N}b'$$

where (a',b') is the NLFT of the truncated sequence F' which coincides with F everywhere except for the N-th entry where we have $F'_N=0$. By induction, $z^{-N}b'$ has highest degree at most -1 (since b' has highest degree at most $N-1$), so the constant term of a is $(1-|F_n)^{-1/2}$ times the constant term of a' and there are no terms of positive order of a. The lowest order term of a' is at least N_--N_++1, while the lowest order term of $z^{-N}b'$ is exactly N_--N_+. Thus for a we have the lowest and highest order coefficients claimed in the lemma.

Similarly, we have

$$b=(1-|F_N|^2)^{-1/2}a'z^NF_N+(1-|F_N|)^{-1/2}b'$$

Only the second summand produces a nonzero coefficient of degree N_- and this is the lowest degree of b. Only the first summand produces a nonzero coefficient or order N_+ and this is the highest order coefficient of b. $\square$

The meromorphic extensions of a and b to the Riemann sphere satisfy the recursion (1.3). On the Riemann sphere, we shall be interested in the open unit disc $D=\{z:|z|<1\}$ and the unit disc at infinity $D^*=\{1/z:|z|<1\}$. Observe that a is holomorphic on D^*.

LEMMA 1.3. *Let* $(a,b)=\widehat{F}$ *for some finite sequence* F*. Then* a *has no zeros in the disc at infinity* D^**.*

PROOF. It suffices to prove the lemma under the assumption $F_n=0$ for $n<0$, because we can translate F and use Lemma 1.1.

The constant term of a is nonzero, and therefore a is not zero at ∞.

For $|z|>1$ we rewrite the recursion

$$\begin{pmatrix} a_n & b_n \end{pmatrix}=\frac{1}{\sqrt{1-|F_n|^2}}\begin{pmatrix} a_{n-1} & b_{n-1}\end{pmatrix}\begin{pmatrix} 1 & F_nz^n \\ \overline{F_n}z^{-n} & 1\end{pmatrix}$$

as

$$\begin{pmatrix} |z|^na_n & b_n \end{pmatrix}=\frac{1}{\sqrt{1-|F_n|^2}}\begin{pmatrix} |z|^na_{n-1} & b_{n-1}\end{pmatrix}\begin{pmatrix} 1 & F_n(z/|z|)^n \\ \overline{F_n}(z/|z|)^{-n} & 1\end{pmatrix}$$

Therefore

$$|z^na_n|^2-|b_n|^2=|z^na_{n-1}|^2-|b_{n-1}|^2\geq|z^{n-1}a_{n-1}|^2-|b_n|^2$$

because $|z| > 1$. Now it follows by induction that

$$|z^n a_n|^2 - |b_n|^2 > 0$$

because this is true for n near $-\infty$. Thus a is not zero at z. $\square$

COROLLARY 1.4. *We have $|a_n(z)| \geq 1$ for $|z| > 1$. Moreover, $a_n(\infty)$ is positive and greater than or equal to* 1*.*

PROOF. Since a has no zeros, the function $\log |a|$ is harmonic on D^*. On the boundary it is non-negative, and by the maximum principle it is non-negative on D^*. It remains to see positivity of $a(\infty)$, but this is clear since $a(\infty)$ is the constant term in the Laurent polynomial of a. $\square$

Next we observe that if (a, b) is the NLFT of a finite sequence, then a is already determined by b.

LEMMA 1.5. *Let b be a Laurent polynomial. Then there exists a unique Laurent polynomial a such that $aa^* = 1 + bb^*$, a has no zeros in D^*, and $a(\infty) > 0$.*

PROOF. Observe that $P = 1 + bb^*$ is a nonzero Laurent polynomial. It can only have poles at 0 and ∞, and by symmetry the order of these two poles have to be equal. Assume P has pole of oder n at 0, clearly $n \geq 0$. By symmetry P has pole of order n at ∞. Since P is clearly has no zeros on $\mathbf{T}$, by symmetry it has exactly n zeros in D and n zeros in D^*.

Uniqueness: In order for aa^* to have the correct order of pole at 0, the Laurent polynomial a has to be a polynomial in z^{-1} of order n. In order for aa^* to have the same zeros as P, it has to have the same zeros as P in D and no other zeros. This determines a up to a scalar factor. The condition $a(\infty) > 0$ determines a up to a positive factor. Since $1 + bb^*$ is nonzero, this positive factor is determined by $1 + bb^* = aa^*$.

Existence: Let a be a polynomial of degree n in $1/z$ whose zeros are exactly the zeros of P in D. Then $a(\infty) \neq 0$. By multiplying by a phase factor we may assume $a(\infty) > 0$. By multiplying by a positive factor we may assume that a^*a coincides with P on at least one point of $\mathbf{T}$. Then $1 + bb^* - aa^*$ has $2n + 1$ zeros, but at most two poles of order n. Thus $1 + bb^* - aa^* = 0$. $\square$

We are now ready to characterize the target space of the Fourier transform of finite sequences.

THEOREM 1.6. *The nonlinear Fourier transform is a bijection from the set of all finite sequences (F_n) in the unit disc into the space of all pairs (a, b) with b an arbitrary Laurent polynomial and a the unique Laurent polynomial which satisfies $aa^* = 1 + bb^*$, $a(\infty) > 0$, and has no zeros in D^*.*

Remark: While not stated explicitly in the theorem, all pairs (a, b) described in the theorem satisfy not only $a(\infty) > 0$ but also $a(\infty) \geq 1$. Moreover, a and b have the same length. This follows from the theorem and the previous discussion.

PROOF. Clearly the Fourier transform maps into the described space by the previous discussion.

We know that the upper and lower degree of F are the same as the upper and lower degree of b. Thus it suffices to prove bijectivity under the assumption of fixed upper and lower degree of F and b. By shifting F and b we may assume both have lower degree 0, and we can use induction on the common upper degree N.

In the case $N = 0$, meaning $F \equiv 0$, and the in case $N = 1$ we have

$$b = F_0(1 - |F_0|^2)^{-1/2}$$

and the map $F_0 \to b$ is clearly a bijection from D to $\mathbf{C}$. Now assume we have proved bijectivity up to upper degree $N - 1$.

We first prove injectivity, i.e., F of upper degree N can be recovered from (a, b). It suffices to show that F_0 can be recovered from (a, b). Then we can by induction recover the truncated sequence F', which coincides with F except for $F'_0 = 0$ from the Fourier transform of F' which can be calculated as

$$(a', b') = (1 - |F_0|^2)^{-1/2}(1, -F_0)(a, b)$$

However, this last identity also implies that

$$0 = b'(0) = (1 - |F_0|^2)^{1/2} b'(0) = b(0) - F_0 a^*(0)$$

Hence

$$F_0 = \frac{b(0)}{a^*(0)}$$

and this quotient is well defined since $a^*(0) \neq 0$. Thus F_0 is determined by (a, b) and we have proved injectivity for upper degree N.

Next, we prove surjectivity. Let b have upper degree N and let a be the unique Laurent polynomial which satisfies $aa^* = 1 + bb^*$, $a(\infty) > 0$, and has no zeros in D^*. We set formally

$$F_0 = \frac{b(0)}{a^*(0)}$$

Observe that $|F_0| < 1$ since b/a^* is holomorphic in D, continuous up to $\mathbf{T}$, and bounded by 1 on $\mathbf{T}$. We calculate formally the truncated Fourier transform data

$$(a', b') = (1 - |F_0|^2)^{-1/2}(1, -F_0)(a, b)$$

Then b' is a polynomial of upper degree at most N and lower degree at least 1.

It now suffices to prove that (a', b') is the nonlinear Fourier transform of a sequence of length $N - 1$. For this it suffices to show that a' is the unique Laurent polynomial such that $1 + b'b'^* = a'a'^*$, $a(\infty) > 0$, and a' has no zeros in D^*.

However,

$$1 + b'b'^* = a'a'^*$$

holds on $\mathbf{T}$ and therefore everywhere since the determinant of the matrix (a, b) coincides with that of (a', b') on $\mathbf{T}$. The recursion

$$(a, b) = (1 - |F_0|^2)^{-1/2}(1, F_0)(a', b')$$

implies

$$a(\infty) = (1 - |F_0|^2)^{-1/2} a'(\infty)$$

and thus $a'(\infty) > 0$. Finally, we observe that

$$a'(z) = (1 - |F_0|^2)^{-1/2}(a - F_0 b^*)$$

does not vanish in D^* since b^*/a is bounded by 1 in D^* and thus a strictly dominates $F_0 b^*$ in D^*. $\square$

1.5. Extension to l^1 sequences

We have defined the Fourier transform for finite sequences. Now, we would like to extend the definition to infinite sequences.

As in the case of the linear Fourier transform, the defining formula actually extends to sequences in $l^1(\mathbf{Z}, D)$, i.e., summable sequences of elements in $\mathcal{D}$.

Define a metric on the space $SU(1,1)$ by

$$\mathrm{dist}(G, G') = \|G - G'\|_{op}$$

This clearly makes $SU(1,1)$ a complete metric space, since $\mathbf{C}^4$ is a complete metric space and $SU(1,1)$ is a closed subset of $\mathbf{C}^4$ with the inherited topology.

Define $L^\infty(\mathbf{T}, SU(1,1))$ to be the metric space of all essentially bounded functions $G : \mathbf{T} \to SU(1,1)$

$$\sup_z \mathrm{dist}(\mathrm{id}, G(z)) < \infty$$

(in the usual sense of the essential supremum) with the distance

$$\mathrm{dist}(G, G') = \sup_z \mathrm{dist}(G(z), G'(z))$$

On the space of all summable sequences in D define the distance

$$\mathrm{dist}(F, F') = \sum_n \|T_n - T_n'\|_{op}$$

where T_n denotes the transfer matrix defined in (1.2). This makes $l^1(\mathbf{Z}, D)$ a complete metric space. We claim that on sets

$$B_\epsilon = \{F_n : \sup_n |F_n| < 1 - \epsilon\}$$

with $\epsilon > 0$ (every element in $l^1(\mathbf{Z}, D)$ is in such a set, and also every Cauchy sequence in $l^1(\mathbf{Z}, D)$ is inside one of these sets) this distance is bi-Lipschitz to

$$\mathrm{dist}'(F, F') = \sum_n |F_n - F'_n|$$

Namely, if F_n and F'_n are in B_ϵ, then

$$\|T_n - T'_n\|_{op} =$$

$$= \left|(1 - |F_n|^2)^{-1/2} - (1 - |F_n|^2)^{-1/2}\right| + \left|(1 - |F_n|^2)^{-1/2} F_n - (1 - |F'_n|^2)^{-1/2} F'_n\right|$$

This is bounded by a constant depending on ϵ. Thus we only need to show equivalence to $|F_n - F'_n|$ if the latter is smaller than a constant depending on ϵ. This however follow easily by Taylor expansion of the nonlinear terms in the expression for $\|T_n - T'_n\|_{op}$.

In particular, we observe that the finite sequences are dense in l^1.

LEMMA 1.7. *With the above metrics, the NLFT on $l_0(\mathbf{Z}, D)$ extends uniquely to a locally Lipschitz map from $l^1(\mathbf{Z}, D)$ to $L^\infty(\mathbf{T}, SU(1,1))$. The NLFT of such sequences can be written as the convergent infinite ordered product of the transfer matrices.*

PROOF. To prove existence and uniqueness of the extension, it suffices to prove the Lipschitz estimate on bounded sets for finite sequences.

Using $\|T_n\| \geq 1$ we have by Trotter's formula:

$$\|\prod_n T_n - \prod_n T'_n\|_{op} \leq [\sum_n \|T_n - T'_n\|_{op}][\prod_n \|T_n\|_{op}][\prod_n \|T'_n\|_{op}]$$

Moreover, we have

$$\prod_n \|T_n\|_{op} \leq \exp[\sum_n [\|T_n\|_{op} - 1]] \leq \exp[\sum_n \|T_n - \mathrm{id}\|_{op}]$$

hence the right-hand side remains bounded on bounded sets of $l^1(\mathbf{Z}, D)$. Thus, on a bounded set we have

$$\| \prod_n T_n - \prod_n T'{}_n\|_{op} \leq C[\sum_n \|T_n - T'{}_n\|_{op}]$$

with C depending on the set. This proves the Lipschitz estimate on bounded sets.

By the abstract theory of metric spaces, the NLFT extends to a locally Lipshitz map on $l^1(\mathbf{Z}, D)$. Given any sequence F, the truncations of the sequence to the interval $[-N, N]$ converge in $l^1(\mathbf{Z}, D)$ to F, and thus the sequence of nonlinear Fourier transforms converges to the NLFT of F.

Convegence in the target space is not only in $L^\infty(\mathbf{T}, SU(1,1))$, but also in the space $C(\mathbf{T}, SU(1,1))$ of continuous functions, which is a closed subspace. This implies that the NLFT of the truncated sequences converge pointwise and uniformly to the NLFT of F.

Observe that if $F \in B_\epsilon$, then the truncations of F_n to intervals $[-N, N]$ remain in B_ϵ, and converge to F in $l^1(\mathbf{Z}, D)$. Thus the products of the transfer matrices converge to the NLFT of F. $\square$

We observe that for $F \in l^1(\mathbf{Z}, D)$ we have

$$\sup_{z \in \mathbf{T}} \|(a(z), b(z))\|_{op} \leq \prod_n \|(1 - |F_n|^2)^{-1/2}(1, F_n)\|$$

or, applying the logarithm to both sides and using Lemma 1.55:

$$\sup_z \operatorname{arccosh}|a(z)| \leq \sum_n \operatorname{arccosh}((1 - |F_n|^2)^{-1/2})$$

Define $g(y) = (\log(\cosh(y)))^{1/2}$. Then g vanishes at 0 and is concave on the positive half axis, and therefore $g(x) + g(y) \geq g(x + y)$ for all $0 \leq x, y$, and the analogue inequality holds for any countable number of summands. Applying g to the last display we thus obtain

$$\sup_z (\log|a(z)|)^{1/2} \leq \sum_n (\log((1 - |F_n|^2)^{-1/2}))^{1/2} \tag{1.10}$$

In the following section, this estimate will be compared to estimates for sequences in spaces $l^p(\mathbf{Z}, D)$ for various p.

1.6. Extension to l^p sequences, $1 < p < 2$

In this section we define the nonlinear Fourier transform of l^p sequences with $1 < p < 2$. The discussion in this section is an extraction from the work of Christ and Kiselev. Mainly we rely on [**6**] and we state and use but not prove theorems from that paper.

Let $F \in l_0(\mathbf{Z}, D)$ be a finite sequence. By the distributive law, we can write

$$\overbrace{F}(z) = \prod_{n \in \mathbf{Z}} (1 - |F_n|^2)^{-1/2}(1, F_n z^n)$$

$$= \left(\prod_{n\in\mathbf{Z}}(1-|F_n|^2)^{-1/2}\right)\left(\sum_{n=0}^{\infty}\left[\sum_{i_1<\cdots<i_n}\prod_{k=1}^{n}(0,F_{i_k}z^{i_k})\right]\right)$$

Here all formally infinite products and sums are actually finite since almost all factors and summands are trivial, and where the summand for $n=0$ on the right-hand side is equal to $(1,0)$.

Observe that the first factor in the last display is independent of z and is a convergent product under the assumption $F\in l^p(\mathbf{Z},D)$ with $1<p<2$. The second factor in the last display is a multilinear expansion, i.e., a Taylor expansion in the sequence F near the trivial sequence $F=0$.

We shall see that for $F\in l^p(\mathbf{Z},D)$, each term in this multilinear expansion is well defined as a measurable function in z and that the multilinear expansion is absolutely summable for almost all $z\in\mathbf{T}$. This allows us to define the nonlinear Fourier transform for l^p sequences as a measurable function on $\mathbf{T}$.

THEOREM 1.8. *Let $1\le p<2$ and let p' be the dual exponent $p/(p-1)$. Let $F\in l^p(\mathbf{Z},D)$. Then the multilinear term*

$$\sum_{i_1<\cdots<i_n}\left[\prod_{k=1}^{n}(0,F_{i_k}z^{i_k})\right] \tag{1.11}$$

is a well defined element of the quasi-metric space $L^{p'/n}(\mathbf{T},M_{2\times 2})$ and depends continuously on the sequence $F\in l^p(\mathbf{Z},D)$. The multilinear expansion

$$\sum_{n=0}^{\infty}\left[\sum_{i_1<\cdots<i_n}\prod_{k=1}^{n}(0,F_{i_k}z^{i_k})\right] \tag{1.12}$$

is absolutely summable for almost every z. Defining

$$(a,b)(z):=\left(\prod_{n\in\mathbf{Z}}(1-|F_n|^2)^{-1/2}\right)\left(\sum_{n=0}^{\infty}\left[\sum_{i_1<\cdots<i_n}\prod_{k=1}^{n}(0,F_{i_k}z^{i_k})\right]\right)$$

we have $|a|^2=1+|b|^2$ almost everywhere, the function a has an outer extension to D^* with $a(\infty)>0$, and we have the estimate

$$\|(\log|a|)^{1/2}\|_{L^{p'}(\mathbf{T})}\le C_p\left\|\,|\log(1-|F_\cdot|^2)|^{1/2}\right\|_{l^p(\mathbf{Z})} \tag{1.13}$$

The case $p=1$ of inequality (1.13) has been observed in (1.10). Indeed, the case $p=1$ of this theorem is considerably easier than the case $p>1$.

The multilinear expansion described in this theorem fails to converge in general if $p=2$, see [**22**]. However, inequality (1.13) remains true for $p=2$ if the nonlinear Fourier transform is defined properly. We will discuss this in subsequent sections. It is an interesting open problem whether C_p in inequality (1.13) can be chosen uniformly in p as p approaches 2.

PROOF. The quasi-metric of the space $L^q(\mathbf{T},M_{2\times 2})$ is defined as

$$\left(\int_{\mathbf{T}}\|G(z)\|_{op}^q\right)^{1/q}$$

To prove that each multilinear term (1.11) is a well defined element in $L^{p'/n}(\mathbf{T},M_{2\times 2})$, it suffices to prove that the multilinear map T_n, originally defined on finite sequences by

$$T_n(F^{(1)}, \dots, F^{(n)})(z) = \sum_{i_1 < \cdots < i_n} \left[\prod_{k=1}^{n} (0, F_{i_k}^{(k)} z^{i_k}) \right]$$

satisfies the a priori estimate

$$\|T_n(F^{(1)}, \dots, F^{(n)})\|_{p'/n} \leq C \prod_{k=1}^{n} \|F^{(k)}\|_p$$

By the general theory of multilinear maps on topological vector spaces, the map T_n extends then uniquely to a continuous map

$$l^p(\mathbf{Z}) \times \cdots \times l^p(\mathbf{Z}) \to L^{p'/n}(\mathbf{T})$$

To prove the above a priori estimate, we use the following theorem formulated slightly differently in [**5**]:

THEOREM 1.9. *Let $1 < p < 2$. Let k_j for $j = 1, \dots, n$ be locally integrable functions on $\mathbf{R} \times \mathbf{R}$ such that the map*

$$K_j f(y) := \int k_j(y, x) f(x)\, dx \tag{1.14}$$

which is defined on bounded compactly supported f satisfies the a priori bound

$$\|K_j f\|_{p'} \leq C \|f\|_p$$

Then the operator T_n defined by

$$T_n(f_1, \dots, f_n) = \int_{x_1 < \cdots < x_n} \prod_{j=1}^{n} k_j(y, x_j) f_j(x_j)\, dx_j$$

satisfies the a priori bound

$$\|T_n(f_1, \dots, f_n)\|_{p'/n} \leq C \prod_{j=1}^{n} \|f_j\|_p$$

For the proof of this theorem we refer to [**5**] with improvements on the constant C in [**6**].

To apply the theorem to the case at hand we need to convert the integral in (1.14) to a sum. This is easily done by considering functions that are constant on each interval $[l, l+1)$ for $l \in \mathbf{Z}$, but some care is to be taken so that the integration in the definition of T_n can be turned into a summation over a discrete set. Specifically, define $k_j(y, x)$ to be zero for $y \notin [0, 2\pi]$ and

$$k_j(x, y) = e^{\pm iy[x/n]}$$

where $[x]$ denotes the largest integer smaller than x and the sign $\pm$ is positive or negative depending on whether j is odd or even. Further we define f_j such that for every integer m, the restriction of f_j to the interval $[nm, n(m+1))$ is equal to $F_m 1_{[n(m+1)-j, n(m+1)-j+1)}$ if j is odd and equal to the complex conjugate of this expression if j is even. With $z = e^{iy}$ one observes that for odd n

$$T_n(f_1, \dots, f_n) = \sum_{i_1 < \cdots < i_n} F_{i_1} z^{i_1} \overline{F_{i_2} z^{i_2}} \cdots F_{i_n} z^{i_n}$$

and for even n

$$T_n(f_1,\dots,f_n) = \sum_{i_1<\dots<i_n} F_{i_1} z^{i_1} \overline{F_{i_2} z^{i_2}} \dots \overline{F_{i_n} z^{i_n}}$$

Thus the desired bound for (1.11) follows from Theorem 1.9.

To obtain good bounds of the multilinear terms and conclude that the expansion (1.12) is absolutely summable almost everywhere we invoke the next theorem from [**6**].

Define a martingale structure on $\mathbf{R}$ to be a collection of intervals E_j^m with $m \geq 0$ and $1 \leq j \leq 2^m$ such that the following conditions are satisfied modulo endpoints

(1) The union $\cup_j E_j^m$ is equal to $\mathbf{R}$ for every m
(2) The intervals E_j^m and $E_{j'}^m$ are disjoint for $j \neq j'$
(3) If $j < j'$, $x \in E_j^m$ and $x' \in E_{j'}^m$, then $x < x'$
(4) For every j, m we have $E_j^m = E_{2j-1}^{m+1} \cup E_{2j}^{m+1}$

Given such a martingale structure, to each locally integrable function f we associate $g_f \in [0,\infty]$ by

$$g_f = \sum_{m=1}^{\infty} \left(\sum_{j=1}^{2^m} | \int_{E_j^m} f|^2 \right)^{1/2}$$

Theorem 1.10. *There is a constant B such that the following holds. Define*

$$T_n(f_1,\dots,f_n) := \int_{x_1<\dots<x_n} \prod_{i=1}^{n} f_i(x_i)\, dx_i \tag{1.15}$$

and let X be a finite set of locally integrable functions on $\mathbf{R}$. Assume that we are given a fixed martingale structure and define

$$g := \max_{f\in X} g_f$$

Then for every $n > 1$ and every $f_1,\dots,f_n \in X$,

$$|T_n(f_1,\dots,f_n)| \leq (n!)^{-1/2} B^n g^n$$

For each parameter y, we apply this theorem with the same k_i and f_i as in the application of the previous theorem. Thus the number g depends on the parameter y. Writing again $z = e^{iy}$ we obtain for even n

$$\sum_{i_1<\dots<i_n} F_{i_1} z^{i_1} \overline{F_{i_2} z^{i_2}} \dots \overline{F_{i_n} z^{i_n}} \leq (n!)^{-1/2} B^n g^n(z)$$

and similarly for odd n. If we can show that for a proper choice of the martingale structure the function $g(z)$ is finite for almost every z, then this implies immediately that the multilinear expansion (1.12) converges for almost all z.

We choose the martingale structure adapted to f_i in the sense of [**6**] (observe that all the f_i are identical up to complex conjugation). As in the remark to Theorem 1.1 of [**6**] one checks that

$$\|g\|_{p'} \leq \|F\|_p$$

Thus in particular the expansion (1.12) converges for almost every z.

The diagonal entry a of (a, b) is only affected by the even terms in the multilinear expansion (1.12). Thus we obtain with the same martingale structure as before

$$|a| \le [\prod_n (1 - |F_n|^2)^{-1/2}] \sum_n ((2n)!)^{-1/2} B^{2n} g^{2n}$$

$$\le [\prod_n (1 - |F_n|^2)^{-1/2}] \sum_n (n!)^{-1} B^{2n} g^{2n} = [\prod_n (1 - |F_n|^2)^{-1/2}] \exp(B^2 g^2)$$

$$(\log |a|)^{1/2} \le C[\sum_n |\log(1 - |F_n|^2)|]^{1/2} + Cg$$

This easily proves (1.13).

To complete the proof of Theorem 1.8 it remains to prove that the nonlinear Fourier transform (a, b) of an l^p sequence, which we have now defined using the multilinear expansion, satisfies $|a|^2 = 1 + |b|^2$ and a has an outer extension to D^*. We shall, for simplicity, only argue for sequences supported on $\mathbf{Z}_{\ge 0}$. The case of sequences on the full line $\mathbf{Z}$ is then a slight variation using truncations at both ends of the sequence.

Assume $F \in l^p(\mathbf{Z}_{\ge 0}, D)$ and consider the nonlinear Fourier transforms

$$(a^{(\le N)}, b^{(\le N)})$$

of the truncations $F^{(\le N)}$. If we can show that $(a^{(\le N)}, b^{(\le N)})$ converge to (a, b) almost everywhere, then we obtain immediately $|a|^2 = 1 + |b|^2$. Moreover, with the additional a priori estimate (1.16) and Lebesgue's dominated convergence theorem we have that $\log |a^{(\le N)}|$ converges to $\log |a|$ in L^1 and one easily concludes that a is outer. $\square$

THEOREM 1.11. *Let $F \in l^p(\mathbf{Z}_{\ge 0}, D)$. With the notation as above, the sequence $(a^{(\le N)}, b^{(\le N)})$ converges for almost every z to (a, b). Moreover, we have the a priori estimate*

$$\| \sup_N \log |a^{(\le N)}| \|_L^{p'} \le C_p \| |\log((1 - |F_n|^2)|^{1/2} \|_{l^p} \tag{1.16}$$

PROOF. We first show that almost everywhere convergence follows from the a priori estimate (1.16).

Let $M > N$ an write

$$(a^{(\le M)}, b^{(\le M)}) = (a^{(\le N)}, b^{(\le N)})(a', b')$$

By continuity of multiplication in $SU(1, 1)$, we have to show smallness of (a', b') for large N, independently of M and outside a set of prescribed small measure.

This however is precisely what (1.16) provides if applied to the tail $F^{(>N)}$.

Thus it remains to prove (1.16). This a priori estimate follows by arguments similar to those given before with the following theorem taken from [**6**], which is a modification of Theorem 1.10.

For a given martingale structure E_j^m we define

$$\tilde{g}(f) = \sum_{m=1}^{\infty} m \left(\sum_{j=1}^{2^m} | \int_{E_j^m} f|^2 \right)^{1/2}$$

THEOREM 1.12. *Define*

$$M_n(f_1,\dots,f_n) := \sup_{y,y'}\left|\int_{y<x_1<\dots<x_n<y'}\prod_{i=1}^n f_i(x_i)\,dx_i\right| \tag{1.17}$$

and let X be a finite set of locally integrable functions on $\mathbf{R}$. Assume we are given a fixed martingale structure and define

$$\tilde g := \max_{f\in X}\tilde g(f)$$

Then for every $n>1$ and every $f_1,\dots,f_n\in X$

$$|M_n(f_1,\dots,f_n)|\le (n!)^{-1/2}B^n\tilde g^n$$

Applying this theorem as before proves Theorem 1.11. This also completes the proof of Theorem 1.8. □

1.7. The nonlinear Fourier transform on $l^2(\mathbf{Z}_{\ge 0})$

1.8. Extension to half-infinite l^2 sequences

Most of this section is an adaption from an article by Sylvester and Winebrenner [**27**].

Assume F is a square summable sequence with values in the open unit disc, i.e., an element of $l^2(\mathbf{Z},D)$.

As for the linear Fourier transform, the defining equation for the nonlinear Fourier transform of F (infinite product of transfer matrices) does not necessarily converge for given $z\in\mathbf{T}$. Indeed, almost everywhere convergence of the partial products - a nonlinear version of a theorem of Carleson - is an interesting open problem, see the discussion in [**23**].

It however converges in a certain L^2 sense. As in the linear theory, the main ingredient to prove this is a Plancherel type identity.

LEMMA 1.13. *Let F_n be a finite sequence of elements in the unit disc. Then with $(a,b)=\widehat{F}$ we have*

$$\int_0^1 \log|a(e^{2\pi i\theta})|\,d\theta = \int_{\mathbf{T}}\log|a(z)| = -\frac{1}{2}\sum_n \log(1-|F_n|^2)$$

Remark 1: Observe that the integrand $\log|a(z)|$ on the left - hand side is positive, as is each summand $-\log(1-|F_n|^2)$ on the right - hand side. Thus this equation has the flavor of a norm identity.

Exercise: prove that in lowest order approximation (quadratic) this becomes the usual Plancherel identity.

Remark 2: This formula appears at least as early as in a 1936 paper by Verblunsky [**32**, p. 291].

PROOF. Since F_n is a finite sequence, $a=a_\infty$ is a polynomial in z^{-1} with constant term

$$\prod_n (1-|F_n|^2)^{-1/2}$$

and non-vanishing in D^* by Lemma 1.3.

Thus we have

$$\int_{\mathbf{T}} \log(a(z)) = \log(a(\infty)) = -\frac{1}{2}\sum_n \log(1-|F_n|^2)$$

Since the right-hand side is real, we also have

$$\int_{\mathbf{T}} \log|a(z)| = -\frac{1}{2}\sum_n \log(1-|F_n|^2)$$

□

Let $l^2(\mathbf{Z}_{\geq 0}, D)$ be the space of sequences supported on the nonnegative integers ("right half-line") with values in D. We now proceed to describe the space $\mathbf{H}$ which we will show is the range of the nonlinear Fourier transform on $l^2(\mathbf{Z}_{\geq 0}, D)$.

Consider the space $\mathbf{K}$ of all measurable $SU(1,1)$ functions on the circle with

$$\int_{\mathbf{T}} \log|a(z)| < \infty \tag{1.18}$$

We can embed this space into the space $L^1(\mathbf{T}) \times L^2(\mathbf{T}) \times L^2(\mathbf{T})$ mapping the function (a,b) to the function $(\log|a|, b/|a|, a/|a|)$. Clearly by our assumption on the space $\mathbf{K}$, $\log|a|$ is in $L^1(\mathbf{T})$, while $b/|a|$ is an essentially bounded measurable function because (b,a) is in $SU(1,1)$ almost everywhere, and $a/|a|$ is also an essentially bounded measurable function.

This embedding is indeed injective, since we can recover the modulus of a almost everywhere from $\log|a|$, we can recover the phase of a almost everywhere from $a/|a|$, thus we can recover a almost everywhere. Then we can recover b almost everywhere from $b/|a|$.

We endow the space $\mathbf{K}$ with the inherited metric. Thus the distance between two functions (a,b) and (a',b') is given by

$$\int_{\mathbf{T}} \log|a| - \log|a'| + \left(\int_{\mathbf{T}} |b/|a| - b'/|a'||^2\right)^{1/2} + \left(\int_{\mathbf{T}} |a/|a| - a'/|a'||^2\right)^{1/2}$$

Indeed, this makes $\mathbf{K}$ a complete metric space. To see this, it is enough to show that the image of the embedding is closed, because the space $L^1 \times L^2 \times L^2$ is complete. However, the image is the subspace of all functions (f,g,h) such that f is real and nonnegative almost everywhere, g satisfies $|ge^f|^2 + 1 = |e^f|^2$ almost everywhere, and h has values in $\mathbf{T}$ almost everywhere. Any limit of a sequence in this subspace satisfies the same constraints almost everywhere, and thus the subspace is closed.

We observe that in the above definition of the metric we could have used quotients of the type $b/a - b'/a'$ instead of $b/|a| - b'/|a'|$ and obtained an equivalent metric. This is because

$$\begin{aligned}|b/|a| - b'/|a'|| &= |(b/a)(a/|a|) - (b/a)(a'/|a'|) + (b/a)(a'/|a'|) - (b'/a')(a'/|a'|)| \\ &\leq |a/|a| - a'/|a'|| + |(b/a) - (b'/a')|\end{aligned}$$

and similarly

$$|b/a - b'/a'| \leq |(b/|a|) - (b'/|a'|)| + |a/|a| - a'/|a'||$$

Here we have used $|b|/|a| < 1$.

Likewise, we could have used quotients b/a^* in the definition of the distance. If G denotes the group $SU(1,1)$ and K is the compact subgroup of diagonal elements, then b/a parameterizes the left residue classes $K\backslash G$ while b/a^* parameterizes the right residue classes G/K.

Some calculations to follow will be slightly simplified if we note that we can work with quasi-metrics rather than metrics. A quasi-metric on a space is a distance function with definiteness, $d(x,y)=0$ implies $x=y$, symmetry, $d(x,y)=d(y,x)$, and a modified triangle inequality

$$d(x,z) \leq C\max(d(x,y),d(y,z))$$

Just like a metric, a quasi-metric defines a topology through the notion of open balls. This topology is completely determined by its convergent sequences. Also, Cauchy sequences are defined, and one can talk about completeness of quasi-metric spaces.

Two quasi-metrics on the same space are called equivalent if there is some strictly monotone continuous function C vanishing at 0 such that

$$d(x,y) \leq C(d'(x,y))$$
$$d'(x,y) \leq C(d(x,y))$$

Two equivalent quasi-metrics produce the same convergent sequences and the same Cauchy sequences.

A quasi-metric equivalent to the above metric is given by

$$\mathrm{dist}((a,b),(a',b')) =$$
$$\int_{\mathbf{T}} |\log|a| - \log|a'|| + \int_T |b/|a| - b'/|a'||^2 + \int |a/|a| - a'/|a'||^2$$

The space $\mathbf{K}$ is too large to be the image of the NLFT. As we shall see, the image of the NLFT lies in a subspace of $\mathbf{K}$ on which the phase of a does not contain any information other than that already contained in $\log|a|$.

More precisely, let $\mathbf{L}$ be the subspace of $\mathbf{K}$ consisting of all pairs (a,b) such that a is the boundary value of an outer function — also denoted by a — on D that is positive at ∞.

The outerness condition together with positivity of a at ∞ can be rephrased as

$$a/|a| = e^{-ig}$$

where

$$g = p.v. \int_{\mathbf{T}} \log|a(\zeta)| \mathrm{Im}(\frac{\zeta+z}{\zeta-z})\, d\zeta$$

i.e., g is the Hilbert transform of $\log|a|$. Recall that the harmonic extension of the Hilbert transform to D vanishes at 0.

Lemma 1.14. *Let F_n be a finite sequence of elements in $\mathcal{D}$ and $(a,b) = \overbrace{\ F\ }$. Then $(a,b) \in \mathbf{L}$.*

Proof. Clearly $(a,b) \in \mathbf{K}$. The function a is holomorphic in a neighborhood of the closure of D^* and has no zeros there. Therefore a and a^{-1} are in $H^\infty(D^*)$ and by Lemma 1.54 in the appendix the function a is outer on D^*. □

Lemma 1.15. *The space $\mathbf{L}$ is closed in $\mathbf{K}$. The restriction of the quasi-metric of $\mathbf{K}$ to $\mathbf{L}$ is equivalent to the following quasi-metric on $\mathbf{L}$:*

$$\mathrm{dist}((a,b),(a',b')) = \int_{\mathbf{T}} |\log|a| - \log|a'|| + \int_T |b/|a| - b'/|a'||^2$$

If (a, b) and (a', b') are in $\mathbf{L}$, then

$$\int_{\mathbf{T}} |a/|a| - a'/|a'||^2 \le C \int_{T} |\log|a| - \log|a'||$$

Namely,

$$\int |a/|a| - a'/|a'||^2$$

$$\le \int_0^2 \lambda|\{|a/|a| - a'/|a'|| > \lambda\}| \, d\lambda$$

$$\le \int_0^2 \lambda|\{|\mathrm{Im}(\log(a) - \log(a'))| > \lambda\}| \, d\lambda$$

where $\log(a)$ and $\log(a')$ are the boundary values of the branches of the logarithm which are real at ∞. Using the weak type 1 bound for the Hilbert transform, we can estimate the last display by

$$\le \int_0^2 \lambda(C\| \log|a| - \log|a'|\|_1/\lambda) \, d\lambda$$

$$\le C\| \log|a| - \log|a'|\|_1$$

This estimate shows that the quasi-metric defined in the lemma is equivalent on $\mathbf{L}$ to the distance defined on $\mathbf{K}$. Moreover, it shows that a sequence in $\mathbf{L}$ which is convergent in $\mathbf{K}$ converges to an element in $\mathbf{L}$ and thus $\mathbf{L}$ is closed.

Observe that on $\mathbf{L}$ we have

$$\mathrm{dist}(\mathrm{id}, (a, b)) \le 3 \int_{\mathbf{T}} \log|a| \tag{1.19}$$

because $aa^* = 1 + bb^*$ and $1 - x \le \log(1/x)$ imply

$$\int_{\mathbf{T}} |b/|a||^2 \le 2 \int_{\mathbf{T}} \log|a|$$

Also observe that for $(a, b) \in \mathbf{L}$, knowledge of the quotient $b/|a|$ (or b/a^* or b/a) is sufficient to recover (a, b). Namely, we can recover $|a|$ almost everywhere using the formula $|a|^2 = 1 + |b|^2$. Then we can recover the argument of a almost everywhere as the Hilbert transform of $\log|a|$. Then we can recover b almost everywhere from a and the quotient $b/|a|$ (or b/a^* or b/a).

Define the space $\mathbf{H}$ to be the subspace of $\mathbf{L}$ of all functions such that b/a^* is the boundary value of an analytic function on $\mathcal{D}$ (also denoted by b/a^*) that is in the Hardy space H^2. Since the Hardy space H^2 (identified as space of functions on $\mathbf{T}$) is a closed subspace of L^2, we have that $\mathbf{H}$ is a closed subspace of $\mathbf{L}$.

If F is a finite sequence supported on the right half-line, i.e., $F_n = 0$ for $n < 0$, then clearly $(a, b) \in \mathbf{H}$.

The following string of lemmas will prove that the nonlinear Fourier transform is a homeomorphism from $l^2(\mathbf{Z}_{\ge 0}, D)$ onto $\mathbf{H}$.

LEMMA 1.16. *Let F be a sequence in $l^2(\mathbf{Z}_{\ge 0}, D)$ and let $F^{(\le N)}$ denote the truncations to $[0, N]$. Then $(a_N, b_N) = \overbrace{F^{(\le N)}}$ is a Cauchy sequence in $\mathbf{H}$.*

Remark: Once this lemma has been established, it is possible to define $\widehat{F}$ to be the limit of this Cauchy sequence.

PROOF. We need the following auxiliary lemma:

LEMMA 1.17. *For $G, G' \in \mathbf{L}$ we have*

$$\mathrm{dist}(GG', G) \leq C\mathrm{dist}(G', \mathrm{id}) + C\left[\mathrm{dist}(G, \mathrm{id})\mathrm{dist}(G', \mathrm{id})\right]^{1/2}$$

PROOF. We have

$$\mathrm{dist}(GG', G) \sim \int_{\mathbf{T}} |\log|aa' + b\overline{b'}| - \log|a|| + \int_{\mathbf{T}} |\frac{ab' + b\overline{a'}}{aa' + b\overline{b'}} - \frac{b}{a}|^2 \tag{1.20}$$

Consider the first summand. We have

$$\begin{aligned} &|\log|aa' + b\overline{b'}| - \log|a|| \\ &= |\log|a'| + \log|1 + (b/a)\overline{(b'/a')}(\overline{a'}/a')|| \\ &\leq |\log|a'|| + |\log|1 + (b/a)\overline{(b'/a')}(\overline{a'}/a')|| \end{aligned}$$

Upon integration over $\mathbf{T}$, the first summand is bounded by $\mathrm{dist}(G', \mathrm{id})$. On the set of all z such that $(b'/a')(z) \geq 1/10$, we estimate

$$\begin{aligned} &|\log|1 + (b/a)\overline{(b'/a')}(\overline{a'}/a')|| \\ &\leq |\log(1 - |b'/a'|)| \leq C|\log|1 - |b'/a'|^2| = 2C|\log|a'|| \end{aligned}$$

which again upon integration is bounded by $\mathrm{dist}(G', \mathrm{id})$. On the set of all z with $(b'/a')(z) \leq 1/10$, we estimate

$$|\log|1 + (b/a)\overline{(b'/a')}(\overline{a'}/a')|| \leq C|b/a||b'/a'|$$

Upon integration over $\mathbf{T}$ and application of Cauchy-Schwarz, this is bounded by the square root of $\mathrm{dist}(G, \mathrm{id})\mathrm{dist}(G', \mathrm{id})$.

We consider the second summand on the right-hand side of (1.20). We claim that

$$\left|\frac{ab' + b\overline{a'}}{aa' + b\overline{b'}} - \frac{b}{a}\right| \leq C\frac{|b'|}{|a'|} + C\left|1 - \frac{a'}{|a'|}\right| \tag{1.21}$$

This will finish the proof, since upon taking the square and integrating, the right-hand side is bounded by $\mathrm{dist}(G', \mathrm{id})$ (here we use that the distance functions on $\mathbf{L}$ and $\mathbf{K}$ are equivalent). The claim is evident if $|b'|/|a'|$ is greater than $1/10$, since the left-hand side of (1.21) is bounded by 2. Assume

$$\frac{1}{10} > \frac{|b'|}{|a'|} > \frac{1}{10}\frac{|b|}{|a|}$$

Then we use triangle inequality on the left-hand side of (1.21) and obtain

$$\left|\frac{ab' + b\overline{a'}}{aa' + b\overline{b'}} - \frac{b}{a}\right| \leq C\left|\frac{ab' + b\overline{a'}}{aa'}\right| + C\frac{|b|}{|a|} \leq C\frac{|b'|}{|a'|}$$

Assume $|b'|/|a'| < \frac{1}{10}|b|/|a|$. Then

$$\begin{aligned} \left|\frac{ab' + b\overline{a'}}{aa' + b\overline{b'}} - \frac{b}{a}\right| &\leq \left|\frac{ab'}{aa' + b\overline{b'}}\right| + \left|(\frac{b\overline{a'}}{aa' + b\overline{b'}} - \frac{b\overline{a'}}{aa'})\right| + \left|\frac{\overline{a'}}{a'} - 1\right|\frac{|b|}{|a|} \\ &\leq C\frac{|b'|}{|a'|} + C\frac{|b'|}{|a'|} + C\left|1 - \frac{a'}{|a'|}\right| \end{aligned}$$

This proves Lemma 1.17. □

We continue the proof of Lemma 1.16. Consider

$$\operatorname{dist}(\prod_{n=0}^{M} T_n, \prod_{n=0}^{N} T_n) = \operatorname{dist}(GG', G)$$

where

$$G = \prod_{n=0}^{N} T_n \quad \text{and} \quad G' = \prod_{n=N+1}^{M} T_n$$

By the Plancherel identity and (1.19) we have

$$\operatorname{dist}(G, \mathrm{id}) \leq \int \log |a_N| \leq C \sum_{n=1}^{N} \big|\log |1 - |F_n|^2|\big| \leq C$$

(since $F \in l^2$) and similarly,

$$\operatorname{dist}(G', \mathrm{id}) \leq C \sum_{n=N+1}^{M} \big|\log |1 - |F_n|^2|\big| \leq \epsilon$$

if $N > N(\epsilon)$ is chosen large enough depending on the choice of ϵ. Thus, for $M > N > N(\epsilon)$, we have by Lemma 1.17

$$\operatorname{dist}(\prod_{n=0}^{M} T_n, \prod_{n=0}^{N} T_n) \leq C\epsilon^{1/2}$$

This shows that $\widehat{F^{(\leq N)}}$ is Cauchy in $\mathbf{H}$.

□

Thus we can define the NLFT on $l^2(\mathbf{Z}_{\geq 0}, D)$ as the limit of the NLFT of the truncated sequences. We have not yet shown any genuine continuity of the NLFT, but we will do that further below. Using Theorem 1.11, one can show that this definition of the NLFT coincides with the old definition on the subset $l^p(\mathbf{Z}_{\geq 0}, D)$ of $l^2(\mathbf{Z}_{\geq 0}, D)$ for $1 \leq p < 2$.

As the distance between the NLFT of the truncated sequence and the NLFT of the full sequence converges to 0, the Plancherel identity continues to hold on all of $l^2(\mathbf{Z}_{\geq 0}, D)$.

Lemma 1.18. *The NLFT is injective on $l^2(\mathbf{Z}_{\geq 0})$.*

Proof. We know for the finite truncations that

$$F_0 = b^{(\leq N)}(0)/a^{(\leq N)*}(0) = \int_{\mathbf{T}} b^{(\leq N)}/a^{(\leq N)*}$$

Where we used that b/a^* has holomorphic extension to a neighborhood of D.

By convergence of the data $(a^{(\leq N)}, b^{(\leq N)})$ in $\mathbf{H}$ we see that $b^{(\leq N)}/a^{(\leq N)*}$ converges in $L^2(\mathbf{T})$ to b/a^*, where (a, b) is the NLFT of F. Thus

$$F_0 = \int_{\mathbf{T}} b/a^*$$

Observe that the quotient b/a^* is sufficient to determine F_0. This is consistent with the earlier observation that this quotient contains the full information of $(a, b) \in \mathbf{H}$.

To proceed iteratively, we need to determine the NLFT $(\tilde{a}, \tilde{b})$ of the "layer stripped" sequence $\tilde{F}$ in $l^2(\mathbf{Z}_{\geq 0}, D)$; this is defined by $\tilde{F}_n = F_n$ for $n > 0$ and $\tilde{F}_0 = 0$. More precisely, we will determine the quotient $\tilde{b}/\tilde{a}^*$ using only the quotient b/a^*. Write $r = b/a^*$, $\tilde{r} = \tilde{b}/\tilde{a}^*$ etc.

Using the product formula for finite sequences we calculate

$$(\tilde{a}^{(\leq N)}, \tilde{b}^{\leq N}) = (1 - |F_0|^2)^{-1/2}(1, -F_0)(a^{(\leq N)}, b^{(\leq N)})$$

$$\tilde{r}^{(\leq N)} = \frac{r^{(\leq N)} - F_0}{-\overline{F_0} r^{(\leq N)} + 1}$$

As N tends to ∞, the equation tends in L^2 norm to

$$\tilde{r} = \frac{r - F_0}{-\overline{F_0} r + 1}$$

For the left-hand side this follows directly from the definitions. For the right-hand side this follows from the fact that the map

$$s \to \frac{s - F_0}{-\overline{F_0} s + 1}$$

has bounded derivative on the closure of D and thus turns L^2 convergence of r into L^2 convergence of $\frac{r - F_0}{-\overline{F_0} r + 1}$ (recall F_0 is fixed and $|F_0| < 1$.)

Thus we can calculate $\tilde{b}/\tilde{a}^*$.

By an inductive procedure (conjugate the new problem by a shift to reduce it to the old problem for sequences starting at 0) we can calculate all F_n. This proves injectivity.

$\square$

The layer stripping method in the proof of this lemma can be used to obtain the following result: If F is in l^2 and if $F^{(\leq N)}$ and $F^{(>N)}$ are the truncations to $[0, N]$ and $[N + 1, \infty)$, then

$$(a, b) = (a^{(\leq N)}, b^{(\leq N)})(a^{(>N)}, b^{(\leq N)})$$

This is clear if $N = 0$. If $N = 1$, this has been observe in the proof of the previous lemma. Then one can use induction to prove this for all N.

For later reference we note that

$$r_{>N} := b^{(>N)}/(a^{(>N)})^*$$

has a holomorphic extension to D which vanishes to order $N + 1$ at 0.

Lemma 1.19. *The NLFT is surjective from l^2 onto the space* $\mathbf{H}$

Proof. Let $(a, b) \in \mathbf{H}$. Then $r = b/a^*$ is an analytic function in D bounded by 1. Moreover, $r(0) < 1$ since the extension of r to $\mathbf{T}$ is strictly less than 1 almost everywhere. Following the calculations in the proof of the previous lemma, we set formally

$$F_0 = r(0)$$

and

$$z\tilde{r} = \frac{r - F_0}{-F_0 r - 1}$$

Being the Möbius transform of a bounded analytic function, the right-hand side is still an analytic function in D bounded by 1 and it vanishes at 0 by construction. Thus, by the Lemma of Schwarz, we can divide by z and calculate formally $\tilde{r}$ which is again an analytic function in D bounded by 1. This procedure can be iterated and gives a sequence F_n.

This iteration process can be applied to any analytic function bounded by 1, regardless of any further regularity of this function. This process is called Schur's algorithm after [**24**].

We show that $\tilde{r}$ as constructed above is of the form $\tilde{b}/\tilde{a}^*$ for some $(a,b) \in \mathbf{H}$. To this end, it suffices to show that $1-|\tilde{r}|^2$ (which is formally $|\tilde{a}|^{-2}$) is log integrable over $\mathbf{T}$. Then we can determine the outer function $\tilde{a}$ and calculate $\tilde{b}$. Observe that as bounded analytic function, $\tilde{r}$ is automatically in $H^2(D)$.

We have

$$\begin{aligned} 1 - |\tilde{r}|^2 &= \frac{|\overline{F_0}r - 1|^2 - |r - F_0|^2}{|\overline{F_0}r - 1|^2} \\ &= \frac{|\overline{F_0}r|^2 - |r|^2 - |F_0|^2 + 1}{|\overline{F_0}r - 1|^2} \\ &= \frac{(1 - |F_0|^2)(1 - |r|^2)}{|\overline{F_0}r - 1|^2} \end{aligned}$$

Observe that $\overline{F_0}r - 1$ is bounded and bounded away from 0 and so is its holomorphic extension to D, thus its logarithm is integrable over $\mathbf{T}$ and equal to the value of the logarithm at 0:

$$\int \log(1 - \overline{F_0}r) = \log(1 - |F_0|^2)$$

Taking logarithms and integrating gives

$$\int \log(1 - |\tilde{r}|^2) = \log(1 - |F_0|^2) + \int (1 - |r|^2) - 2\log(1 - |F_0|^2)$$

$$\int \log(1 - |\tilde{r}|^2) = -\log(1 - |F_0|^2) + \int \log(1 - |r|^2)$$

Discussing the signs we obtain

$$\int |\log(1 - |\tilde{r}|^2)| = -|\log(1 - |F_0|^2)| + \int |\log(1 - |r|^2)| \tag{1.22}$$

Thus $\log(1 - |\tilde{r}|^2)$ is integrable.

We can iterate to calculate formally F_n. Using (1.22) inductively, we obtain

$$\sum_{n=0}^{\infty} |\log(1 - |F_n|^2)| \le \int |\log(1 - |r|^2)| \tag{1.23}$$

Thus the sequence F_n we calculated is in $l^2(\mathbf{Z}_{\ge 0}, D)$ and it is a candidate for the preimage of (a,b) under the NLFT.

Let $(\tilde{a}, \tilde{b})$ denote the NLFT of F_n. We will show that indeed, $(a,b) = (\tilde{a}, \tilde{b})$ and we have equality in (1.23).

As noted earlier,

$$(\tilde{a}, \tilde{b}) = (\tilde{a}^{(\le N)}, \tilde{b}^{(\le N)})(\tilde{a}^{(>N)}, \tilde{b}^{(>N)})$$

where the factors on the right-hand side are defined by the usual truncations.

Observe that by unwinding Schur's algorithm introduced above, we obtain

$$(a,b) = (\tilde{a}^{(\le N)}, \tilde{b}^{(\le N)})(a^{(>N)}, b^{(>N)})$$

where $(a^{(>N)}, b^{(>N)})$ is the unique element in $\mathbf{H}$ such that $b^{(>N)})/a^{(>N)}$ is equal to the N-th function in Schur's algorithm.

Thus in the expression

$$(\tilde{a}, \tilde{b})^{-1}(a, b)$$

we can cancel factors to obtain

$$(\tilde{a}, \tilde{b})^{-1}(a, b) = (\tilde{a}^{(>N)}, \tilde{b}^{(>N)})^{-1}(a^{(>N)}, b^{(>N)}) \tag{1.24}$$

Consider an off- diagonal entry of the right-hand side of (1.24):

$$(\tilde{a}^{(>N)})^* b^{(>N)} - \tilde{b}^{(>N)}(a^{(>N)})^*$$

This is a Nevanlinna function on D. The Taylor coefficients of this function at 0 vanish up to order N. By (1.24), this function does not actually depend on N, therefore all Taylor coefficients at 0 vanish. Thus the function vanishes on D, and so do its radial limits on $\mathbf{T}$ almost everywhere. Thus the off diagonal coefficients of

$$(\tilde{a}, \tilde{b})^{-1}(a, b)$$

vanish and we have for some function c

$$(a, b) = (\tilde{a}, \tilde{b})(c, 0)$$

This function c has to have modulus 1 almost everywhere on $\mathbf{T}$, since the last display can be read as identity between $SU(1,1)$ valued measurable functions on $\mathbf{T}$. Calculating a diagonal entry in the last display, we obtain

$$a = \tilde{a}c$$

Since a and $\tilde{a}$ are outer, so is c. However, any outer function with constant modulus on $\mathbf{T}$ is constant. Moreover, c is positive since a and $\tilde{a}$ are positive at ∞. This proves $c = 1$. □

LEMMA 1.20. *The NLFT is a continuous map from $l^2(\mathbf{Z}_{\geq 0}, D)$ to $\mathbf{H}$.*

PROOF. Fix $F \in l^2(\mathbf{Z}_{\geq 0}, D)$ and choose $\epsilon > 0$. Choose N very large depending on ϵ.

For F' close to F depending on N, ϵ we write

$$\mathrm{dist}((a, b), (a', b')) \leq \mathrm{dist}((a, b), (a^{(\leq N)}, b^{(\leq N)})) +$$

$$+ \mathrm{dist}((a^{(\leq N)}, b^{(\leq N)}), ({a'}^{(\leq N)}, {b'}^{(\leq N)})) + \mathrm{dist}(({a'}^{(\leq N)}, {b'}^{(\leq N)}), (a', b'))$$

We intend to argue that all the terms on the right-hand side are less than $\epsilon/3$.

By the definition of (a, b), the first term can be made small by choosing N large enough. Likewise the third term can be made small, since the distance between the truncation and the full Fourier transform depends only on the l^2 norm of the sequence F' and the l^2 norm of the tail of this sequence, which can be both controlled by choosing N large enough and F' close enough to F. Thus it remains to control the middle term.

Consider the space of D-valued sequences on $[0, N]$ with the l^2 norm. Since the space is finite dimensional, the l^2 norm is equivalent to the l^1 norm.

Observe that for two matrices (a, b) and (a', b') in $SU(1,1)$ we have

$$|\log|a| - \log|a'|| \leq |a - a'| \leq \|(a, b) - (a', b')\|_{op}$$

and

$$|b/|a| - b'/|a'|| \leq |b - b'| \leq \|(a, b) - (a', b')\|_{op}$$

Thus, if $F'^{(\leq N)}$ is sufficiently close to $F^{(\leq N)}$ w.r.t. l^2 and thus l^1, then

$$\sup_z \left|\log |a^{(\leq N)}| - \log |a'^{(\leq N)}|\right| + \sup_z \left|b^{(\leq N)}/|a^{(\leq N)}| - b'^{(\leq N)}/|a'^{(\leq N)}|\right|$$

is small, and thus

$$\operatorname{dist}((a',b'),(a^{(\leq N)},a^{(\leq N)}))$$

is small. $\square$

We remark that this proof does not give any uniform continuity. The weak point in the argument is the comparison of the l^2 with the l^1 norm of a finite sequence without any good control over the length of the sequence.

LEMMA 1.21. *The inverse of the NLFT is a continuous map from* $\mathbf{H}$ *to* $l^2(\mathbf{Z}_{\geq 0}, D)$.

PROOF. We first prove that all F_n depend continuously on (a,b). This is clear for F_0 since

$$F_0 = \int b/a^*$$

and the integral is continuous in the L^2 norm of b/a^*.

To use induction, we need to show that (see the proof of injectivity)

$$r \to \frac{r - F_0}{-\overline{F_0}r + 1}$$

is a jointly continuous mapping of $F_0 \in D$ and $r \in H^2$ into H^2. This follows from the fact that the Möbius transform with F_0 provides a Lipschitz distortion on the closed unit disc, and the distortion depends continuously on F_0.

Now let (a,b) be given and let F be its inverse NLFT. Given ϵ, we can find N very large so that the $[N,\infty)$ tail of F is very small. Let (a',b') be close to (a,b) and let F' be the inverse NLFT. Then we can assume for all $n \leq N$

$$\log(1-|F_n|^2) - \log(1-|F_n'|^2)$$

is much smaller than ϵ, by continuity of all F_n individually.

Next we have

$$\sum_{n>N} |\log(1-|F_n'|^2)| = \sum_n |\log(1-|F_n'|^2)| - \sum_{n\leq N} |\log(1-|F_n'|^2)|$$

$$\leq \sum_n |\log(1-|F_n|^2)| - \sum_{n\leq N} |\log(1-|F_n|^2)| + \epsilon \leq 2\epsilon$$

Here we have used that (a,b) and (a',b') are close and therefore, by the Plancherel identity, the quantities

$$\sum_n \log(1-|F_n|^2)$$

$$\sum_n \log(1-|F'_n|^2)$$

are close. Also we have used the previously observed continuity for individual F_n.

Now it is straight forward to obtain

$$\sum_n |\log(1-|F_n|^2) - \log(1-|F'_n|^2)| < 4\epsilon$$

by considering separately $n > N$ and $n \leq N$. $\square$

1.9. Higher order variants of the Plancherel identity

The main ingredient in the l^2 theory of the nonlinear Fourier transform described in the previous section is the Plancherel identy

$$\int_{\mathbf{T}} \log|a| = -\frac{1}{2}\sum_{n\in\mathbf{Z}} \log(1-|F_n|^2)$$

Both sides of the identity are equal to $a(\infty)$, which on the left-hand side is expressed as a Cauchy integral and on the right-hand side is expressed in terms of the sequence F by solving explicitly the recursion for $a(\infty)$.

There are higher order identities of this type, which arise from calculating higher derivatives of $\log(a)$ at ∞. These identities are nonlinear analogues of Sobolev identities, i.e., identities between expressions for Sobolev norm of a function in terms of the function itself and in terms of its Fourier transform.

We discuss a^* instead of a. The contour integral for $\log(a^*)^{(k)}(0)$ can easily be written in closed form:

$$\log(a^*)(z) = \int_{\mathbf{T}} \frac{\zeta+z}{\zeta-z}\log|a|(\zeta) = \int_T \left[\frac{2\zeta}{\zeta-z}-1\right]\log|a|(\zeta)$$

Taking derivatives in z we obtain for $k>0$

$$\log(a^*)^{(k)}(z) = \int_T \frac{2\zeta k!}{(\zeta-z)^{k+1}}\log|a|(\zeta)$$

$$\log(a^*)^{(k)}(0) = \int_T \frac{2k!}{\zeta^k}\log|a|(\zeta)$$

Solving the recursion in terms of the F_n is harder to do in closed form. Such formulae are stated in Case's paper [**4**], see also [**19**].

We shall calculate only the cases $k=1,2$. For an application to the theory of orthogonal polynomials, see [**10**].

LEMMA 1.22. *For F a square summable sequence we have*

$$2\int_{\mathbf{T}} z^{-1}\log|a|(z) = \sum_n \overline{F_n}F_{n+1}$$

$$4\int_{\mathbf{T}} z^{-2}\log|a|(z) = -\sum_n(\overline{F_n}F_{n+1})^2 + 2\sum_n \overline{F_n}(1-|F_{n+1}|^2)F_{n+2}$$

PROOF. We first reduce to the case of compactly supported seqeunces F by approximating an arbitrary sequence by its truncations F_n. At least for half infinite sequences we have already seen that $\log|a_n|$ converges to $\log|a|$ in L^1 norm, and in the next section we will discuss and establish the same fact for general sequences in $l^2(D,\mathbf{Z})$. Thus the left-hand side of each of the identities in the lemma is well approximated by the truncations. Also the right-hand side clearly converges if F is in l^2. Thus it suffices to show the identities for compactly supported F.

Assume F is compactly supported. We expand the product

$$\prod_{n\in\mathbf{Z}}(1-|F_n|^2)^{-1/2}[(1,0)+(0,F_nz^n)]$$

Only the terms of even order in F contribute to the diagonal elements and thus

$$a^*(z)\prod_n(1-|F_n|^2)^{1/2}$$

$$= 1 + \sum_{n_1<n_2} \overline{F_{n_1}} F_{n_2} z^{n_2-n_1} + \sum_{n_1<n_2<n_3<n_4} \overline{F_{n_1}} F_{n_2} \overline{F_{n_3}} F_{n_4} z^{n_2-n_1} z^{n_4-n_3} + \dots$$

Since $n_1 < n_2$ and $n_3 < n_4$ etc., we see that the bilinear term in F is has lowest order z and the four-linear term has lowest order z^2, while all other terms have order at least z^3. Thus for the purpose of calculating the first two derivatives at ∞ we only need to consider the terms that are explicitly written.

Indeed, we have

$$a^*(z) \prod_n (1 - |F_n|^2)^{1/2} = 1 + \sum_n \overline{F_n} F_{n+1} z$$

$$+ \sum_n \overline{F_n} F_{n+2} z^2 + \sum_{n_1+1<n_2} \overline{F_{n_1}} F_{n+1} \overline{F_{n_2}} F_{n_2+1} z^2 + O(z^3)$$

Considering the case $k = 1$ we have

$$\log(a^*)'(0) = \frac{(a^*)'(0)}{a^*(0)} = \sum_n \overline{F_n} F_{n+1}$$

This proves the first identity of Lemma 1.22.

Considering the case $k = 2$ we have

$$\log(a^*)''(0) = \frac{(a^*)''(0)}{a(0)} - \frac{(a^*)'(0)^2}{a^*(0)^2}$$

$$= 2 \sum_n \overline{F_n} F_{n+2} + 2 \sum_{n_1+1<n_2} \overline{F_{n_1}} F_{n+1} \overline{F_{n_2}} F_{n_2+1} - [\sum_n \overline{F_n} F_{n+1}]^2$$

$$= -\sum_n (\overline{F_n} F_{n+1})^2 + 2 \sum_n \overline{F_n} (1 - |F_{n+1}|^2) F_{n+2}$$

This proves the second identity of Lemma 1.22. □

1.10. The nonlinear Fourier transform on $l^2(\mathbf{Z})$

1.11. The forward NLFT on $l^2(\mathbf{Z})$

We have defined the nonlinear Fourier transform for square summable sequences supported on the nonnegative integers. Indeed, we have shown that it is a homeomorphism onto $\mathbf{H}$, the space of all measurable $SU(1,1)$ valued function (a, b) such that a has an outer extension to D^*, $a(\infty) > 0$, and b/a^* has a holomorphic extension to D which is in the Hardy space $H^2(D)$.

Demanding that property (1.8) of Lemma 1.1 continues to hold for infinite sequences, we define for F supported on $n \le 0$:

$$\overbrace{F}(z) := (a^*(z^{-1}), b(z^{-1}))$$

where (a, b) is the Fourier transform of the reflected sequence $\tilde{F}$ with $\tilde{F}_n = F_{-n}$.

It is clear that the nonlinear Fourier transform thus defined is a homeomorphism from $l^2(\mathbf{Z}_{\le 0}, D)$ to $\mathbf{H}^*$, the latter denoting the space of all $SU(1,1)$ valued measurable functions (a, b) such that a has an outer extension to D^*, $a(\infty) > 0$, and b/a has a holomorphic extension to D^* which is in the Hardy space $H^2(D^*)$.

Let $\mathbf{H}_0^*$ be the space of all elements in $\mathbf{H}^*$ such that $b(\infty) = 0$. By the shifting property (1.5) of Lemma 1.1, it is easy to deduce that this space is the homeomorphic image of $l^2(\mathbf{Z}_{\le -1}, D)$.

If F_n is any square summable sequence in $l^2(\mathbf{Z}, D)$, then we can cut it as

$$F_n = F_n^{(\leq -1)} + F_n^{(\geq 0)}$$

where $F_n^{(\leq -1)} = 0$ for $n \geq 0$ and $F_n^{(\geq 0)} = 0$ for $n \leq -1$. Then we define a measurable $SU(1,1)$ valued function on $\mathbf{T}$ by

$$\overbrace{F} := \overbrace{F^{(\leq -1)}}\ \overbrace{F^{(\geq 0)}} \tag{1.25}$$

in accordance with property (1.6) of Lemma 1.1. We shall use the suggestive notation

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

for (1.25).

It is easy to verify that the NLFT defined by (1.25) on $l^2(\mathbf{Z}, D)$ satisfies the properties of Lemma 1.1. The properties of Lemma 1.1 imply that the exact location of the cut we used in (1.25) is not relevant for the definition.

As noted previously, the definition of the NLFT on $l^2(\mathbf{Z}_{\geq 0}, D)$ was consistent with earlier definitions on the subset $l^p(\mathbf{Z}_{\geq 0}, D)$ for $p < 2$. Passing from the half-line to the full line, since definition (1.25) and the old definition of the NLFT on $l^p(\mathbf{Z}, D)$ are consistent with Lemma 1.1, the two definitions coincide on $l^p(\mathbf{Z}, D)$.

Lemma 1.23. *The NLFT is continuous from $l^2(\mathbf{Z}, D)$ to $\mathbf{L}$. The Plancherel identity*

$$\int_{\mathbf{T}} \log |a(z)| = -\frac{1}{2} \sum_n \log(1 - |F_n|^2)$$

holds.

Proof. Recall that $\mathbf{L}$ is the space of all $SU(1,1)$ valued measurable functions (a, b) such that a has an outer extension to D^* and $a(\infty) > 0$.

First we check that the nonlinear Fourier transform indeed maps to $\mathbf{L}$. We need to verify that a has an outer extension to D^* and that $a(\infty) > 0$. But we have

$$a = a_- a_+ + b_- b_+^*$$

and all functions on the right-hand side extend holomorphically to D^* with

$$b_-(\infty) = 0$$

Therefore

$$a(\infty) = a_-(\infty) a_+(\infty) > 0$$

Moreover,

$$a = a_- a_+ \left(1 + \frac{b_-}{a_-} \frac{b_+^*}{a_+}\right)$$

and the first two factors on the right-hand side are outer. The last factor has positive real part on on D^* because the extensions of b_-/a_- and b_+^*/a_+ to D^* are bounded by 1. Thus the Herglotz representation theorem applies to the last factor, which then can be seen to be in $H^p(D^*)$ for all $p < 1$. The reciprocal of the last factor has also positive real part and is also in $H^p(D^*)$ for all $p < 1$. Thus the last factor is an outer function.

The proof of continuity invokes Lemma 1.17. Given F, we can use the independence of the cut in Definition (1.25) to cut F and any nearby F' at a very large integer N (depending only on F), so that the tail to the right of N of both F and F' is negligible by the Plancherel identity and Lemma 1.17. Then we can apply continuity of the nonlinear Fourier transform on the (shifted) half line to show that the parts of F and F' to the left of N have nearby nonlinear Fourier transforms.

The same argument also proves the Plancherel identity. $\square$

We now observe

LEMMA 1.24. *The nonlinear Fourier transform is not injective on $l^2(\mathbf{Z}, D)$.*

PROOF. We claim that

$$(a,b) = \left(\frac{2z}{z-1}, \frac{z+1}{z-1}\right)$$

is in $\mathbf{H} \cap \mathbf{H}^*$. Therefore it has nonzero preimages in $l^2(\mathbf{Z}_{\geq 0})$ and in $l^2(\mathbf{Z}_{\leq 0})$, and since these preimages are not finite sequences (a is not a Laurent polynomial), these two preimages are necessarily distinct members of $l^2(\mathbf{Z}, D)$.

It remains to prove the claim. We observe

$$a(z)a^*(z) - b(z)b^*(z)$$

$$= \frac{(2z)(2z^{-1})}{(z-1)(z^{-1}-1)} - \frac{(z+1)(z^{-1}+1)}{(z-1)(z^{-1}-1)}$$

$$= \frac{4}{-z+2-z^{-1}} - \frac{z+2+z^{-1}}{-z+2-z^{-1}} = 1$$

The function a is outer on D^* since it is in $H^p(D^*)$ for all $p < 1$ and its reciprocal is in $H^\infty(D^*)$. We also have $a(\infty) = 2 > 0$.

Moreover, both

$$\frac{b(z)}{a(z)} = \frac{z+1}{2z}$$

and

$$\frac{b^*(z)}{a(z)} = -\frac{z+1}{2z}$$

are holomorphic in D^* and in $H^2(D^*)$. This proves the claim. $\square$

We now discuss the inverse problem, i.e., finding a (sometimes not unique) sequence F whose nonlinear Fourier transform is a given (a, b).

Given data $(a, b) \in \mathbf{L}$, we need to factorize it

$$(a_-, b_-)(a_+, b_+) = (a, b)$$

with $(a_-, b_-) \in \mathbf{H}_0^*$ and $(a_+, b_+) \in \mathbf{H}$. Any such factorization is in bijective correspondence to a sequence $F \in l^2(\mathbf{Z}, D)$ whose truncations satisfy

$$\widehat{F_{\leq -1}} = (a_-, b_-)$$

$$\widehat{F_{\geq 0}} = (a_+, b_+)$$

Thus the inverse problem for the nonlinear Fourier transform is a matrix factorization problem with (mainly but not exclusively) holomorphicity conditions on the matrix factors.

Observe that the corresponding linear problem is the decomposition of a function $f \in L^2(\mathbf{T})$ as the sum of a function in the Hardy space H^2 and a function in the conjugate Hardy space $\overline{H}_0^2$ (where the index 0 stands for functions with mean zero).

Finding a factorization of a matrix function on $\mathbf{T}$ into a product of two matrix functions, one extending to D and one extending to D^* is called a Riemann-Hilbert problem.

Our factorization is a somewhat twisted Riemann-Hilbert problem, because the matrices both have entries which extend to D and D^*. Moreover, the factorization problem is constrained in that there are algebraic relations between the matrix entries and there is an outerness condition on a_-, a_+ and a normalization condition at ∞.

However, the factorization problem can be reduced to a more genuine Riemann-Hilbert problem by the following algebraic manipulations. The factorization equation together with the determinant condition can be rewritten as

$$\begin{pmatrix} {a_+}^* & -b_+ \\ b_-^* & a_-^* \end{pmatrix} \begin{pmatrix} a_+ & b_+ \\ b_+^* & a_+^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ b^* & a^* \end{pmatrix}$$

Here the second row comes from the factorization problem while the first row comes from

$$(a_+, b_+)^{-1}(a_+, b_+) = (1, 0)$$

Similarly we obtain

$$\begin{pmatrix} a_+^* & -b_+ \\ -b_+^* & a_+ \end{pmatrix} \begin{pmatrix} a_+ & -b_- \\ b_+^* & a_- \end{pmatrix} = \begin{pmatrix} 1 & -b \\ 0 & a \end{pmatrix}$$

Multiplying the two equations and using the determinant condition on (a, b) gives

$$\begin{pmatrix} {a_+}^* & -b_+ \\ {b_-}^* & {a_-}^* \end{pmatrix} \begin{pmatrix} a_+ & -b_- \\ {b_+}^* & a_- \end{pmatrix} = \begin{pmatrix} 1 & -b \\ b^* & 1 \end{pmatrix}$$

In the last equation, all entries of the first factor on the left-hand side extend to D while all entries of the second factor on the left-hand side extend to D^* (the function b_- is in addition required to vanish at ∞). The Riemann-Hilbert problem is still constrained in that the entries of the two matrices on the left are dependent, however the constraints can be subsumed in the statement that the factorization should be invariant under the map

$$T : \begin{pmatrix} a & b \\ c & d \end{pmatrix} \to \begin{pmatrix} a^* & -c^* \\ -b^* & d^* \end{pmatrix}$$

This map reverses the order of multiplication, $T(G)T(G') = T(G'G)$, and one can easily check that this symmetry produces all algebraic constraints between the two factors.

Observe that while a no longer appears explicitly in the factorization problem, a^* and a formally coincide with the determinants of the two factors and thus taking determinants everywhere we formally obtain the equation

$$a^* a = 1 + bb^*$$

Observe that for any solution of this Riemann-Hilbert problem we can produce another solution by multipying the first factor by a function c of modulus 1 on $\mathbf{T}$ with holomorphic extension to D and the second factor by c^*. To obtain any hope for uniqueness, one has to make the additional analytic assumption that say all functions of the first matrix are in $N^+(D)$ (as defined in the appendix) and all entries of the second matrix are in $N^+(D^*)$ and that the determinants of the two matrices are outer on D and D^* respectively. A sharper constraint is to require that the diagonal entries of the two factors are outer functions on D and D^* respectively.

Then the only obvious ambiguity left is a scalar factor, which can be normalized by requiring the determinants of both matrices to be positive at 0 and ∞ respectively.

1.12. Existence and uniqueness of an inverse NLFT for bounded a

We shall now prove existence and uniqueness of the solution to the factorization problem under the additional assumption that a is bounded. We shall use the original formulation rather than the more genuine Riemann-Hilbert problem.

LEMMA 1.25. *If $(a, b) \in \mathbf{L}$ and in addition a is a bounded function, then there is a unique $F_n \in l^2(\mathbf{Z}, D)$ such that*

$$\widehat{F_n} = (a, b)$$

PROOF. By the half-line theory it suffices to find and show uniqueness of a decomposition

$$(a_-, b_-)(a_+, b_+) = (a, b) \tag{1.26}$$

such that the factors on the left-hand side are in $\mathbf{H}_0^*$ and $\mathbf{H}$ respectively.

We first prove the following, which does not require a to be bounded.

LEMMA 1.26. *Let $(a, b) \in \mathbf{L}$. For any factorization of the Riemann-Hilbert problem*

$$(a_-, b_-)(a_+, b_+) = (a, b)$$

with $(a-, b-) \in \mathbf{H}_0^*$ and $(a_+, b_+) \in \mathbf{H}$, we have that a_-/a and a_+/a are functions in $H^2(D^*)$.

PROOF. As a_-/a and a_+/a are outer, it suffices to show that the boundary values of these functions on $\mathbf{T}$ are in L^2.

The Riemann-Hilbert problem gives

$$a = a_- a_+ [1 - (b_-/a_-)(b_+^*/a_+)]$$

or equivalently,

$$a_- a_+/a = [1 - (b_-/a_-)(b_+^*/a_+)]^{-1} \tag{1.27}$$

We first show that the real part of the right-hand side is in $L^1(\mathbf{T})$.

This function extends to D^* with positive real part, because the quotients b_-/a_- and b_+^*/a_+ are strictly bounded by 1 on D^*. By the Theorem 1.49 of Herglotz discussed in the appendix, the real part of (1.27) is the harmonic extension of a positive measure. Almost everywhere on $\mathbf{T}$, the real part of the function $a_- a_+/a$ coincides with the density of the absolutely continuous part of this measure and is thus in $L^1(\mathbf{T})$.

As

$$\mathrm{Re}(\frac{a_- a_+}{a}) + \mathrm{Re}(\frac{b_- b_+^*}{a}) = 1$$

from the Riemann-Hilbert factorization, we also have that

$$\mathrm{Re}\left[\frac{a_- a_+}{a} - \frac{b_- b_+^*}{a}\right] \tag{1.28}$$

is absolutely integrable.

The Riemann-Hilbert factorization can be rewritten in terms of the equations

$$(a_-, b_-) = (a, b)(a_+^*, -b_+)$$

$$(a_+, b_+) = (a_-^*, -b_-)(a, b)$$

These give

$$a_- = aa_+^* - bb_+^*$$
$$b_+ = a_-^* b - b_- a^*$$

Which in turn give

$$a_+^* = \frac{a_-}{a} + \frac{bb_+^*}{a}$$
$$b_- = -\frac{b_+}{a^*} + \frac{a_-^* b}{a^*}$$

Thus we can write for (1.28) on $\mathbf{T}$:

$$\mathrm{Re}\left[\frac{a_- a_-^*}{aa^*} + \frac{a_- b^* b_+}{aa^*} + \frac{b_+^* b_+}{aa^*} - \frac{b_+^* a_-^* b}{aa^*}\right]$$
$$= \mathrm{Re}\left[\frac{a_- a_-^*}{aa^*} + \frac{b_+^* b_+}{aa^*}\right]$$

In the last line we have cancelled two terms inside the brackets which added up to a purely imaginary quantity on $\mathbf{T}$. From integrability of the last line, we observe that

$$a_-/a \in L^2(\mathbf{T})$$

and also, since $a_+ a_+^* = 1 + b_+ b_+^*$ and $|a| > 1$ on $\mathbf{T}$,

$$a_+/a \in L^2(\mathbf{T})$$

This proves the lemma. □

We now prove the uniqueness part of Lemma 1.25.

By Lemma 1.26, it suffices to prove uniqueness under the additional assumption that a_+, b_+^*, a_-, b_- are in $H^2(D^*)$.

We rewrite the Riemann-Hilbert problem as

$$(a_-, b_-) = (a, b)(a_+{}^*, -b_+)$$

Since a is nonvanishing on $\mathbf{T}$, we can rewrite the second equation as

$$\frac{b_-}{a} = -b_+ + \frac{b}{a} a_+$$

Let P_D be the orthogonal projection from $L^2(T)$ to $H^2(D)$. Then the previous display implies

$$b_+ = P_D\left(\frac{b}{a} a_+\right)$$

since b_+ is already in $H^2(D)$ and b_-/a is in $H^2(D^*)$ with vanishing constant term.

Next, we have again from the Riemann-Hilbert factorization

$$\frac{a_-^*}{a^*} = a_+ - \frac{b^*}{a^*} b_+$$

Applying the orthogonal projection P_{D^*} from $L^2(T)$ to $H^2(D^*)$, we obtain

$$a_+ = \frac{a_-(\infty)}{a(\infty)} + P_{D^*}\left(\frac{b^*}{a^*} b_+\right)$$

Here we have used that a_+ is already in $H^2(D^*)$ and the quotient a_-^*/a^* is in $H^2(D)$ and thus its P_D projection is equal to its constant term, which is real and equal to $a_-(\infty)/a(\infty)$.

Observe that evaluating the extension of

$$a = a_- a_+ + b_- b_+^*$$

at ∞ gives

$$a(\infty) = a_-(\infty) a_+(\infty) + 0$$

Thus we can can rewrite the expression for a_+ as

$$a_+ = \frac{1}{a_+(\infty)} + P_{D^*}\left(\frac{b^*}{a^*} b_+\right)$$

For any constant c, the affine linear map

$$(A, B) \mapsto \left(c + P_{D^*}\left(\frac{b^*}{a^*} B\right), P_D\left(\frac{b}{a} A\right)\right)$$

is a contraction in $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ (Hilbert space sum). Namely, P_D and P_{D^*} have norm 1 in $L^2(\mathbf{T})$, while multiplication by b/a or b^*/a^* have norm strictly less than 1 in $L^2(\mathbf{T})$. Here we use that a is bounded and thus

$$\left|\frac{b}{a}\right| \leq \left|1 - \frac{1}{|a|^2}\right|^{1/2} \leq 1 - \epsilon$$

Therefore, by the contraction mapping principle, this map has a unique fixed point (A_c, B_c). Indeed, by linearity, this fixed point is (cA_1, cB_1).

From the above it follows that (a_+, b_+) is equal to this unique fixed point for some constant c. To prove uniqueness of (a_+, b_+), it therefore suffices to show that we can determine c uniquely.

The phase of the constant c is determined by the requirement

$$a_+(\infty) = cA_1(\infty) > 0$$

The modulus of c can then be determined by

$$cA_1(\infty) = \frac{1}{cA_1(\infty)} + P_D\left(\frac{b}{a} b_+^*\right)(\infty)$$

and thus

$$cA_1(\infty)^2 = 1 + P_D(\frac{b}{a} b_+^*)(\infty) cA_1(\infty)$$

The second summand on the right-hand side is necessarily real and positive since the left-hand side is larger than 1. This gives a quadratic equation for $cA_1(\infty)$ with a positive and a negative solution. Since $cA_1(\infty)$ is necessarily positive, it is therefore uniquely determined. Thus we can recover $(a_+, b_+) = (cA_1, cA_1)$ completely from (a, b). By matrix division, we also obtain (a_-, b_-). Thus the solution to the Riemann Hilbert problem is unique.

It remains to prove existence of the solution to the Riemann-Hilbert problem. In the next section, we will prove existence without assuming boundedness of a. However, the proof of existence for bounded a is much easier. Therefore we choose to present it here.

Consider again the above fixed point equation and let $(A, B) \in H^2(D^*) \oplus H^2(D)$ be the unique solution for $c = 1$.

Observe that by interpolation, the linear map is also a contraction on $H^{2+\epsilon} \oplus H^{2+\epsilon}$ for small ϵ. Namely, the map is bounded in any space $H^p \oplus H^p$ with $2 < p < \infty$. The operator norm may be large for any fixed p, but interpolating this estimate with the estimate for $H^2 \oplus H^2$, where the operator norm is less than 1, gives for sufficiently small ϵ an operator norm on $H^{2+\epsilon} \oplus H^{2+\epsilon}$ which is still less than 1. Hence the unique solution in $H^2 \oplus H^2$ is actually in the subspace $H^{2+\epsilon} \oplus H^{2+\epsilon}$, since this subspace also contains a solution by the contraction mapping principle. Hence the regularity of the solution to the Riemann-Hilbert problem will be slightly better than Lemma 1.26 suggests. We shall not need this extra regularity in the current proof.

We claim that the function

$$AA^* - BB^*$$

is constant on $\mathbf{T}$. Since it is manifestly real, it suffices to show that it is in the Hardy space $H^1(D)$ (being in a Hardy space H^p with $p \geq 1$ makes the linear Fourier coefficients supported on a half-line, while being real makes the moduli of the Fourier coefficients symmetric about 0). Use the fixed point equation to write the function as

$$= A[1 + P_D(\frac{b}{a}B^*)] + B^* P_D(\frac{b}{a}A)$$

$$= A[1 - (\mathrm{id} - P_D)(\frac{b}{a}B^*)] + B^*(\mathrm{id} - P_D)(\frac{b}{a}A)$$

Here the two terms involving the identity operators that have artificially been inserted are negatives of each other. Observe that $\mathrm{id} - P_D$ is projecting onto the space $H_0^2(D^*)$ of functions in the Hardy space $H^2(D^*)$ with vanishing constant coefficient.

Thus the entire last displayed expression is an element in $H^1(D^*)$, since it is the sum of products of functions in $H^2(D^*)$.

Moreover, we observe that the constant coefficient of this expression is that of A:

$$\int_{\mathbf{T}} AA^* - BB^* = \int_{\mathbf{T}} A = A(\infty)$$

Thus the constant coefficient of A is real. Indeed, it is positive, as we see from the following calculation:

$$\int_{\mathbf{T}} AA^* + BB^*$$

$$= \int_{\mathbf{T}} A(1 + P_D(\frac{b}{a}B^*)) + B^* P_D(\frac{b}{a}A)$$

$$= \int_{\mathbf{T}} A(1 + \frac{b}{a}B^*) + B^* \frac{b}{a}A$$

In the last line we have dropped the projection operators, because the operands are integrated against functions in the Hardy space $H^2(D^*)$. However, estimating the last display using $|b/a| \leq 1$ on $\mathbf{T}$, we obtain:

$$\int_{\mathbf{T}} AA^* + BB^* \leq \int_{\mathbf{T}} A + 2\int_{\mathbf{T}} |A||B|$$

or

$$\int_{\mathbf{T}} (|A| - |B|)^2 \leq A(\infty)$$

Thus the constant coefficient of A is nonnegative.

Indeed, in the above string of inequlities, identity holds only if $|A| = |B| = 0$ almost everywhere on $\mathbf{T}$, since $|b/a|$ is strictly less than 1 almost everywhere. However, $A = B = 0$ is inconsistent with the fixed point equation. Therefore, we have strict inequality and the constant coefficient of A is strictly positive.

Set

$$a_+(z) := A(z)[A(\infty)]^{-1/2}$$

$$b_+(z) := B(z)[A(\infty)]^{-1/2}$$

Then

$$|a_+|^2 - |b_+|^2 = 1$$

almost everywhere on $\mathbf{T}$ and $a_+ \in H^2(D^*)$ and $b_+ \in H^2(D)$.

Now we can define a_- and b_- by

$$(a_-, b_-) := (a, b)(a_+{}^*, -b_+)$$

but to complete the proof we need to show that $(a_-, b_-) \in \mathbf{H}_0^*$. We also need to show that a_+ is outer.

Clearly we have $a_- a_-^* = 1 + b_- b_-^*$ almost everywhere on $\mathbf{T}$ since the other matrices in the equation are $SU(1,1)$ almost everywhere.

Next we check that a_- and b_- have the correct holomorphicity properties. From the fixed point equations,

$$a_- = aa_+^* - bb_+^*$$

$$= a\frac{1}{a_+(\infty)} + aP_D(\frac{b}{a}b_+{}^*) - bb_+^*$$

$$= a\frac{1}{a_+(\infty)} - a(\mathrm{id} - P_D)(\frac{b}{a}b_+{}^*)$$

Clearly this is an element of $H^2(D^*)$. Moreover, the constant term obeys

$$a_-(\infty) = \frac{a(\infty)}{a_+(\infty)}$$

and so is positive as required.

Similarly, using the fixed point equation for b_+,

$$b_- = -ab_+ + ba_+$$

$$= -aP_D(\frac{b}{a}a_+) + ba_+$$

$$= (1 - P_D)(\frac{b}{a}a_+)$$

Thus b_- is in $H^2(D^*)$

To prove that we have indeed solved the Riemann-Hilbert problem, we have to verify that a_- and a_+ are outer.

Consider the equation

$$a = a_- a_+ + b_- b_+^*$$

Every function in this equation is holomorphic in D^*. We divide by the outer function a to obtain

$$1 = a_- a_+ a^{-1} + b_- b_+{}^* a^{-1}$$

since the first summand on the right is larger in modulus on $\mathbf{T}$ than the second, we conclude that

$$\mathrm{Re}(a_- a_+ a^{-1}) \geq 1/2$$

almost everywhere on $\mathbf{T}$. This implies that the function $a_- a_+ a^{-1}$, which is in $H^1(D^*)$ and thus equal to its Poisson integral on D^*, has real part larger than $1/2$ on D^*. Then the reciprocal function $\frac{a}{a_- a_+}$ is in $H^\infty(D^*)$ and

$$\frac{1}{a_+} = \left(\frac{a}{a_- a_+}\right)\left(\frac{a_-}{a_+}\right)$$

is in $H^2(D^*)$ by Lemma 1.26. By Lemma 1.54, a_+ is outer. Likewise one concludes that a_- is outer.

This completes the proof of Lemma 1.25. □

1.13. Existence of an inverse NLFT for unbounded a

In this section we prove that for every $(a, b) \in \mathbf{L}$, there exists a factorization

$$(a_-, b_-)(a_+, b_+) = (a, b)$$

with $(a_-, b_-) \in \mathbf{H}_0^*$ and $(a_+, b_+) \in \mathbf{H}$. We shall call such a factorization a Riemann-Hilbert factorization. This factorization is not necessarily unique. If it is unique, then we say that (a, b) has a unique Riemann-Hilbert factorization.

In the previous section we used the Banach fixed point theorem to produce a Riemann-Hilbert factorization when $a \in H^\infty(D^*)$. This same approach does not work in the general case; we shall instead use the Riesz representation theorem for linear functionals on a Hilbert space. In general there will be several choices of Hilbert space to work with, which will cause non-uniqueness of the Riemann-Hilbert factorization.

We introduce two examples of such Hilbert spaces, which in general may be different and will turn out to be extremal examples. They are vector spaces over the real (not complex) numbers. Given $(a, b) \in \mathbf{L}$, we consider the following inner product on pairs of measurable functions on $\mathbf{T}$:

$$\langle (A', B'), (A, B)\rangle := \int_{\mathbf{T}} \mathrm{Re}[A'(A^* - \frac{b}{a}B^*) + (B')^*(B - \frac{b}{a}A)] \tag{1.29}$$

whenever the integral on the right-hand side is absolutely integrable. We emphasize that absolute integrability is only required for the real part of the algebraic expression in the integrand.

This inner product is positive definite, as we see from the following calculation:

$$\begin{aligned}
\|(A, B)\| := \langle (A, B), (A, B)\rangle &\geq \int_{\mathbf{T}} |A|^2 + |B|^2 - 2|b/a||A||B| \\
&\geq \int_{\mathbf{T}} |b/a|(|A| - |B|)^2 + \int (1 - |b/a|)(|A|^2 + |B|^2) \\
&\geq \frac{1}{2}\int_{\mathbf{T}} (1 - |b/a|^2)(|A|^2 + |B|^2) \\
&= \frac{1}{2}\int_{\mathbf{T}} (|A|^2 + |B|^2)|a|^{-2} \geq 0
\end{aligned}$$

Equality holds in the last estimate if and only if A and B vanish almost everywhere on $\mathbf{T}$. Therefore the inner product is positive definite. Moreover, we have seen that the integrand in (1.29) is nonnegative almost everywhere if $(A, B) = (A', B')$.

The above calculation also shows that a necessary condition for for the inner product (1.29) to be defined is $A/a \in L^2(\mathbf{T})$ and $B/a \in L^2(\mathbf{T})$.

Define $H_{\max}$ to be the space of all pairs (A, B) such that $A/a \in H^2(D^*)$ and $B/a^* \in H^2(D)$ and $\|(A, B)\| < \infty$ with respect to the inner product (1.29). This space is evidently a pre-Hilbert space with inner product (1.29). It is indeed a Hilbert space, because for any Cauchy sequence, the boundary values of A/a and B/a^* converge in $L^2(\mathbf{T})$ and thus remain in $H^2(D^*)$ and $H^2(D)$ respectively. By an application of Fatou's lemma, the limit has again finite norm and thus is in $H_{\max}$. The previous display shows that $H_{\max}$ is continuously embedded in $aH^2(D^*) \times a^*H^2(D)$.

The space $H^2(D^*) \times H^2(D)$ is contained in $H_{\max}$, because

$$\langle (A, B), (A, B)\rangle \le 2\int_{\mathbf{T}} |A|^2 + |B|^2$$

Define $H_{\min}$ to be the closure of $H^2(D^*) \times H^2(D)$ in $H_{\max}$. As we will see, there are examples of data (a, b) for which the space $H_{\min}$ is strictly contained in $H_{\max}$.

We now introduce the real linear functional to which the Riesz representation theorem will be applied. It takes the same form on $H_{\max}$ and $H_{\min}$ and is given by

$$\lambda : (A, B) \to \mathrm{Re}[A(\infty)]$$

Observe that this linear functional is indeed continuous on $H_{\max}$ and thus also $H_{\min}$, since it is even continuous on the larger space $aH^2(D^*) \times a^*H^2(D)$, as can be seen immediately from

$$\mathrm{Re}A(\infty) = a(\infty)\mathrm{Re}\int_{\mathbf{T}} A/a$$

Let $(A_{\min}, B_{\min})$ be the unique element in $H_{\min}$ which produces this linear functional in $H_{\min}$ and is guaranteed to exist by the Riesz representation theorem:

$$\langle (A_{\min}, B_{\min}), (A, B)\rangle = \lambda(A, B)$$

for all $(A, B) \in H_{\min}$. Let $(A_{\max}, B_{\max})$ be the unique element in $H_{\max}$ which produces this linear functional in $H_{\max}$.

THEOREM 1.27. *Let $(a, b) \in \mathbf{L}$. Then there exists a factorization*

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

with $(a_-, b_-) \in \mathbf{H}_0^*$ and $(a_+, b_+) \in \mathbf{H}$. Moreover, with $(A_{\max}, B_{\max})$ and $(A_{\min}, B_{\min})$ defined as above, two possible choices of such a Riemann-Hilbert factorization are given by

$$(a_+, b_+) := (A_{\min}, B_{\min})A_{\min}(\infty)^{-1/2} \tag{1.30}$$

and

$$(a_+, b_+) := (A_{\max}, B_{\max})A_{\max}(\infty)^{-1/2} \tag{1.31}$$

with the corresponding (a_-, b_-), which are easily determined by matrix division.

PROOF. The proofs that (1.30) and (1.31) give Riemann Hilbert factorizations are very similar. We shall formulate the proof for (1.30) and comment on the changes needed to prove (1.31).

Define L to be the space of all pairs (A, B) of measurable functions such that $A/a \in H^2(D^*)$, $B/a^* \in L^2(\mathbf{T})$, and $\|(A, B)\| < \infty$.

We claim that $H^2(D^*) \times L^2(\mathbf{T})$ is dense in L. Indeed, let $(A,B) \in L$ be orthogonal to all elements of $H^2(D^*) \times L^2(\mathbf{T})$. Choosing $A' = 0$ and B' of modulus one such that $(B')^*(B - \frac{b}{a}A)$ is nonnegative real, we have

$$0 = \langle (A', B'), (A, B)\rangle = \int |B - \frac{b}{a}A|$$

and thus

$$B - \frac{b}{a}A = 0 \tag{1.32}$$

almost everywhere. Now choosing $A' \in H^2$ arbitrary and $B' = 0$ we obtain

$$0 = \langle (A', B'), (A, B)\rangle$$

$$= \int_{\mathbf{T}} A'(A^* - \frac{b}{a}B^*) = \int_{\mathbf{T}} A'A^*(1 - |b/a|^2) = \int_{\mathbf{T}} \frac{A'}{a}\frac{A^*}{a^*}$$

By Beurling's theorem (see [**16**]), since a is outer, the set of all A'/a with $A' \in H^2(D^*)$ is dense in $H^2(D^*)$. Thus, redefining A',

$$\int_{\mathbf{T}} A' \frac{A^*}{a^*} = 0$$

for all $A' \in H^2(D^*)$. But $A^*/a^* \in H^2(D)$, thus $A^*/a^* = 0$. Hence $A = B = 0$ by (1.32) and we have shown that the orthogonal complement of $H^2(D^*) \times L^2(\mathbf{T})$ is trivial, thus proving our claim.

Define H_n to be the closure of $H^2(D^*) \times z^n H^2(D)$ in L, in particular $H_0 = H_{\min}$. (Here is the main difference in proving the theorem for $(A_{\min}, B_{\min})$ and $(A_{\max}, B_{\max})$. To prove the theorem for $(A_{\max}, B_{\max})$, one would need to define H_n to be the space of all $(A,B) \in L$ such that $B/a^* \in z^n H^2(D)$.)

Since evaluation of $z^{-n}B$ at 0 is a continuous functional on H_n,

$$z^{-n}B = a^*(0) \int_{\mathbf{T}} z^{-n}B/a^*$$

we see that H_{n+1} is precisely the subspace of H_n of all (A,B) such that $z^{-n}B$ vanishes at 0. Thus H_{n+1} has real co-dimension two in H_n.

Let H_∞ be the intersection of all H_n for $n \in \mathbf{Z}$, then it is clear that H_∞ consists of pairs (A,B) such that B vanishes to infinite order at 0 and thus is identically equal to 0. Finiteness of the norm of (A,B) is then equivalent to $A \in H^2(D^*)$ and thus evidently $H_\infty = H^2(D^*) \times \{0\}$.

Let $H_{-\infty}$ be the closure of the union of all H_n for $n \in \mathbf{Z}$. Since every element of $H^2(D^*) \times L^2(\mathbf{T})$ can be approximated by a sequence of elements in spaces H_n with decreasing n, we see that $H_{-\infty}$ is equal to the closure of $H^2(D^*) \times L^2(\mathbf{T})$ which is all of L.

Let (A_n, B_n) be the element which represents the linear functional λ in the subspace H_n. It is easy to see that

$$(A_\infty, B_\infty) = (1, 0)$$

$$(A_{-\infty}, B_{-\infty}) = a(\infty)(a, b)$$

Observe that the operation $(A,B) \to (B^*z^n, A^*z^n)$ is a bijection on the space $H^2(D^*) \times z^n H^2(D)$, and extends to a bijective isometry of H_n.

We claim

$$(A_{n+1}, B_{n+1}) = (A_n, B_n) - F_n(B_n^* z^n, A_n^* z^n) \tag{1.33}$$

for a certain complex number $F_n \in D$.

Indeed, since λ is non-zero (it is so on H_∞), we have

$$\mathrm{Re} A_n(\infty) = \langle (A_n, B_n), (A_n, B_n)\rangle > 0 \tag{1.34}$$

and there is a unique $F_n \in \mathbf{C}$ such that the off-diagonal entry on the right-hand side of (1.33) vanishes to order $n+1$ at 0. For later reference we pause to argue that taking real part on the left-hand side of (1.34) is superfluous since $A_n(\infty)$ itself is positive. Namely, $(A, B) \to (cA, cB)$ is an isometry of H_n for $|c| = 1$ and since

$$\mathrm{Re}[cA_n] = \langle (A_n, B_n), (cA_n, cB_n)\rangle$$

is maximized for $c = 1$, we have that $A_n(\infty) > 0$.

Now we observe that

$$(B_n^* z^n, A_n^* z^n)$$

is orthogonal to H_{n+1}. Namely, let $(A, B) \in H_{n+1}$, then

$$\langle (B_n^* z^n, A_n^* z^n), (A, B)\rangle = \langle (A_n, B_n), (B^* z^n, A^* z^n)\rangle = \lambda(B^* z^n, A^* z^n) = 0$$

Thus the right-hand side of (1.33) is the orthogonal projection of (A_n, B_n) onto H_{n+1} and thus indeed equal to (A_{n+1}, B_{n+1}).

We now verify that $|F_n| < 1$. This simply follows from the fact that

$$\|(B_n^* z^n, A_n^* z^n)\| = \|(A_n, B_n)\|$$

and the fact that the terms in (1.33) form a Pythagorean triple and $(A_{n+1}, B_{n+1}) \neq 0$. Another consequence is that

$$\|(A_{n+1}, B_{n+1})\| = (1 - |F_n|^2)^{1/2} \|(A_n, B_n)\| \tag{1.35}$$

Each vector (A_n, B_n) is the orthogonal projection of $(A_{-\infty}, B_{-\infty})$ onto H_n, and the projection of (A_n, B_n) onto H_∞ is (A_∞, B_∞). Thus the length of each vector (A_n, B_n) is squeezed between two finite numbers

$$\|(A_{-\infty}, B_{-\infty})\| \geq \|(A_n, B_n)\| \geq \|(A_\infty, B_\infty)\|$$

By an inductive argument using (1.35) we see that

$$\prod_n (1 - |F_n|^2)^{1/2}$$

is a convergent product and thus the sequence $F = (F_n)$ is in $l^2(\mathbf{Z}, D)$.

Let $(\tilde{a}, \tilde{b})$ be the nonlinear Fourier transform of F. We need to show that $(a, b) = (\tilde{a}, \tilde{b})$. We claim that

$$(A_n, B_n) \prod_{k \geq n} (1 - |F_n|^2)^{1/2}$$

is the nonlinear Fourier transform $(\tilde{a}^{(\geq n)}, \tilde{b}^{(\geq n)})$ of the truncated sequence $F^{(\geq n)}$.

Consider

$$((\tilde{a}^{(\geq n)})^*, -\tilde{b}^{(\geq n)})(A_n, , B_n) \prod_{k \geq n} (1 - |F_k|^2)^{1/2} \tag{1.36}$$

where we have used the convention to read each vector as the first row of a positive scalar multiple of a $SU(1,1)$ matrix.

Observe that (1.33) reads as

$$(A_{n+1}, B_{n+1}) = (1, -F_n z^n)(A_n, B_n)$$

Therefore, by the recursion equation for $(a^{(\geq n)}, b^{(\geq n)})$, the quantity (1.36) is independent of the parameter n.

There is an increasing sequence n_k of integers such that pointwise almost everywhere on $\mathbf{T}$ we have

$$(\tilde{a}^{(\geq n_k)}, \tilde{b}^{(\geq n_k)}) \to (1,0)$$

$$(A_{n_k}, B_{n_k}) \to (A_\infty, B_\infty) = (1,0)$$

as $k \to \infty$. For the first limit this follows from convergence of the sequence $F^{(\geq n)}$ to 0 in l^2 and thus convergence of $\log|\tilde{a}^{(\geq n)}|$ in L^1, convergence of the phase $\tilde{a}^{(\geq n)}/|\tilde{a}^{(\geq n)}|$ in L^2, and convergence of $\tilde{b}^{(\geq n)}/\tilde{a}^{(\geq n)}$ in L^2. For the second limit this follows from convergence of A_n/a and B_n/a^* in L^2. Thus (1.36) is equal to $(1,0)$ almost everywhere. Taking a similar limit as $n \to -\infty$ we observe

$$(\tilde{a}^*, -\tilde{b})(A_{-\infty}, B_{-\infty}) \prod_k (1-|F_k|^2)^{1/2} = (1,0)$$

This proves that (a,b) is a positive scalar multiple of $(\tilde{a}, \tilde{b})$, but since both are $SU(1,1)$ valued they are indeed equal.

Finally, we observe from splitting the sequence F as $F^{(<0)}+F^{(\geq 0)}$ that $(\tilde{a}^{(\geq 0)}, \tilde{b}^{(\geq 0)})$ is the right factor of a Riemann-Hilbert factorization of (a,b), and that we have

$$(A_{\min}, B_{\min}) = \tilde{a}_0(\infty)(\tilde{a}_0, \tilde{b}_0)$$

This completes the proof that (1.30) produces a Riemann-Hilbert factorization.

The proof for $(A_{\max}, B\max)$ is similar with changes as indicated above. $\square$

The above construction of a Riemann-Hilbert factorization easily provides the following strengthening. Let $(a,b) \in \mathbf{L}$ and

$$(a,b) = (a_-, b_-)(a_+, b_+)$$

be any Riemann-Hilbert factorization. Then the vector $(A_{\min}, B_{\min})$ constructed as above with respect to (a,b) is identical to the vector $(A_{+,\min}, B_{+,\min})$ constructed as above with respect to (a_+, b_+). To see this, it suffices to show that the inner products

$$\langle (A',B'),(A,B)\rangle := \int_{\mathbf{T}} \mathrm{Re}[A'(A^* - \frac{b}{a}B^*) + (B')^*(B - \frac{b}{a}A)]$$

$$\langle (A',B'),(A,B)\rangle_+ := \int_{\mathbf{T}} \mathrm{Re}[A'(A^* - \frac{b_+}{a_+}B^*) + (B')^*(B - \frac{b_+}{a_+}A)]$$

coincide on the space $H^2(D^*) \times H^2(D)$. Indeed, by polarization it suffices to show this for $(A,B) = (A',B')$. However, the difference of these inner products is then given by

$$\mathrm{Re} \int_{\mathbf{T}} 2[\frac{b}{a} - \frac{b_+}{a_+}]AB^*$$

We have

$$b_- = -ab_+ + ba_+$$

$$\frac{b_-}{aa_+} = \frac{b}{a} - \frac{b_+}{a_+}$$

Observe that here the right-hand side is bounded on $\mathbf{T}$, while the left-hand side is in $H^2_0(D^*)$ and thus in $H^\infty_0(D^*)$. The difference of the inner products is then given by

$$\mathrm{Re} \int_{\mathbf{T}} 2\frac{b_-}{aa_+}AB^*$$

Since A and B^* are in $H^2(D^*)$, this difference is equal to 0.

THEOREM 1.28. *Let $(a,b) \in \mathbf{L}$. Then there is a unique factorization*

$$(a,b) = (a_{--}, b_{--})(a_{\mathbf{o}}, b_{\mathbf{o}})(a_{++}, b_{++})$$

such that

$$(a_{--}, b_{--}) \in \mathbf{H}^*_0$$

$$(a_{\mathbf{o}}, b_{\mathbf{o}}) \in \mathbf{H}^*_0 \cap \mathbf{H}$$

$$(a_{++}, b_{++}) \in \mathbf{H}$$

and (a_{--}, b_{--}) and (a_{++}, b_{++}) do not have any Riemann-Hilbert factorizations other than

$$(a_{++}, b_{++}) = (1,0)(a_{++}, b_{++})$$

and

$$(a_{--}, b_{--}) = (a_{--}, b_{--})(1,0)$$

Moreover, we have the sub-factorization property: Any Riemann-Hilbert factorization

$$(a,b) = (a_-, b_-)(a_+, b_+) \tag{1.37}$$

comes with further (obviously unique) Riemann-Hilbert factorizations

$$(a_-, b_-) = (a_{--}, b_{--})(a_{-\mathbf{o}} b_{-\mathbf{o}})$$

$$(a_+, b_+) = (a_{\mathbf{o}+}, b_{\mathbf{o}+})(a_{++} b_{++})$$

where (a_{--}, b_{--}) and (a_{++}, b_{++}) are as above and

$$(a_{-\mathbf{o}} b_{-\mathbf{o}}), (a_{\mathbf{o}+}, b_{\mathbf{o}+}) \in \mathbf{H}^*_0 \cap \mathbf{H}$$

We have

$$(a_{\mathbf{o}}, b_{\mathbf{o}}) = (a_{-\mathbf{o}} b_{-\mathbf{o}})(a_{\mathbf{o}+}, b_{\mathbf{o}+}) \tag{1.38}$$

The passage from the Riemann-Hilbert factorization (1.37) of (a,b) to the Riemann-Hilbert factorization (1.38) of $(a_{\mathbf{o}}, b_{\mathbf{o}})$ constitutes a bijective correspondence between the Riemann-Hilbert factorizations of (a,b) and the Riemann-Hilbert factorizations of $(a_{\mathbf{o}}, b_{\mathbf{o}})$.

Before proving the theorem, we prove the following lemma.

LEMMA 1.29. *If $(a,b) \in \mathbf{H}$ and*

$$(a,b) = (a_-, b_-)(a_+, b_+)$$

is a Riemann-Hilbert factorization, then $(a_-, b_-) \in \mathbf{H}$. Conversely, if

$$(a_-, b_-) \in \mathbf{H}^*_0 \cap \mathbf{H}$$

$$(a_+, b_+) \in \mathbf{H}$$

then

$$(a,b) := (a_-, b_-)(a_+, b_+) \in \mathbf{H}$$

Observe that by reflection there is an analogous lemma with $(a,b) \in \mathbf{H}_0^*$ and $(a_+,b_+) \in \mathbf{H}_0^*$.

Proof. Assume we have a Riemann-Hilbert factorization of $(a,b) \in \mathbf{H}$. Then

$$b_+ = a_-^* b - b_- a^*$$

$$\frac{b_-}{a_-^*} = \frac{b}{a^*} - \frac{b_+}{a_-^* a^*} \tag{1.39}$$

Every summand on the right-hand side is in $H^2(D)$, hence so is the expression on the left-hand side. This proves the first statement of the lemma.

Next, assume

$$(a_-,b_-) \in \mathbf{H}_0^* \cap \mathbf{H}$$
$$(a_+,b_+) \in \mathbf{H}$$

Clearly the product (a,b) is in $\mathbf{L}$ by Lemma 1.23. Then $b/a^* \in \mathbf{H}(D)$ follows again from (1.39). This proves the second statement of the lemma. □

Now we can prove Theorem 1.28.

Proof. With the notation of Theorem 1.27, set

$$(a_{++},b_{++}) = A_{\min}(\infty)^{-1/2}(A_{\min},B_{\min})$$

As we have observed in the discussion prior to the statement of Theorem 1.28, for any Riemann-Hilbert factorization

$$(a,b) = (a_-,b_-)(a_+,b_+)$$

we have a Riemann-Hilbert factorization

$$(a_+,b_+) = (a_{\mathbf{o}+},b_{\mathbf{o}+})(a_{++},b_{++})$$

The lemma just shown implies that

$$(a_{\mathbf{o}+},b_{\mathbf{o}+}) \in \mathbf{H}_0^* \cap \mathbf{H}$$

Thus we have shown the sub-factorization property for (a_{++},b_{++}).

By the sub-factorization property, (a_{++},b_{++}) is the only possible right factor in a Riemann-Hilbert factorization of (a,b) which does not have a Riemann-Hilbert factorization other than identity times itself. We claim that (a_{++},b_{++}) indeed does not have any further Riemann-Hilbert factorization. Assume we have a Riemann-Hilbert factorization

$$(a_{++},b_{++}) = (\tilde{a},\tilde{b})(a_{+++},b_{+++})$$

Then by an application of Lemma 1.29 we observe that (a_{+++},b_{+++}) is a right factor of a Riemann-Hilbert factorization of (a,b), and thus by the sub-factorization property

$$(a_{+++},b_{+++}) = (\tilde{a}^*,-\tilde{b})(a_{++},b_{++})$$

is also a Riemann-Hilbert factorization. Thus both $(\tilde{a},\tilde{b})$ and its inverse are in $\mathbf{H} \cap \mathbf{H}_0^*$. Thus $\tilde{b}/\tilde{a}$ is in $H^2(D) \cap H_0^2(D^*)$. Therefore $\tilde{b}/\tilde{a} = 0$ and consequently $(\tilde{a},\tilde{b}) = (1,0)$. This proves that (a_{++},b_{++}) does not have any nontrivial Riemann-Hilbert factorization.

Since by Lemma 1.29 any triple factorization of (a,b) as in the Theorem gives a Riemann-Hilbert factorization

$$(a,b) = [(a_{--},b_{--})(a_{\mathbf{o}},b_{\mathbf{o}})][(a_{++},b_{++})]$$

we have uniquely identified the factor (a_{++}) as the only possible in such a triple factorization. Similarly we can uniquely identify the factor (a_{--}, b_{--}), and by an application of the sub-factorization property we actually obtain the triple factorization from knowledge of these two factors.

Finally, we observe that any Riemann-Hilbert factorization of (a, b) gives a Riemann-Hilbert factorization of $(a_{\mathbf{o}}, b_{\mathbf{o}})$ as described in the theorem, and vice versa every Riemann-Hilbert factorization of $(a_{\mathbf{o}}, b_{\mathbf{o}})$ necessarily has two factors in $\mathbf{H}_0^* \cap \mathbf{H}$ by Lemma 1.29 and therefore comes from a Riemann-Hilbert factorization of (a, b). This proves the theorem. $\square$

For any $(a, b) \in \mathbf{L}$ call $\log |a(\infty)|$ the energy of (a, b). The energy of (a, b) is a nonnegative real number. Indeed, it is positive unless $(a, b) = (1, 0)$. If

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

is a Riemann-Hilbert factorization, then we have additivity of the energies

$$\log |a(\infty)| = \log |a_+(\infty)| + \log |a_-(\infty)|$$

(evaluate $a = a_- a_+ + b_- b_+^*$ at ∞).

The sub-factorization property shows that the right factor (a_{++}, b_{++}) of the triple factorization minimizes the energy among all right factors of Riemann-Hilbert factorizations of (a, b), and indeed is a unique minimizer. Since (a_{++}, b_{++}) was constructed through a minimal Hilbert space $H_{\min}$ in Theorem 1.27, it is natural to guess that the solution constructed from the space $H_{\max}$ in Theorem 1.27 maximizes the energy. This is the main content of the following lemma:

Lemma 1.30. *With the notation of Theorems 1.27 and 1.28 we have*

$$(a_{\mathbf{o}}, b_{\mathbf{o}})(a_{++}, b_{++}) = (A_{\max}, B_{\max}) A_{\max}(\infty)^{-1/2}$$

Proof. Consider the notation of Theorem 1.27 and the space $H_{\max}$. We claim that for every Riemann-Hilbert factorization

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

we have

$$(a_+, b_+) \in H_{\max}$$

By Lemma 1.26, a_+/a and b_+^*/a are in $H^2(D^*)$. Therefore it suffices to show that $\|(a_+, b_+)\|$ is finite. However, we have

$$\begin{aligned}\|(a_+, b_+)\| &= \int \mathrm{Re}[a_+(a_+^* - \frac{b}{a} b_+^*) + b_+^*(b_+ - \frac{b}{a} a_+^*)] \\ &= \int \mathrm{Re}[\frac{a_+ a_-}{a} - \frac{b_+^* b_-}{a}]\end{aligned}$$

and the right-hand side has been shown to be finite in the proof of Lemma 1.26. Therefore, $(a_+, b_+) \in H_{\max}$.

We have the quantitative estimate

$$\begin{aligned}\|(a_+, b_+)\| &= -1 + 2\int \mathrm{Re}[\frac{a_+ a_-}{a}] \\ &\leq -1 + 2a_+(\infty) a_-(\infty) a(\infty)^{-1} = 1\end{aligned}$$

In the inequality we have used the observation in the proof of Lemma 1.26 that $a_- a_+/a$ has positive real part, and thus the real part of its value at ∞ is the total mass of the positive measure given by the Herglotz representation theorem. The total mass of this measure dominates the total mass of its absolutely continuous part $\mathrm{Re}(a_+ a_-/a)$.

Applying the linear functional λ to $(a_+, b_+) \in H_{\max}$ gives

$$\begin{aligned} a_+(\infty) &= \langle (a_+, b_+), (A_{\max}, B_{\max}) \rangle \\ &\leq \|(a_+, b_+)\| \|A_{\max}, B_{\max})\| \\ &\leq A_{\max}(\infty)^{1/2} \end{aligned}$$

This proves that (a_+, b_+) has smaller energy than the solution $(A_{\max}, B_{\max}) A_{\max}(\infty)^{-1/2}$ of the Riemann-Hilbert factorization theorem. Thus $(A_{\max}, B_{\max}) A_{\max}(\infty)^{-1/2}$ maximizes the energy of the right factor of a Riemann-Hilbert factorization of (a, b). On the other hand, by symmetry (a_{--}, b_{--}) (uniquely) minimizes the energy of the left factor of a Riemann-Hilbert factorization, just as (a_{++}, b_{++}) minimizes the energy of the right factor. By additivity of the energy this proves the lemma. $\square$

1.14. Rational functions as Fourier transform data

1.15. The Riemann-Hilbert problem for rational functions

The class $\mathbf{L}$ of nonlinear Fourier transform data of l^2 sequences contains elements (a, b) such that a and b are rational functions in z. We call (a, b) rational if a and b are rational.

Indeed, any pair (a, b) of rational functions is an element of $\mathbf{L}$ if $aa^* = 1 + bb^*$, $a(\infty) > 0$, and the function a has no zeros and poles in D^*. These pairs can easily be parameterized by the function b, as the following lemma states:

LEMMA 1.31. *For each rational function b there is precisely one rational function a such that $aa^* = 1 + bb^*$, a has no zeros and poles in D^*, and $a(\infty) > 0$. This is the unique function a such that $(a, b) \in \mathbf{L}$.*

For rational $(a, b) \in \mathbf{L}$, we have $(a, b) \in \mathbf{H}$ if and only if b has no poles in D, and $(a, b) \in \mathbf{H}_0$ if and only if in addition $b(0) = 0$. Likewise, we have $(a, b) \in \mathbf{H}^*$ if and only if b has no poles in D^* and we have $(a, b) \in \mathbf{H}_0^*$ if and only if in addition $b(\infty) = 0$.

PROOF. Let b be a rational function. Consider the rational function $g = 1 + bb^*$. Then $g(z) = g^*(z)$ and the zeros and poles of g are symmetric about $\mathbf{T}$: if z is a pole of order n, then so is z^* and likewise for the zeros. Moreover, there are no zeros of g on $\mathbf{T}$ and the poles of g on $\mathbf{T}$ are of even order.

Let a be a rational function whose zeros and poles in D are precisely the zeros and poles of g in D with the same order, whose poles on $\mathbf{T}$ are precisely the poles of g but with half the order, and which has no zeros on $\mathbf{T} \cup D^*$ and no poles on D^*. Thus poles and zeros of a are completely specified and a is determined up to a scalar factor. We assume a to be positive at ∞, which determines the phase of this scalar factor.

Consider

$$f = (aa^*)^{-1}(1 + bb^*)$$

Then this rational function evidently has no zeros and no poles and therefore it is constant. Since it is positive on $\mathbf{T}$, we may normalize a with a positive factor such that $f = 1$.

We claim that $(a, b) \in \mathbf{L}$. Certainly $aa^* = 1 + bb^*$ by construction. The function a is holomorphic in D^* with no zeros in D^*. Any rational function with these properties is outer on D^*. To see this, it suffices by multiplicativity of outer functions to show that functions of the form $(1/z - 1/z_0)$ with $z_0 \in D \cup \mathbf{T}$ are outer, which is easy to verify.

This proves that there exists an a with $(a, b) \in \mathbf{L}$. Uniqueness follows very generally from the fact that the normalized outer function a is determined by $|a|$ almost everywhere on $\mathbf{T}$, and the latter is determined by b. This proves the first statement of the lemma.

Clearly holomorphicity of b in D is necessary for $(a, b) \in \mathbf{H}(D)$. However, if b is holomorphic in D, which is the same as saying b has no poles in D, then b/a is a rational function holomorphic in D and bounded by 1 almost everywhere on $\mathbf{T}$, and thus holomoprhic in a neighborhood of $D \cup \mathbf{T}$ and thus in $H^2(D)$. This proves $(a, b) \in \mathbf{H}(D)$. The statement about $\mathbf{H}_0(D)$ is clear. This together with the symmetric statement for D^* proves the remaining statements of the lemma. □

The next lemma states that solving the problem of Riemann-Hilbert factorization does not leave the class of rational functions.

LEMMA 1.32. *Assume $(a, b) \in \mathbf{L}$ is rational. Given any factorization*

$$(a_-, b_-)(a_+ b_+) = (a, b)$$

with $(a_-, b_-) \in \mathbf{H}^2(D^*)$ and $(a, b) \in \mathbf{H}(D)$, then (a_-, b_-) and (a_+, b_+) are also rational.

PROOF. Recall from Lemma 1.26 that

$$\frac{a_-}{a}, \frac{b_-}{a}, \frac{a_+}{a}, \frac{b_+^*}{a} \in H^2(D^*)$$

In particular

$$\int_{\mathbf{T}} \left| \frac{a_+}{a} \right|^2 (r\cdot) \leq C$$

for $r \geq 1$. Now $aa^* = 1 + bb^*$ implies that the poles of b on $\mathbf{T}$ have the same order as the poles of a on $\mathbf{T}$. Thus b/a, which is also a rational function, is actually holomorphic on a neighborhood of $\mathbf{T}$.

Using

$$a_-^* = a_+ a^* - b^* b_+$$

we obtain

$$\frac{a_+}{a^*} = \frac{1}{a^*} \frac{a_-^*}{a^*} + \frac{b}{a^*} \frac{b_+}{a^*}$$

On the right-hand side, the functions a_-^*/a^* and b_+/a^* are in the Hardy space $H^2(D)$, while the rational functions $1/a^*$ and b^*/a^* are holomorphic in a neighborhood of $\mathbf{T}$.

Therefore, a_+ has a meromorphic extension to D which is holomorphic in an annulus $1 - \epsilon < |z| < 1$ for some small ϵ and satisfies

$$\int_{\mathbf{T}} \left| \frac{a_+}{a^*} \right|^2 (r\cdot) \leq C$$

for $1 - \epsilon < r \leq 1$

Observe that a and a^* have comparable moduli in a small neighborhood of $\mathbf{T}$ since the quotients a/a^* and a^*/a are holomorphic near $\mathbf{T}$.

Thus

$$\int_{\mathbf{T}} \left| \frac{a_+}{a} \right|^2 (r\cdot) \leq C$$

for $1 - \epsilon < r < 1$ for some small ϵ, and the same estimate has been observed previously for $r \geq 1$.

We claim that holomorphicity of a_+/a in a neighborhood of $\mathbf{T}$ with possible exception on $\mathbf{T}$ together with the above estimates implies that a_+/a is indeed holomorphic across $\mathbf{T}$.

In the current situation that a_+/a is in addition meromorphic in D and D^* with finitely many poles we can argue as follows.

We may remove the poles of a_+/a in D by the following recursive procedure. If a_+/a has a pole at $z_\infty \in D$, then we subtract a constant from a_+/a so that the new function has a zero at a distinct point $z_0 \in D$, and then we multiply the function by $(z - z_\infty)/(z - z_0)$. This reduces the order of the pole at z_∞ and leaves the order of all other poles unchanged. Iterating this procedure we obtain a function g which is holomorphic in D and D^*. The above L^2 estimates prevail throughout this iteration, possibly with different constants C, so g is in $H^2(D) \cap H^2(D^*)$. Therefore g is constant, and we conclude that a_+/a is rational. The estimates near $\mathbf{T}$ then imply that it has no poles on $\mathbf{T}$.

More generally, the claim can be proved using the theorem of Morera: a function is holomorphic in a disc if the Cauchy integral over each triangle vanishes. For triangles which avoid $\mathbf{T}$ this is obvious for a_+/a, and for triangles which intersect $\mathbf{T}$ one obtains vanishing of the Cauchy integral by approximating the triangle by shapes avoiding $\mathbf{T}$ and then using maximal function estimates to pass to the limit.

This proves that a_+ is rational, and one can argue similarly that b_+, a_-, b_- are rational.

This proves Lemma 1.32. $\square$

The lemma just proved reduces the Riemann-Hilbert problem for rational (a, b) to a purely algebraic problem in the class of rational functions. Even better, the following lemma states that the solution functions a_-, b_-, a_+, b_+ are in a sense subordinate to a, b. This reduces the Riemann-Hilbert problem to a finite dimensional problem.

If f is meromorphic near $z \in \mathbf{C}$, denote by $\mathrm{ord(f,z)}$ the order of the pole of f at z. Thus

$$f(\zeta)(z-\zeta)^{\mathrm{ord(f,z)}}$$

is holomorphic at z and does not vanish at z. For rational functions we define the order at ∞ in the usual manner using a change of coordinates on the Riemann sphere. Observe that the order is a negative number if f vanishes at z.

Call a rational function g subordinate to another rational function f on a certain domain if for all points z in the domain such that $\mathrm{ord(g,z)} > 0$ we have $\mathrm{ord(f,z)} \geq \mathrm{ord(g,z)}$. We call g subordinate to f if g is subordinate to f on the whole Riemann sphere. Clearly, if we fix f, the set of rational functions g subordinate to f is a finite dimensional vector space.

LEMMA 1.33. *Let $(a, b) \in \mathbf{L}$ be rational. Then a is subordinate to bb^*.*
If $(a_-, b_-) \in \mathbf{H}_0$ and $(a_+, b_+) \in \mathbf{H}$ such that

$$(a_-, b_-)(a_+, b_+) = (a, b)$$

then the rational functions b_- and b_+ are subordinate to b.

PROOF. Since $(a, b) \in \mathbf{L}$, we have $aa^* = 1 + bb^*$. Since a^* does not vanish in $D \cup \mathbf{T}$, we see that a is subordinate to bb^* on a neighborhood of $D \cup \mathbf{T}$. Since a

has no poles on D^*, it is subordinate to bb^* on D^* and thus on the whole Riemann sphere.

Now let (a_-, b_-) and (a_+, b_+) be a Riemann-Hilbert factorization as in the lemma. Then

$$b_- = -ab_+ + ba_+$$

$$b_+ = -\frac{b_-}{a} + b\frac{a_+}{a}$$

On $D^* \cup \mathbf{T}$, the functions b_-/a and a_+/a have no poles, since they are in $H^2(D^*)$. The last display then implies that b_+ is subordinate to b on a neighborhood of $D^* \cup \mathbf{T}$. Since b_+ has no poles on D, it is subordinate to b on the whole Riemann sphere.

Similarly one proves that b_- is subordinate to b. □

LEMMA 1.34. *Let $(a, b) \in \mathbf{L}$ be rational. Then there exists a unique Riemann-Hilbert factorization*

$$(a, b) = (a_-, b_-)(a_+, b_+) \tag{1.40}$$

such that b_+ does not have any poles on $\mathbf{T}$. The factor (a_+, b_+) coincides with the factor (a_{++}, b_{++}) in the triple factorization of Theorem 1.28. Similarly, there exists a unique Riemann-Hilbert factorization (1.40) such that b_- does not have any poles on $\mathbf{T}$. For this factorization, the factor (a_-, b_-) coincides with the factor (a_{--}, b_{--}) in the triple factorization of Theorem 1.28.

PROOF. If there exists a Riemann-Hilbert factorization (1.40) such that b_+ has no pole on $\mathbf{T}$, then the factor (a_+, b_+) has to coincide with (a_{++}, b_{++}). Namely, it is clear that (a_+, b_+) has no further nontrivial Riemann-Hilbert factorization by Lemma 1.25. This implies that (a_+, b_+) is equal to (a_{++}, b_{++}).

In particular we have proved that the requirement that b_+ has no poles on $\mathbf{T}$ makes the Riemann-Hilbert factorization unique.

It remains to show that such a Riemann-Hilbert factorization exists.

We set up a Banach fixed point argument as in the proof of Lemma 1.25. Recall from the proof of that lemma that, for every constant c, the affine linear mapping

$$T : (A, B) \to (c + P_{D^*}(\frac{b^*}{a^*}B), P_D(\frac{b}{a}A))$$

is a weak contraction mapping in the sense that

$$\|T(A, B) - T(A', B')\| \leq \|(A, B) - (A', B')\|$$

where the norms are with respect to $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$. Indeed, unless $(A', B') = (A, B)$, the inequality is strict since multiplication by b/a strictly lowers the L^2 norm of any nonzero element. This implies that on any invariant finite dimensional subspace of $L^2(\mathbf{T})$, the mapping is a strict contraction in the sense

$$\|T(A, B) - T(A', B')\| \leq (1 - \epsilon)\|(A, B) - (A', B')\|$$

for some ϵ depending on the subspace. This can be seen by a compactness argument.

Consider the finite dimensional space V of all rational (A, B) such that B is subordinate to b, A is subordinate to b^*, and A and B have no poles on $\mathbf{T}$. This is clearly a subspace of $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$. For any rational function f without poles on $\mathbf{T}$ the projection $P_D f$ is up to an additive constant the sum of the principal parts of the poles of f in D^*, while $P_{D^*} f$ is up to an additive constant the sum of the principal parts of the poles of f in D. Therefore, $P_D f$ is a rational function subordinate to f with no poles in D and $P_{D^*}f$ is a rational function subordinate to f with no poles in D^*.

We observe that for $(a, b) \in \mathbf{H}$ the quotient b/a has no poles on $\mathbf{T}$ and is subordinate to b on D^*. Thus for any $(A, B) \in V$ we have that $P_{D^*}(b^*B/a^*)$ is subordinate to b^* and $P_D(bA/a)$ is subordinate to b. Thus V is invariant under the mapping T for any c. Since T is a strict contraction mapping on V, there exists a fixed point in V under this mapping. Using this fixed point (A, B) for $c = 1$, we can as in the proof of Lemma 1.25 produce a right factor

$$(a_+, b_+) = A(\infty)^{-1}(A, B)$$

to the Riemann-Hilbert factorization problem for (a, b). Clearly b_+ has no poles on $\mathbf{T}$, so this is the desired right factor.

The symmetric statement concerning left factors is proved similarly. This completes the proof of Lemma 1.34. □

The above shows that there is a very satisfactory description of the set of rational elements in $\mathbf{L}$ which qualify to be left, middle, or right factors in a triple factorization as in Theorem 1.28. Namely, possible left (middle, right) factors are exactly those rational $(a, b) \in \mathbf{L}$ for which b has only poles in D, ($\mathbf{T}$, D^*).

It remains to study the possible factorizations of a rational middle factor in the triple factorization. Thus we are reduced to study the Riemann-Hilbert problem for rational $(a, b) \in \mathbf{H}_0^* \cap \mathbf{H}$. Any factorization consists again of rational factors in $\mathbf{H}_0^* \cap \mathbf{H}$.

This problem too has a very satisfactory answer, though the formulation of the answer is a little more involved.

Before we proceed further, we shall briefly digress on the maximum principle. The maximum principle says that any nonconstant holomorphic function on D which is continuous on $D \cup \mathbf{T}$ attains its maximum only on $\mathbf{T}$. If the function is actually differentiable on $D \cup \mathbf{T}$, then the following lemma gives more precise information. It is a version of the maximum principle which may be less well known.

LEMMA 1.35. *Let f be a nonconstant holomorphic map from D to itself and assume that f and f' have continuous extensions to the boundary $\mathbf{T}$. Thus f and f' map $D \cup \mathbf{T}$ to $D \cup \mathbf{T}$.*

If f attains its maximum at $z \in \mathbf{T}$, then $f'(z) = z^* \omega f(z)$ for some strictly positive ω.

PROOF. Multiplying f by a constant phase factor, if necessary, we may assume $f(z) = 1$. Consider the real part u of f. It has a maximum at z. In particular, u has zero derivative in the direction tangential to $\mathbf{T}$. Therefore, the gradient of u has to be radial and is either 0 or outward pointing. This proves

$$\frac{\partial u}{\partial x} + i \frac{\partial u}{\partial y} = \omega z$$

for some $\omega \geq 0$ Thus, by the Cauchy Riemann equations,

$$f'(z) = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} = \frac{\partial u}{\partial x} - i \frac{\partial u}{\partial y} = \omega z^* = \omega z^* f(z)$$

It remains to show that ω is not zero, i.e., that the harmonic function u does not have vanishing derivative at z. Assume by a rotation that $z = 1$. It will suffice to find some function $\tilde{u}$ which dominates u in the intersection of D with a neighborhood of 1, such that $\tilde{u}$ is differentiable at 1 with nonvanishing derivative.

Since u is not constant, we find two points on $\mathbf{T}$ where u is strictly less than 1. The two points divide the circle into wo arcs C_1 and C_2. Let L be the line connecting the two points. Assume w.l.o.g. that C_1 contains 1 and let z_0 be a point of C_2. Define

$$\tilde{u}(\zeta) = 1 + \epsilon \mathrm{Re} \frac{\zeta + z_0}{\zeta - z_0}$$

for some small $\epsilon > 0$. Since $\tilde{u}$ is 1 on the arc C_1, it dominates u there. Since u is strictly less than 1 in the (compact) line L, we can choose ϵ small enough so that $\tilde{u}$ dominates u on the line. By the (easy) maximum principle, $\tilde{u}$ dominates u inside the enterior of $C_2 \cup L$. It remains to prove that $\tilde{u}$ has nonvanishing derivative at 1. This however can be done easily by direct inspection. $\square$

We continue to study Riemann-Hilbert factorizations

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

for rational $(a, b) \in \mathbf{H} \cap \mathbf{H}_0^*$. Thus b and a have only poles on $\mathbf{T}$. Indeed, they have the same poles as $1 + bb^* = aa^*$ shows.

By Lemma 1.33, the functions a_+ and a_- can only have poles where a has poles. Consider the identity

$$a = a_- a_+ \left(1 + \frac{b_-}{a_-} \frac{\overline{b_+}}{a_+}\right) \tag{1.41}$$

The function

$$\frac{b_-}{a_-} \frac{\overline{b_+}}{a_+} \tag{1.42}$$

maps $D \cup \mathbf{T}$ to itself. Therefore, the last factor on the right-hand side of (1.41) can only vanish at point $z \in \mathbf{T}$ when z is a maximum of the function (1.42) on $D \cup \mathbf{T}$. By Lemma 1.35, the last factor in (1.41) can only have a simple zero at z.

Therefore, for every pole z of a, we have either

$$ord(a, z) = ord(a_-, z) + ord(a_+, z) \tag{1.43}$$

or

$$ord(a, z) = ord(a_-, z) + ord(a_+, z) - 1 \tag{1.44}$$

We say that the pole z is *split* if (1.43) holds, and we say that it is shared if (1.44) holds. If z is a shared pole, then both functions b_-/a_- and b_+^*/a_+ have modulus one at z. Therefore, both functions a_+ and a_- have a pole at z and by (1.44) both poles have order at most $\mathrm{ord(a, z)}$.

For each pole z of a we define

$$n := \mathrm{ord(a, z)}, \quad \mathrm{n}^- := \mathrm{ord(a_-, z)}, \quad \mathrm{n_j^+} := \mathrm{ord(a_+, z)}$$

Define the functions

$$A_+ := 1 - \frac{b_+ b^*}{a_+ a^*} = \frac{a_-^*}{a_+ a^*} = \frac{1}{a_+ a_+^*} \frac{1}{1 + \frac{b_-^*}{a_-^*} \frac{b_+}{a_+^*}}$$

$$A_- := 1 - \frac{b_- b^*}{a_-^* a} = \frac{a_+}{a_-^* a} = \frac{1}{a_- a_-^*} \frac{1}{1 + \frac{b_-}{a_-}\frac{b_+^*}{a_+}}$$

On $\mathbf{T}$, the functions A_+ and A_- have positive real part except possibly where a has a pole (use the first representation for A_+, A_-). There, A_+ vanishes of order $n + n_- - n_+$ and A_- vanishes of order $n + n_+ - n_-$ (use the second representation for A_+, A_-). In particular, A_+ vanishes of order $2n^+$ if the pole is split or $2n^+ - 1$ if the pole is shared.

For each shared pole z, we define μ^+ and μ^- by the asymptotic expansions

$$A_+(\zeta) = -\mu^+ z(\zeta - z)^{n^+ - 1}(\frac{1}{\zeta} - \frac{1}{z})^{n^+} + O(\zeta - z)^{2n^+}$$

$$A_-(\zeta) = -\mu^- z^*(\zeta - z)^{n^-}(\frac{1}{\zeta} - \frac{1}{z})^{n^- - 1} + O(\zeta - z)^{2n^-})$$

We claim that μ^+ and μ^- are positive.

To see this for A_+, we set

$$a_+(\zeta) = \gamma(\zeta - z)^{-n^+} + O(\zeta - z)^{-n^+ + 1}$$

and, by Lemma 1.35,

$$(1 + \frac{b_-^* b_+}{a_-^* a_+^*})'(z) = (\frac{b_-^* b_+}{a_-^* a_+^*})'(z) = (\frac{b_-^* b_+}{a_-^* a_+^*})(z) z^* \mu = -z^* \mu$$

for some positive μ. Using the third representation of A_+ we obtain

$$\mu^+ = 1/(\mu|\gamma|^2)$$

awhich shows that μ^+ is positive. The proof that μ^- is positive is similar.

Write

$$\frac{1}{aa^*}(\zeta) = \mu(\zeta - z)^n(\frac{1}{\zeta} - \frac{1}{z})^n + O(\zeta - z)^{2n+1}$$

Then the identity

$$A_+ A_- = \frac{1}{aa^*}$$

shows that

$$\mu^+ \mu^- = \mu$$

Our goal is to see that the parameters n^+ and n^- for all poles together with the parameters μ^+ and μ^- for all shared poles parameterize the Riemann-Hilbert factorizations of (a, b). We shall first adress the easier statement that all Riemann-Hilbert factorizations are uniquely determined by these parameters.

LEMMA 1.36. *Let (a, b) be rational in $\mathbf{H} \cap \mathbf{H}_0^*$. Then any Riemann-Hilbert factorization*

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

is uniquely determined if the parameters n^+ and n^- for all poles and the parameters μ^+, and μ^- for all shared poles are specified.

PROOF. We assume to get a contradiction that there are two Riemann Hilbert factorizations with right factors (a_+, b_+) and $(\tilde{a}_+, \tilde{b}_+)$ respectively, which have the same parameters listed in the lemma.

Define (c, d) by

$$(\tilde{a}_+, \tilde{b}_+) = (c, d)(a_+, b_+)$$

Our task is to show that $(c, d) = (1, 0)$. It suffices to show that d is constant. Then d has to be constant 0 as one can see from evaluating the defining equation for (c, d) at 0 and using that $b_+, \tilde{b}_+$ vanish at 0 while a_+^* does not. Then $c = \tilde{a}_+/a_+$ is an outer function on D^* and of constant modulus 1 on $\mathbf{T}$, and thus it is constant. This constant is equal to 1 as one can see from evaluating c at ∞.

Observe that d is a rational function and can only have poles where a has poles. Therefore, it suffices to show that d is holomorphic at all poles of a.

Fix a pole z. Define $r := b/a$ and similarly $r_+, r_-, \tilde{r_+}$. Then we have

$$r - r_+ = \frac{1}{r^*}(1 - r^* r_+ - (1 - r^* r)) = \frac{1}{r^*}(A_+ - (1 - r^* r))$$

Observe that r^* has modulus one at z and $1 - rr^* = (aa^*)^{-1}$ vanishes of order $2n$ at z. Moreover, A_+ vanishes at least of order $2n^+ - 1$ at z and its Taylor coefficient of order $2n^+$ at z is determined by μ^+.

Therefore, $r - r_+$ vanishes at least of order $2n^+ - 1$ at z and its Taylor coefficient of order $2n$ is determined by r and μ^+. The same holds for $r - \tilde{r}_+$, and by taking differences we see that $r_+ - \tilde{r}_+$ vanishes of order $2n^+$ at z. Since a_+ has a pole of order n^+ at z, we see that

$$(r_+ - \tilde{r}_+)a_+\tilde{a}_+ = b_+\tilde{a}_+ - \tilde{b}_+ a_+ = d$$

has no pole at z. □

THEOREM 1.37. *Assume $(a, b) \in \mathbf{H} \cap \mathbf{H}_0^*$ is rational. Let $z_j \in \mathbf{T}$, $j = 1, \ldots, N$ be the distinct poles of a and denote the order of the pole z_j by n_j.*

Assume we are given numbers $0 \le n_j^+, n_j^- \le n_j$ for $j = 1, \ldots, N$ such that for each j either

$$n_j^+ + n_j^- = n_j$$

(split case) or

$$n_j^+ + n_j^- - 1 = n_j$$

(shared case). Assume further that for each j in the shared case we are given positive real numbers μ_j^+, μ_j^- with

$$\mu_j^+ \mu_j^- = \mu_j$$

where μ_j is defined by

$$\frac{1}{aa^*}(\zeta) = \mu_j(\zeta - z_j)^{n_j}(\frac{1}{\zeta} - \frac{1}{z_j})^{n_j} + O(\zeta - z_j)^{2n_j+1}$$

Then there exists a unique Riemann-Hilbert factorization

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

such that

$$\mathrm{ord}(a_+, z_j) = n_j^+$$
$$\mathrm{ord}(a_-, z_j) = n_j^-$$

and, if j is in the shared case,

$$A_+(\zeta) = -\mu_j^+ z_j(\zeta - z_j)^{n_j^+ - 1}(\frac{1}{\zeta} - \frac{1}{z_j})^{n_j^+} + O((\zeta - z_j)^{2n_j^+})$$

$$A_-(\zeta) = -\mu_j^- z_j^*(\zeta - z_j)^{n_j^-}(\frac{1}{\zeta} - \frac{1}{z_j})^{n_j^- - 1} + O((\zeta - z_j)^{2n_j^-})$$

All Riemann-Hilbert factorizations of (a, b) are obtained in this way.

Proof. Our previous discussion of the parameters n_j^+,n_j^-,μ_j^+ and μ_j^- already implies that every Riemann-Hilbert factorization of (a, b) comes with parameters as described in the theorem and the parameters determine the factorization uniquely.

It thus remains to show that for a given set of parameters such a Riemann-Hilbert factorization exists.

It is enough to consider the case when b is nonzero at ∞, because

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

is equivalent to

$$(a, bz^n) = (a_-, b_-z^n)(a_+, b_+z^n)$$

and thus one can reduce the case of b vanishing at ∞ to the case of b not vanishing at ∞.

We first prove existence of a Riemann-Hilbert factorization in the easier case when all poles are split. We write

$$b(z) = b(\infty)\left[\prod(z - y_k)\right]\left[\prod(z - z_j)^{-n_j}\right]$$

where y_k are the zeros of b counted with multiplicities.

For each j, consider points $z_j^+ \in D^*$ and $z_j^- \in D$ close to z_j. It shall be enough to consider $z_j^\pm$ such that they avoid the zeros of b and are all pairwise distinct as well as distinct from all $(z_j^\pm)^*$. Consider the perturbation $\tilde{b}$ of b defined by

$$\tilde{b}(z) = b(\infty)\left[\prod(z - y_k)\right]\left[\prod(z - z_j^+)^{-n_j^+}\right]\left[\prod(z - z_j^-)^{-n_j^-}\right]$$

This function has the same zeros with multiplicities as b, but the poles are at perturbed locations.

Since the zeros y_k are fixed, we have an upper bound on

$$\int_{\mathbf{T}} \log_+ |\tilde{b}|$$

uniformly in the choice of the points $z_j^\pm$.

By Lemma 1.31, there is a unique rational $\tilde{a}$ such that $(\tilde{a}, \tilde{b}) \in \mathbf{L}$. The equation $1 + \tilde{b}\tilde{b}^* = \tilde{a}\tilde{a}^*$ implies that there is a uniform upper bound on

$$\int_{\mathbf{T}} \log_+ |\tilde{a}|$$

and since $\tilde{a}$ is outer on D^* we have a uniform upper bound on $\tilde{a}(\infty)$. Trivially, we also have the lower bound $1 \le |\tilde{a}(\infty)|$.

Since $\tilde{a}$ is bounded on $\mathbf{T}$, Lemma 1.25 gives a unique Riemann-Hilbert factorization

$$(\tilde{a}, \tilde{b}) = (\tilde{a}_-, \tilde{b}_-)(\tilde{a}_+, \tilde{b}_+)$$

Applying Lemma 1.33 thoroughly to the situation at hand, we conclude that $\tilde{a}_+$ has poles only at the points $(z_j^+)^*$ with order at most n_j^+, while $\tilde{a}_-$ has poles only at the points z_j^- with order at most n_j^-. Moreover, since $|\tilde{a}_+(\infty)|$ and $|\tilde{a}_-(\infty)|$ are bounded below by one and $\tilde{a}_+(\infty)\tilde{a}_-(\infty) = \tilde{a}(\infty)$, we obtain that $\tilde{a}_+(\infty)$ and $\tilde{a}_-(\infty)$ are in a fixed compact set avoiding 0.

We can write

$$\tilde{a}_+(z) = \tilde{a}_+(\infty)\left[\prod(z - x_k^+)\right]\left[\prod(z - (z_j^+)^*)^{-\tilde{n}_j^+}\right]$$

where x_k^+ are the zeros of $\tilde{a}_+$ with multiplicities and $\tilde{n}_j^+ \le n_j^+$.

Now we consider a sequence of choices of $z_j^{\pm}$ such that for each j both points z_j^+ and z_j^- converge to z_j. For ea Since the zeros x_k^+ remain in the compact set $D \cup T$ and the value $\tilde{a}_+(\infty)$ remains in a compact set away from 0, there is a subsequence for which the $\tilde{n}_j^+$ are constant, each zero x_k^+ converges (assuming the zeros are appropriatelyenumerated), and the value $\tilde{a}_+(\infty)$ converges. For this subsequence, $\tilde{a}_+$ converges uniformly on compact sets away from the poles of a to a limit a_+. The zeros of a_+ are still in $D \cup \mathbf{T}$ an the poles are on $\mathbf{T}$. Thus a_+ is still outer. Clearly also $a_+(\infty)$ is positive as a limit of positive numbers.

Similarly, one can choose a further subsequence so that all other terms in the identity

$$(\tilde{a}, \tilde{b}) = (\tilde{a}_-, \tilde{b}_-)(\tilde{a}_+, \tilde{b}_+)$$

converge uniformly on compact sets away from the poles of a. In the limit, we obtain a Riemann-Hilbert factorization

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

By construction, the poles of a_+ and of a_- at z_j are at most of order n_j^+ and n_j^- respectively. Since the sum of these orders for any Riemann-Hilbert factorization has to be at least n_j, the orders of a_+ and a_- at z_j are exactly n_j^+ and n_j^-. Thus we have proved existence of a Riemann-Hilbert factorization for the given parameters in the completely split case.

Now we modify the above argument so that it works in the case when there are shared poles.

For each split pole z_j, we choose again z_j^+ and z_j^- exactly as before. For every shared pole, we choose z_j^+ and z_j^- as before but with the additional constraint that

$$\mu_j^+ |(z_j^+)^* - z_j|^{2n_j^+} = \mu_j^- |z_j^- - z_j|^{2n_j^-} \tag{1.45}$$

We define

$$\tilde{b}(z) = b(\infty) \left[\prod (z - y_k)\right] \left[\prod_{z_j \text{ shared}} (z - z_j)\right] \left[\prod (z - z_j^+)^{-n_j^+}\right] \left[\prod (z - z_j^-)^{-n_j^-}\right]$$

Compared to the completely split case, we have defined $\tilde{b}$ to have an additional zero at each shared pole z_j. Since for each shared pole we have $n_j^+ + n_j^- = n_j + 1$, the numerator and denominator of the rational function defining $\tilde{b}$ have the same degree and $\tilde{b}(\infty)$ is again finite.

As before, we obtain a unique Riemann-Hilbert factorization

$$(\tilde{a}, \tilde{b}) = (\tilde{a}_-, \tilde{b}_-)(\tilde{a}_+, \tilde{b}_+)$$

Then we let $z_j^{\pm}$ tend to z_j respecting the additional constraint (1.45) for each shared pole. As before, we can choose a subsequence so that all quantities in the Riemann-Hilbert factorization converge uniformly on compact sets away from the poles z_j. In the limit, we obtain a Riemann-Hilbert factorization

$$(a, b) = (a_-, b_-)(a_+, b_+)$$

We need to show that this factorization has the given parameters $n_j^{\pm}$ and $\mu_j^{\pm}$. As before, for each split pole z_j the order of poles of the limits a_+ and a_- are at most n_j^+ and n_j^- and thus have to be exactly n_j^+ and n_j^-.

We consider a shared pole z_j. We calculate

$$\tilde{b} = \tilde{a}_- \tilde{b}_+ + \tilde{b}_- \tilde{a}_+^*$$

$$\frac{\tilde{b}}{\tilde{a_-} \tilde{a}_+^*} = \frac{\tilde{b}_+}{\tilde{a}_+^*} + \frac{\tilde{b}_-}{\tilde{a}_-}$$

At the shared pole z_j, the left-hand side vanishes:

$$0 = \frac{\tilde{b}_+(z_j)}{\tilde{a}_+^*(z_j)} + \frac{\tilde{b}_-(z_j)}{\tilde{a}_-(z_j)} \tag{1.46}$$

Since for every element $(a', b') \in \mathbf{L}$, the modulus of b'/a' on $\mathbf{T}$ can be expressed in terms of the modulus of a, we conclude from (1.46) that

$$|\tilde{a}_+(z_j)| = |\tilde{a}_-(z_j)| \tag{1.47}$$

In a compact neighborhood of z_j avoiding the poles z_k with $k \neq j$, we can write

$$\tilde{a}_+(z) = (z - (z_j^+)^*)^{-n_j^+} \tilde{h}^+(z)$$

$$\tilde{a}_-(z) = (z - z_j^-)^{-n_j^-} \tilde{h}^-(z)$$

where $\tilde{h}^+$ and $\tilde{h}^-$ converge uniformly on the neighborhood to functions h^+ and h^-. The equation (1.47) then becomes

$$|z_j - (z_j^+)^*|^{-n_j^+} |\tilde{h}^+(z_j)| = |z_j - z_j^-|^{-n_j^-} |\tilde{h}^+(z_j)|$$

By choice of the $z_j^\pm$ we thus have for some constant $\tilde{c}$:

$$|\tilde{h}^+(z_j)| = \tilde{c}(\mu_j^+)^{-1/2}$$

$$|\tilde{h}^-(z_j)| = \tilde{c}(\mu_j^-)^{-1/2}$$

Taking the limit, we obtain

$$|h^+(z_j)| = c(\mu_j^+)^{-1/2}$$

$$|h^-(z_j)| = c(\mu_j^-)^{-1/2}$$

for some constant c. We claim that c is not zero. Assume to get a contradiction that it is zero, then h^+ and h^- vanish at z_j. From the equations

$$a_+(z) = (z - z_j)^{-n_j^+} h^+(z)$$

$$a_-(z) = (z - z_j)^{-n_j^-} h^-(z)$$

we see that a_+ and a_- have order of pole at most $n_j^+ - 1$ and $n_j^- - 1$ at z_j. The sum of these orders is less than n_j, a contradiction. Therefore c is not zero.

Then h^+ and h^- do not vanish at z_j and a_+ and a_- have poles of exact order n_j^+ and n_j^- at z_j. Thus we are indeed in the case of a shared pole.

Moreover, we calculate in the limit

$$\frac{|a_+|^2}{|a_-|^2} = \frac{\mu_j^-}{\mu_j^+} |z - z_j|^{-2n_j^+ + 2n_j^-} + O(|z - z_j|^{-2n_j^+ + 2n_j^- + 1})$$

Now we use $\mu_j^+ \mu_j^- = \mu_j$ and

$$\frac{1}{|a|^2} = \mu_j^2 |z - z_j|^{2n_j} + O(|z - z_j|^{2n_j + 1})$$

to obtain

$$\frac{|a_+|^2}{|a_-|^2|a|^2} = |\mu_j^-|^2|z - z_j|^{-2n_j^+ + 2n_j^- + 2n_j} + O(|z - z_j|^{-2n_j^+ + 2n_j^- + 2n_j + 1}$$

Comparing with the asymptotics for A_- and doing the analogue calculation for A_+ we conclude that the shared pole z_j has indeed the parameters μ_j^+ and μ_j^-. □

1.16. Orthogonal polynomials

1.17. Orthogonal polynomials

In this lecture, we describe how the nonlinear Fourier transform on the half-line relates to orthogonal polynomials. The material on orthogonal polynomials is folklore, a standard reference is [**28**] and a more recent introduction with interesting applications is [**8**].

Let μ be any compactly supported positive measure on the plane $\mathbf{C}$ with the normalization $\|\mu\| = 1$.

Let H be the Hilbert space completion of the linear span of the set of functions $z^0 = 1, z^1, z^2, \ldots$ (monomials) under the inner product

$$\langle f, g\rangle = \int f\overline{g}\, d\mu$$

If a finite set $z^0, \ldots, z^n$ of monomials is linearly dependent in this Hilbert space, then necessarily μ has finite support. Namely, linear dependence means that some linear combination of these monomials, which is nothing but a polynomial $P(z)$, is equivalent to 0 in the Hilbert space. This means

$$\|P\| = \int |P(z)|^2\, d\mu = 0$$

As $|P|^2$ and $d\mu$ are positive, this can only happen if the support of μ is contained in the null set of P, which is finite. Conversely, if μ has finite support, then it is easy to find a polynomial which vanishes on the support and thus is equivalent to 0 in the Hilbert space.

From now on the standing assumption is that μ has infinite support.

We can apply the Gram-Schmidt orthogonalization process to the sequence of vectors z^n. Let

$$\phi_0 = z^0 = 1$$

and let ϕ_n be the unique polynomial of degree n which has unit length in the Hilbert space H, has positive highest coefficient, and is orthogonal to all polynomials of degree less than n. Since the space of polynomials of degree less than n has exact codimension 1 in the space of polynomials of degree less than or equal to n, such a polynomial ϕ_n exists and clearly has exact degree n. The orthogonality condition determines ϕ_n up to a scalar factor. The modulus of this scalar factor is determined by the requirement that ϕ_n has unit length, and the phase is determined by the requirement that the highest coefficient of ϕ_n, which is the coefficient in front of z^n, is positive.

The polynomials ϕ_n form an orthonormal set. Indeed, they form an orthonormal basis of H since they span the same subspace as the monomials z^n, which by definition span the full space H.

The linear operator

$$T : f \to zf$$

originally defined on all polynomials f, is bounded with respect to the norm on H because the function z is bounded on the support of μ. Therefore, the operator T extends to a unique bounded operator on H.

We can express the map T in the basis ϕ_n. This means we represent elements in H as infinite linear combinations $\sum a_j\phi_j$ and let T act on the column vector $(a_j)_{j\geq 0}$ by matrix multiplication from the left by a matrix $(J_{ij})_{i,j\geq 0}$:

$$(a_j) \to (\sum_j J_{ij}a_j)$$

As $z\phi_j$ is a polynomial of degree $j+1$, we have $J_{ij} = 0$ if $i > j+1$. Moreover, $z\phi_j$ has positive highest coefficient, and thus $J_{j+1,j} > 0$. We call a matrix a Hessenberg matrix if it satisfies these two constraints, namely positivity on the subdiagonal and vanishing below the subdiagonal.

The requirement that Hessenberg matrices be positive on the subdiagonal is a matter of convenient normalization to produce uniqueness results. One can conjugate a Hessenberg matrix by a diagonal unitary matrix and thus obtain an equivalent matrix which vanishes below the subdiagonal but is merely nonzero on the subdiagonal. In the literature, one often calls these more general matrices Hessenberg matrices.

We will specialize to two cases.

Case 1: The measure μ is supported on the real line $\mathbf{R}$. In this case $T : f \to zf$ is selfadjoint since

$$\langle Tf, g\rangle = \int zf\overline{g}\,d\mu = \int f\overline{zg}\,d\mu = \langle f, Tg\rangle$$

where we have used that z is real on the support of μ. Thus J is a selfadjoint matrix, and since it is Hessenberg, it is tridiagonal ($J_{ij} = 0$ if $|i-j| > 1$). A Hessenberg matrix which is self adjoint is called a Jacobi matrix.

Observe that the j-th column vector of a Jacobi matrix has two real parameters not determined by the previous columns: the subdiagonal is a positive number by definition of Hessenberg matrices and independent of the previous columns, the diagonal entry has to be a real number by selfadjointness and is independent of the previous columns, while the superdiagonal entry is determined by the previous column.

We know that T and thus J have to be bounded operators. For tridiagonal matrices J_{ij}, boundedness is equivalent to $\sup_{ij}|J_{ij}| < \infty$.

Case 2: The measure is supported on the circle $\mathbf{T}$. In this case we have $T^*T = \mathrm{id}$ since

$$\langle T^*Tf, g\rangle = \langle Tf, Tg\rangle = \int zf\overline{zg}\,d\mu = \int f\overline{g}\,d\mu = \langle f, g\rangle$$

Observe that $T^*T = \mathrm{id}$ does not imply that T is unitary, since $TT^* = \mathrm{id}$ may fail. For example, if the measure μ is a normalized Lebesgue measure on $\mathbf{T}$, then the monomials z^n form an orthonormal set. The operator T in this case is the standard shift operator and is obviously not surjective since the image of T contains only functions with zero constant term.

Unlike in the selfadjoint case, the matrix J is not sparse above the diagonal. However, we still have two new real parameters per column. The condition $J^*J = \mathrm{id}$ implies that the column vectors are orthonormal. Thus the first n entries of the n-th column have to be orthogonal to the previous columns, and thus there is only a complex parameter in D (the vector of the first n entries must have norm less than 1) as degree of freedom, and then the $(n+1)$-st entry is determined since it is positive and makes the column have unit length.

Observe that boundedness of a matrix satisfying $J^*J = \mathrm{id}$ is automatic.

THEOREM 1.38. *The above construction of a Jacobi matrix provides a bijective correspondence between*

(1) *A compactly supported positive measure μ supported on the real line with $\|\mu\| = 1$ and infinite support.*
(2) *A tridiagonal selfadjoint matrix $J = (J_{ij})_{i,j\geq 0}$ that has strictly positive elements on the subdiagonal and has finite operator norm.*

PROOF. We need to show that the map from measures to Jacobi matrices described above is injective and surjective.

The matrix J determines the moments of $d\mu$ because

$$\int z^n d\mu = \langle T^n\phi_0, \phi_0\rangle = \langle J^n e_0, e_0\rangle$$

where e_n denotes the n-th standard unit vector whose n-th entry is 1 and all other entries are 0. But then by the Stone-Weierstrass theorem, this determines $d\mu$. Hence the map $\mu \mapsto J$ is injective.

To show surjectivity, let J be a Jacobi matrix and write down the Neumann series

$$(J-z)^{-1} = -\sum_{n=0}^{\infty} \frac{J^n}{z^{(n+1)}}$$

Since J is bounded, for $|z|$ greater than the operator norm of J the Neumann series converges and provides a proper meaning to the left-hand side of the last equation.

We can form the function

$$m(z) = \frac{1}{\pi}\left\langle (J-z)^{-1}e_0, e_0\right\rangle$$

This function is holomorphic in a neighborhood of ∞ on the Riemann sphere since the Neumann series converges there. Moreover,

$$m(z) = -1/z + O(z^{-2})$$

as $z \to \infty$.

We would like to show that m can be extended to the open upper half-plane and has positive imaginary part there. Let z have positive imaginary part and let z_0 be purely imaginary with sufficiently large (in modulus) negative imaginary part. Then the Neumann series

$$(J-z)^{-1} = ((J-z_0)-(z-z_0))^{-1} = \sum_{n=0}^{\infty} -\frac{(J-z_0)^n}{(z-z_0)^{n+1}}$$

converges because

$$|z_0 - z|^2 = |z_0 - \mathrm{Im}(z)|^2 + |\mathrm{Re}(z)|^2 = |z_0|^2 + 2|z_0||\mathrm{Im}(z)| + O(1)$$

while

$$\begin{aligned} \|z_0 - J\|^2 &= \sup_{\|f\|=1} \langle (J-z_0)f, (J-z_0)f\rangle \\ &= \sup_{\|f\|=1} \langle z_0 f, z_0 f\rangle + \langle Jf, Jf\rangle = |z_0|^2 + O(1) \end{aligned}$$

This shows that both $(J-z)^{-1}$ and m extend holomorphically to the upper half-plane.

For $\mathrm{Im}(z) > 0$, let $\phi_z := (J-z)^{-1}e_0$. We observe

$$\mathrm{Im}(m(z)) = \frac{1}{\pi}\langle \phi_z, (J-z)\phi_z\rangle = \frac{1}{\pi}\mathrm{Im}(z)\|\phi_z\|^2$$

Thus m has positive imaginary part in the upper half plane.

By the Herglotz representation theorem there is a positive measure μ on the real line such that

$$m(z) = \frac{1}{\pi}\int \frac{1}{\zeta - z}\, d\mu(\zeta)$$

The measure is compactly supported on $\mathbf{R}$ since m is holomorphic in a neighborhood of ∞.

As $m(z)$ has the asymptotics

$$z^{-1} + O(z^{-2})$$

near ∞, we have $\|\mu\| = 1$.

We prove that μ is not supported on a finite set. The most important ingredient in the following argument is that J has no zeros on the subdiagonal $i = j+1$.

Assume to get a contradiction that the measure is supported on a finite set. Then the integral representation for m shows that m is a rational function and therefore there is a polynomial

$$p(z) = \sum_{k=0}^{m} a_k z^k$$

such that $p(z)m(z)$ is a polynomial. The latter polynomial has the form

$$-\frac{1}{\pi}\sum_{k=0}^{m} a_k z^k \sum_{n=0}^{\infty} z^{-(n+1)}\langle J^n e_0, e_0\rangle$$

Since this is a polynomial, for sufficiently large n the coefficient in the Laurent series has to be zero. This leads to the identity

$$\sum_{k=0}^{m} a_k \left\langle J^{n+k}e_0, e_0\right\rangle = 0$$

for large n. By selfadjointness of J, we also obtain for large n and all $m \geq 0$

$$\sum_{k=0}^{m} a_k \left\langle J^{n+k}e_0, J^m e_0\right\rangle = 0$$

Since the vectors $J^m e_0$ with $m \geq 0$ span the full Hilbert space $l^2(\mathbf{Z}_{\geq 0})$, this gives

$$\sum_{k=0}^{m} a_k J^{n+k} e_0 = 0$$

This is however absurd if J is a Hessenberg matrix with positive elements on the subdiagonal.

To complete the proof of surjectivity, it remains to show that the measure μ that we have constructed gives rise to the matrix J. By construction we have for large $|z|$

$$\int (\zeta - z)^{-1}\, d\mu(\zeta) = \left\langle (J-z)^{-1}e_0, e_0\right\rangle$$

Comparing coefficients in the Neumann series gives for $n \geq 0$

$$\int \zeta^n \, d\mu(\zeta) = \langle J^n e_0, e_0 \rangle$$

By self adjointness of J we also have

$$\int \zeta^n \, d\mu(\zeta) = \langle J^n e_0, J^m e_0 \rangle$$

The same identity holds with J replaced by the matrix $\tilde{J}$ constructed from the measure μ by the Gram-Schmidt orthogonalization process. Therefore, it suffices to show that the numbers

$$\langle J^n e_0, J^m e_0 \rangle \tag{1.48}$$

determine J. We claim that knowing these numbers, we can express all of the standard basis vectors e_n as linear combination of $J^0 e_0, \dots J^n e_0$. This is clear for $n = 0$. Assume by induction we can express $e_0, \dots, e_n$ in terms of $J^0 e_0, \dots, J^n e_0$. We can calculate the coefficients $\left\langle J^{n+1} e_0, e_j \right\rangle$ for $j \leq n$ by expressing e_j in terms of $J^0 e_0, \dots, J^j e_j$ and using (1.48). As $J^{n+1} e_0$ is a linear combination of $e_0, \dots, e_{n+1}$, it remains to determine the coefficient in front of e_{n+1}. This coefficient is positive, hence we can then determine this coefficient by determining the length of $J^{n+1} e_0$, which we can do by (1.48). This poves the claim.

Using the claim, we can calculate $\langle J e_n, e_m \rangle$ for all n, m and thus obtain all coefficients of J. This completes the proof of surjectivity.

□

In the case of measures on the circle we obtain

Theorem 1.39. *The above construction of a Hessenberg matrix provides a bijective correspondence between*

(1) *A positive measure μ supported on the circle $\mathbf{T}$ with $\|\mu\| = 1$ and infinite support.*
(2) *A matrix $J = (J_{ij})_{i,j \geq 0}$ which satisfies $J_{ij} = 0$ for $i > j + 1$, $J_{(j+1),j} > 0$ for all j, and $J^* J = \mathrm{id}$.*

Proof. The proof is a reprise of the arguments in the proof of the previous theorem. We shall only describe the inverse map and leave the other details as an exercise.

We can use Neumann series to define

$$\frac{z + J}{z - J} = (z + J) \sum_{n=0}^{\infty} J^n z^{-(n+1)}$$

for $|z| > 1$. Then we define

$$m(z) = \left\langle \frac{z + J}{z - J} e_0, e_0 \right\rangle$$

on D^*. Setting

$$\phi_z := (z - J)^{-1} e_0$$

we have

$$\begin{aligned} \mathrm{Re} m(z) &= \mathrm{Re} \langle (z + J)\phi_z, (z - J)\phi_z \rangle \\ &= \langle z\phi_z, z\phi_z \rangle - \langle J\phi_z, J\phi_z \rangle = (|z|^2 - 1)\|\phi_z\|^2 \end{aligned}$$

Thus m has positive real part on D^* and by the Herglotz representation theorem is the extension of a unique positive measure μ:

$$m(z) = \int_{\mathbf{T}} \frac{z+\zeta}{z-\zeta}\, d\mu$$

This is the desired measure. $\square$

Assume J is a matrix with $J^*J = 1$ and $J_{ij} = 0$ for $i-1 > j$. Then we have $JJ^*J = J$ and thus $JJ^* = 1$ on the image of J. We claim that the image of J has codimension at most 1, and in case that the codimension is exactly 1, then e_0 complements the image:

$$\mathrm{image}(J) + \mathrm{span}(e_0) = l^2(\mathbf{Z}_{\geq 0})$$

The claim simply follows from the fact that

$$e_0, Je_0, J^2e_0, \dots, J^ne_0$$

evidently span the same space as

$$e_0, e_1, e_2, \dots e_n$$

If the codimension of the image is 0, then J is unitary. This happens if and only if e_0 is in the image of J. A criterion when this happens can be derived from Szegö's theorem stated below. Namely, J is unitary if and only if both sides of the identity in Szegö's theorem are zero.

Theorem 1.40. *(Szegö) Let μ be a measure on $\mathbf{T}$ with a.c. part $wd\theta$ (we denote by $d\theta$ Lebesgue measure on $\mathbf{T}$ normalized so that $\int_{\mathbf{T}} d\theta = 1$). Then*

$$\inf_f \int |1-f|^2\, d\mu = \exp \int_{\mathbf{T}} \log |w|\, d\theta$$

where f runs through all polynomials in z with zero constant term and $\|f\| = 1$ in $L^2(\mu)$.

In case the left-hand side is nonzero, it has the meaning of $|\langle u, 1\rangle|^2$ where u is the unit vector perpendicular to the image of J. The right-hand side becomes zero if $\int \log |w| = -\infty$.

We will prove this theorem in the next section in the case that both sides are finite.

1.18. Orthogonal polynomials on T and the nonlinear Fourier transform

In this section we relate orthogonal polynomials on the circle $\mathbf{T}$ to the nonlinear Fourier transform on the half-line.

The nonlinear Fourier transform (a, b) of a sequence in $l^2(\mathbf{Z}_{\geq 1}, D)$ gives rise to an analytic function b/a^* which maps D to itself and vanishes at 0. Via a Möbius transform of the target space D, such an analytic functions is in unique correspondence with an analytic function m on D which has positive real part and is equal to 1 at 0:

$$m(z) = \frac{1 - b/a^*}{1 + b/a^*}$$

By the Herglotz representation theorem, this function is uniquely associated with a positive measure on $\mathbf{T}$ with total mass 1. By Theorem 1.39, assuming for the moment that this measure does not have finite support, the measure is in unique correspondence with a Hessenberg matrix. The following theorem states that the diagonal and subdiagonal entries of the Hessenberg matrix can easily be expressed in terms of the sequence F.

THEOREM 1.41. *Let $F \in l^2(\mathbf{Z}_{\geq 1}, D)$ and let (a,b) be the nonlinear Fourier transform of F. Let μ be the positive measure on $\mathbf{T}$ whose harmonic extension to D is equal to the real part of*

$$m(z) = \frac{1 - b/a^*}{1 + b/a^*}$$

Then the Hessenberg matrix associated to μ via Theorem (1.39) satisfies

$$J_{i(i-1)} = (1 - |F_i|^2)^{1/2} \tag{1.49}$$

$$J_{ii} = -F_i \overline{F_{i+1}} \tag{1.50}$$

for $i \geq 1$ and

$$J_{00} = -\overline{F_1} \tag{1.51}$$

The main point of this theorem is that taking the forward nonlinear Fourier transform is equivalent to calculating the spectral data (the measure μ) of a Hessenberg matrix, while taking the inverse nonlinear Fourier transform or "layer stripping method" is equivalent to the Gram-Schmidt orthogonalization process.

PROOF. For F a sequence in $l^2(\mathbf{Z}_{\geq 1}, D)$ with nonlinear Fourier transform (a,b), we consider the truncated sequences $F_{\leq n}$ and their nonlinear Fourier transforms

$$(a_{\leq n}, b_{\leq n}) = (a_n, b_n)$$

Define for $n \geq 0$

$$\phi_n(z) = z^n[a_n(z) + b_n^*(z)] \tag{1.52}$$

By Lemma 1.2, ϕ_n is a polynomial in z of exact degree n, and the highest coefficient of ϕ_n is equal to the constant coefficient of a_n and therefore it is positive. In particular, $\phi_0 = 1$.

We can write (1.52) in matrix form (with our standing convention to complete the second row of a matrix by applying the $*$ operation to and reversing the order of the entries of the first row) as

$$(z^{-n}\phi_n, z^n\phi_n^*) = (1,1)(a_n, b_n) \tag{1.53}$$

Thus we obtain a recursion formula by multiplying a transfer matrix from the right:

$$(z^{-(n+1)}\phi_{n+1}, z^{n+1}\phi_{n+1}^*) = (z^{-n}\phi_n, z^n\phi_n^*)(1 - |F_{n+1}|^2)^{-1/2}(1, F_{n+1}z^{n+1})$$

In particular, we have the identity

$$z^{-(n+1)}\phi_{n+1} = (1 - |F_{n+1}|^2)^{-1/2}[z^{-n}\phi_n + \overline{F_{n+1}} z^{-1}\phi_n^*]$$

which can be rewritten as

$$z\phi_n = (1 - |F_{n+1}|^2)^{1/2}\phi_{n+1} - \overline{F_{n+1}} z^n \phi_n^* \tag{1.54}$$

This recursion formula (1.54) together with $\phi_0 = 1$ contains all information on the sequence ϕ_n, but it will be convenient to rewrite the recursion in a different form.

In (1.53), we may also multiply from the right by an inverse transfer matrix to obtain

$$(z^{-(n-1)}\phi_{n-1}, z^{n-1}\phi_{n-1}^*) = (z^{-n}\phi_n, z^n\phi_n^*)(1-|F_n|^2)^{-1/2}(1, -F_n z^n)$$

$$z^{-(n-1)}\phi_{n-1} = (1-|F_n|^2)^{-1/2}[z^{-n}\phi_n - \overline{F_n}\phi_n^*]$$

$$\phi_n = (1-|F_n|^2)^{1/2} z\phi_{n-1} + \overline{F_n} z^n \phi_n^* \tag{1.55}$$

Let the collection of polynomials ϕ_n be fomally an orthonormal basis of a Hilbert space H. Thus abstractly the Hilbert space is the set of all linear combinations of the ϕ_n with square summable coefficients, and the inner product is given by the standard inner product on the space of square summable sequences. Further below we will identify the measure on $\mathbf{T}$ with respect to which the ϕ_n are the Gram-Schmidt orthogonal polynomials, but for now we only need H abstractly defined.

We claim that $T: f \to zf$, originally defined on the set of polynomials, extends to an isometry on H. This will follow once we have shown that the ϕ_n are the Gram-Schmidt orthogonal polynomials of a measure supported on $\mathbf{T}$. However, we find the following proof instructive, and the proof will also help to calculate the diagonal and subdiagonal entries of the Hessenberg matrix representing T.

We shall prove by induction the following two statements

$$\|z\phi_{n-1}\| = 1 \tag{1.56}$$

$$\langle z\phi_{n-1}, \phi_n\rangle = (1-|F_n|^2)^{1/2} \tag{1.57}$$

Observe that equation (1.52) applied for $n=-1$ gives $\phi_{-1} = z^{-1}$. In this sense the above two statements for $n=0$ reduce to the fact $\|\phi_0\| = 1$.

Assume by induction that (1.56) and (1.57) are true for some n. Then by pairing (1.55) with $z\phi_{n-1}$ we see that the two summands on the right-hand side of (1.55) are orthogonal. As the coefficients in (1.55) form a Pythagorean triple, we conclude

$$\|z^n\phi_n^*\| = 1 \tag{1.58}$$

Using this in (1.54) together with the obvious fact that ϕ_{n+1} is orthogonal to $z^n\phi_n^*$, we obtain $\|z\phi_n\| = 1$ and $\langle z\phi_n, \phi_{n+1}\rangle = (1-|F_{n+1}|^2)^{1/2}$. This concludes the induction step.

Now we prove by induction on n that

$$\langle z\phi_n, z\phi_m\rangle = 0 \tag{1.59}$$

for $m < n$. Fix n and assume that (1.59) is true for all indices smaller than n. By eliminating $z^n\phi_n^*$ from (1.54) and (1.55), $z\phi_n$ becomes a linear combination of terms manifestly orthogonal to $z\phi_m$ if $m < n-1$. If $m = n-1$, we use (1.55) and (1.57) to observe that $z^n\phi_n^*$ is perpendicular to $z\phi_{n-1}$ and we insert this into (1.54). This proves (1.59) for the index n.

Therefore, the orthonormal basis ϕ_n is mapped to an orthonormal set via the operation $T: f \to zf$. This proves that T is an isometry.

The operator T in the basis ϕ_n is given by a Hessenberg matrix J, i.e., $J_{ij} = 0$ for $i > j+1$ and $J_{j+1,j} > 0$ (recall the highest coefficient of ϕ_n is positive).

Indeed, we read from the above recursions that the Hessenberg matrix satisfies (1.49), (1.50), and (1.51). Only the proof of (1.50) is slightly more involved. It follows from adding $\overline{F_n}$ times (1.54) and $\overline{F_{n+1}}$ times (1.55) to obtain

$$\overline{F_n}z\phi_n + \overline{F_{n+1}}\phi_n = \overline{F_n}(1-|F_{n+1}|^2)^{1/2}\phi_{n+1} + \overline{F_{n+1}}(1-|F_n|^2)^{1/2}z\phi_{n-1}$$

Taking the ϕ_n component everywhere gives

$$\overline{F_n}\,\langle z\phi_n, \phi_n\rangle + \overline{F_{n+1}} = \overline{F_{n+1}}(1-|F_n|^2)$$

Which proves (1.50).

We know from the abstract theory discussed in the previous section that there is a positive measure μ on the circle with $\|\mu\| = 1$ such that the ϕ_n are the Gram-Schmidt orthogonal polynomials with respect to this measure. We need to show that the Herglotz function m of this measure satisfies

$$m(z) = \frac{1-s(z)}{1+s(z)}$$

where we have set $s = b/a^*$.

We first argue that it suffices to prove the claim under the additional assumption that the sequence F is compactly supported. To pass to the case of general F, we approximate s by s_n. Clearly s_n converges on the disc D pointwise to s since we know convergence of (a_n, b_n) to (a, b) in $\mathbf{H}$. Thus the measures μ_n (defined as above by the function $s_n = b_n/a_n^*$) converge weakly to the measure μ (defined as above by the function s). Observe that the polynomials ϕ_m defined as $z^m(a_m + b_m^*)$ do only depend on the first m coefficients of F and thus are the same as if defined by the truncated sequences as long as the truncation parameter n is larger than or equal to m.

Suppose we can prove that the inner product $\langle \phi_m, \phi_{m'}\rangle$ with respect to the measure μ_n for all sufficiently large n is equal to the Kronecker delta of m and m'. By passing to the weak limit, the inner product with respect to μ is also the Kronecker delta. Thus the ϕ_n are an orthonormal basis for μ.

We now prove the orthonormality property of ϕ_n under the additional assumption that F_n is a finite sequence. Then a, b and m are analytic across the circle $\mathbf{T}$, and the measure μ associated to the Herglotz function m is absolutely continuous with respect to normalized Lebesgue measure and has density $\mathrm{Re}(m)$. We can write

$$\begin{aligned}\mathrm{Re}(m) &= \frac{1}{2}\left[\frac{1-s}{1+s} + \frac{1-s^*}{1+s^*}\right]\\ &= \frac{1-ss^*}{(1+s)(1+s^*)}\\ &= \frac{1}{(a^*+b)(a+b^*)}\end{aligned}$$

We need to show that with respect to this measure,

$$\phi_n = z^n(a_n + b_n^*)$$

is orthogonal to all z^k with $k < n$ and has length 1.

Consider the decomposition (we write a_+ for $a_{>n}$ etc.)

$$\begin{pmatrix} a & b \end{pmatrix} = \begin{pmatrix} a_n & b_n \end{pmatrix}\begin{pmatrix} a_+ & b_+ \end{pmatrix}$$

or equivalently

$$\begin{pmatrix} a & b \end{pmatrix}\begin{pmatrix} a_+^* & -b_+ \end{pmatrix} = \begin{pmatrix} a_n & b_n \end{pmatrix}$$

We obtain

$$\begin{aligned} a_n + b_n^* &= aa_+^* - bb_+^* + b^* a_+^* - a^* b_+^* \\ &= (a + b^*) a_+^* - (a^* + b) b_+^* \end{aligned}$$

And thus

$$\int_{\mathbf{T}} \frac{z^{-k} z^n (a_n + b_n^*)}{(a^* + b)(a + b^*)}$$

$$= \int_{\mathbf{T}} z^{n-k} \left[\frac{a_+^*}{a^* + b} - \frac{b_+^*}{a + b^*} \right] \tag{1.60}$$

However,

$$\frac{a_+^*}{a^* + b}$$

is holomorphic on D and

$$\frac{b_+^*}{a + b^*}$$

is holomorphic on D^* with a zero of order $n+1$ at ∞. Thus the above integral (1.60) is zero for $0 \le k < n$.

For $n = k$ we write for (1.60)

$$\int_{\mathbf{T}} \frac{a_+^*}{a^* + b} - \frac{b_+^*}{a + b^*}$$

$$= \frac{a_+^*(0)}{a^*(0) + b(0)} = \frac{1}{a_n(\infty)}$$

As the highest order coefficient of ϕ_n is $a_n(\infty)$, we obtain

$$\|\phi_n\|^2 = a_n(\infty) \langle z^n, \phi_n \rangle = 1$$

Thus we have verified that μ is the measure with respect to which the ϕ_n are orthonormal.

We remark that if F is a finite sequence and n is the order of the largest nonzero element of F, then the density of the measure is given by

$$\frac{1}{(a_n^* + b_n)(a_n + b_n^*)} = \frac{1}{|\phi_n|^2}$$

□

We are now ready to prove Szegö's theorem under the assumption that F_n is a square integrable sequence and thus $\log |a|$ is integrable.

The measure μ splits as $\mu = \mu_s + \mu_{ac}$ where μ_s is singular with respect to the Lebesgue measure and μ_{ac} is absolutely continuous with density $w \in L^1(\mathbf{T})$.

Since $\mathrm{Re}(m)$ is equal to w almost everywhere on $\mathbf{T}$, we have

$$\int \log w = \int \log\left(\frac{1}{(a^* + b)(a + b^*)}\right)$$

$$= \int \log \frac{1}{|a|^2} + \int \log \frac{1}{1+s} + \int \log \frac{1}{1+s^*}$$

The last two integrals vanish because the function $1/(1+s)$ is holomorphic and outer on D and equal to 1 at 0. Thus

$$\int \log w = -2 \int \log |a|$$

On the other hand, $\inf_f \|1-f\|$ on the left-hand side of Szegö's theorem is equal to the modulus squared of the inner product $\langle u, 1\rangle$ where u is a unit vector perpendicular to the image of $T : f \to zf$. (We shall momentarily establish that under the assumption of square integrable F_n this image has indeed codimension 1.) We claim that the vector u is the strong limit as $n \to \infty$ of the vectors

$$z^n \phi_n^*$$

These vectors are unit vectors by (1.58) and perpendicular to $z\phi_0, \ldots, z\phi_{n-1}$ by (1.55). Thus the strong limit of these vectors has to be a unit vector perpendicular to the image of $T : f \to zf$.

To show the existence of the strong limit, we observe that applying the star operation to (1.55) and multiplying by z^n we obtain

$$z^n\phi_n^* = (1-|F_n|^2)^{1/2} z^{n-1}\phi_{n-1}^* - F_n\phi_n \tag{1.61}$$

This recursion implies that

$$z_n\phi_n^* = \sum_{k=0}^{n} c_{k,n} F_k \phi_k$$

for some constants $c_{k,n}$ bounded by 1. Using orthogonality of the vectors ϕ_n we obtain that the sequence $(z_n\phi_n^*)$ has a limit u.

Using (1.61) and orthogonality of ϕ_n to ϕ_0 for $n > 1$, we obtain

$$\langle u, 1\rangle = \prod_{i=1}^{\infty} (1-|F_n|^2)^{1/2}$$

Thus the left-hand side of Szegö's theorem is $\prod_{i=1}^{\infty}(1-|F_n|^2)$ and the identity of Szegö's theorem follows by the nonlinear Plancherel identity.

This proves Szegös theorem in the setting of square summable F. It can be shown that in case F is not square summable, orthogonal polynomials can still be defined using the Hessenberg matrix associated to the sequence F, and both sides of Szegö's theorem vanish.

1.19. Jacobi matrices and the nonlinear Fourier transform

Orthogonal polynomials on $\mathbf{R}$ and Jacobi matrices are related to the nonlinear Fourier transform on $l^2(\mathbf{Z}_{\geq 1}, [-1,1])$, i.e., the space of bounded real valued sequences F_n on the the half-line. Since Jacobi matrices are a discrete model for Schrödinger operators, the material discussed here also relates to the spectral theory of Schrödinger operators. We only touch upon the subject, entry points to some related current literature are [**19**], [**7**], [**25**].

Our main concern is again to show that on the one hand one can parameterize a class of Jacobi matrices easily by sequences $F \in l^2(\mathbf{Z}_{\geq 1}, [-1,1])$, and on the other hand one can determine the measure μ associated to such a Jacobi matrix easily from the nonlinear Fourier transform (a, b) of F.

We first observe that if F is a sequence in $l^2(\mathbf{Z}, [-1,1])$, then the nonlinear Fourier transform (a, b) of F satisfies

$$a(z^*) = a^*(z), \quad b(z^*) = b^*(z) \tag{1.62}$$

by property (1.9) stated in Lemma 1.1. Conversely, any element $(a, b) \in \mathbf{H}$ satisfying (1.62) has a real sequence $F \in l^2(\mathbf{Z}_{\geq 0}, D)$ as preimage under the nonlinear Fourier transform.

THEOREM 1.42. *Let $F \in l^2(\mathbf{Z}_{\geq 1}, [-1,1])$ and let (a,b) be the nonlinear Fourier transform of F. Let μ be the positive measure on $[-2,2]$ whose harmonic extension to the upper half plane is the imaginary part of the function m defined by*

$$m(w+w^*) = \frac{1}{w-w^*}\frac{1-b(w)/a^*(w)}{1+b(w)/a^*(w)}$$

Then the Jacobi matrix associated to μ via Theorem 1.38 satisfies

$$J_{n,n} = (F_{2n+1}(1+F_{2n}) - F_{2n-1}(1-F_{2n})) \tag{1.63}$$

$$J_{n+1,n} = (1+F_{2n})^{1/2}(1-|F_{2n+1}|^2)^{1/2}(1-F_{2n+2})^{1/2} \tag{1.64}$$

for $n \geq 1$ and

$$J_{1,0} = 2^{1/2}(1-|F_1|^2)^{1/2}(1-F_2)^{1/2} \tag{1.65}$$

$$J_{0,0} = 2F_1 \tag{1.66}$$

We remark that the elements of the sequence F_n with even n enter into the formulas for J_{ij} in a different manner from the elements with odd index. Interesting special cases occur when either all F_n with even index n vanish, or all F_n with odd index n vanish, but we do not further elaborate on this here.

PROOF. We study orthogonal polynomials for measures supported on the interval $[-2,2]$. By a simple scaling argument, it is no restriction if we fix the length of this interval. One can relate these polynomials to orthogonal polynomials on the circle $\mathbf{T}$ using the conformal map

$$w \to y = w + w^*$$

from D to the Riemann sphere slit at $[-2,2]$. This map is sometimes called the Joukowski map. The Joukowski map extends to the boundary $\mathbf{T}$ of D and maps $\mathbf{T} \setminus \{-1,1\}$ two-to-one onto the interval $(-2,2)$ and maps $\{-1,1\}$ to one to one to the endpoints $\{-2,2\}$ of that interval.

Using this map on $\mathbf{T}$, one can push forward measures on $\mathbf{T}$ to measures on $[-2,2]$. This provides a bijection from measures μ' on $\mathbf{T}$ with the symmetry $\mu'(w) = \mu'(w^*)$ to measures μ on $[-2,2]$. The relation between μ' and μ is given by the formula

$$\int f(w+w^*)\, d\mu'(w) = \int f(y)\, d\mu(y)$$

We shall assume that the measure μ' is positive and normalized to $\|\mu'\| = 1$. Then the same normalization holds for μ.

By the Herglotz representation theorem, there is a holomorphic function m' on D whose real part is the harmonic extension of μ'. By the Herglotz representation theorem for the upper half-plane, there is a function m on the upper half plane whose imaginary part is the harmonic extension of the compactly supported measure μ on $\mathbf{R}$. We claim

$$m'(w) = (w-w^*)m(w+w^*)$$

Namely, using the Poisson kernel for the disc, we have

$$m'(w) = \int \frac{v+w}{v-w}\, d\mu'(v)$$

Using the symmetry of μ', we obtain

$$\begin{aligned}
m'(w) &= \frac{1}{2}\int_{\mathbf{T}} \frac{v+w}{v-w} + \frac{v^*+w}{v^*-w}\, d\mu'(v) \\
&= \int_{\mathbf{T}} \frac{1-w^2}{(v-w)(v^*-w)}\, d\mu'(v) \\
&= \int_{\mathbf{T}} \frac{w^*-w}{(vw^*-1)(v^*-w)}\, d\mu'(v) \\
&= (w-w^*)\int_{\mathbf{T}} [(v+v^*)-(w+w^*)]^{-1}\, d\mu'(v) \\
&= (w-w^*)\int_{\mathbf{R}} [y-(w+w^*)]^{-1}\, d\mu(y) \\
&= (w-w^*)m(w+w^*)
\end{aligned}$$

This proves the claim.

We now discuss how the orthogonal polynomials in the variable $y = w + w^*$ can be identified with orthogonal polynomials in the variable w on $\mathbf{T}$ with respect to μ'. As in the previous section, let ϕ_n denote the orthogonal polynomials in the variable w on $\mathbf{T}$ with respect to the measure μ'. Thus

$$\phi_n(w) = w^n[a_n(w) + b_n^*(w)]$$

where (a_n, b_n) are the truncated nonlinear Fourier transforms of a sequence F_n supported on $\mathbf{Z}_{\geq 1}$. Due to the symmetry of μ' we have

$$\phi^*(w) = \phi(w^*), \quad a^*(w) = a(w^*), \quad b^*(w) = b(w^*)$$

and the sequence F_n is real.

As ϕ_{2n} is orthogonal to all monomials of degree up to $2n-1$, the function

$$\psi_n(w) := (w^*)^n \phi_{2n}(w)$$

is orthogonal to all functions w^k with $-n \leq k \leq n-1$. Consequently, ψ_n is also orthogonal to all

$$(w+w^*)^k \quad , \quad 0 \leq k \leq n-1 \tag{1.67}$$

By symmetry under $w \to w^*$, also $\psi_n^*(w)$ is orthogonal to all (1.67), and so is

$$\Psi_n = \psi_n + \psi_n^* = [w^n(a+b^*) + w^{-n}(a^*+b)] \tag{1.68}$$

However, Ψ_n is itself symmetric under $w \to w^*$, and thus a polynomial in $y = w+w^*$ of degree n. Therefore, up to an as of yet unspecified scalar factor, Ψ_n is the n-th orthogonal polynomial with respect to μ.

To determine the scalar factor we calculate

$$\begin{aligned}
&\int \Psi_n(w)\Psi_n^*(w)\, d\mu' \\
&= \int (\psi_n + \psi_n^*)(\psi_n + \psi_n^*) \\
&= 2 + 2\mathrm{Re}\int \psi_n^2\, d\nu \\
&= 2 + 2\mathrm{Re}\int w^{-2n}\phi_{2n}\phi_{2n}
\end{aligned}$$

If $n = 0$, this is simply equal to 4.

Assume $n > 0$. Since ϕ_{2n} is an orthogonal polynomial in the variable w with respect to μ', we see that the last display is equal to

$$= 2 + 2\mathrm{Re}\int \phi_{2n}(w)w^{-2n}c_0 \tag{1.69}$$

where c_0 is the constant coefficient of ϕ_{2n}, which can be obtained from (1.55) and (1.61)

$$c_0 = \overline{F_{2n}}\prod_{k=1}^{2n}(1-|F_k|^2)^{-1/2}$$

However, again by the fact that ϕ_{2n} is an orthogonal polynomial, (1.69) is equal to

$$= 2 + 2\mathrm{Re}\int \phi_{2n}(w)\phi_{2n}^*(w)c_0(\overline{c_{2n}})^{-1}$$

where c_{2n}^{-1} is the highest coefficient of ϕ_{2n}. From (1.61) and the value of c_0 stated above we obtain

$$c_n = \prod_{k=1}^{2n}(1-|F_k|^2)^{-1/2}$$

Thus, since F_n is real,

$$\|\Psi_n\|^2 = 2(1+F_n)$$

Define

$$\Phi_0 = 1$$

and, for $n \geq 1$,

$$\Phi_n = 2^{-1/2}(1+F_{2n})^{-1/2}\Psi_n$$

Then Φ_n is the n-th orthogonal polynomial with respect to μ. Observe that the expression for Φ_n in the case of generic n remains correct if we set $F_0 = 1$.

We calculate the Jacobi matrix associated to the polynomials Φ_n. Assume first $n \geq 1$. We can write for (1.68)

$$(\Psi_n, \Psi_n) = (1,1)(a,b)(w^n,0)(1,1)$$

Then we have the recursion equations

$$\begin{aligned}
&(1-|F_{2n+1}|^2)^{1/2}(1-|F_{2n+2}|^2)^{1/2}(\Psi_{n+1},\Psi_{n+1})\\
&= (1,1)(a,b)(1,F_{2n+1}w^{2n+1})(1,F_{2n+2}w^{2n+2})(w^{n+1},0)(1,1)\\
&= (1,1)(a,b)(w^n,0)(w,F_{2n+1})(1,F_{2n+2})(1,1)
\end{aligned}$$

$$= (1,1)(a,b)(w^n,0)([w+F_{2n+1}][(1+F_{2n+2}],0)(1,1) \tag{1.70}$$

In the last step we have used that for any real number γ and any (a,b) we have

$$(a,b)(\gamma,\gamma) = (a+b,0)(\gamma,\gamma)$$

Similarly,

$$\begin{aligned}
&(1-|F_{2n-1}|^2)^{1/2}(1-|F_{2n}|^2)^{1/2}(\Psi_{n-1},\Psi_{n-1})\\
&= (1,1)(a,b)(1,-F_{2n}w^{2n})(1,-F_{2n-1}w^{2n-1})(w^{n-1},0)(1,1)\\
&= (1,1)(a,b)(w^n,0)(1,-F_{2n})(w^*,-F_{2n-1})(1,1)
\end{aligned}$$

$$= (1,1)(a,b)(w^n,0)(w^*-F_{2n}w-F_{2n-1}+F_{2n}F_{2n-1},0)(1,1) \tag{1.71}$$

Multiplying (1.70) by $(1+F_{2n})/(1+F_{2n+2})$ and adding to (1.71) we obtain on the right-hand side

$$(1,1)(a,b)(w^n,0)(w+w^*+F_{2n+1}(1+F_{2n})-F_{2n-1}(1-F_{2n}),0)(1,1)$$

$$= (w + w^* + F_{2n+1}(1 + F_{2n}) - F_{2n-1}(1 - F_{2n}))(\Psi_n, \Psi_n)$$

where we have pulled a real scalar matrix out of the product. Collecting the terms and expressing Ψ_n in terms of Φ_n gives for $n \geq 2$:

$$(1 + F_{2n})^{1/2}(1 - |F_{2n+1}|^2)^{1/2}(1 - F_{2n+2})^{1/2}\Phi_{n+1}$$
$$+(1 + F_{2n-2})^{1/2}(1 - |F_{2n-1}|^2)^{1/2}(1 - F_{2n})^{1/2}\Phi_{n-1}$$
$$= (w + w^*)\Phi_n + (F_{2n+1}(1 + F_{2n}) - F_{2n-1}(1 - F_{2n}))\Phi_n$$

This identity shows that we can express multiplication by $y = w + w^*$ in the basis Φ_n by a matrix J with

$$J_{n,n} = (F_{2n+1}(1 + F_{2n}) - F_{2n-1}(1 - F_{2n}))\Phi_n$$

and

$$J_{n+1,n} = (1 + F_{2n})^{1/2}(1 - |F_{2n+1}|^2)^{1/2}(1 - F_{2n+2})^{1/2}$$

for $n \geq 1$. To obtain the value for $J_{1,0}$, we review the above calculation for $n = 1$ and observe that it remains correct if Φ_0 is replaced by $2^{1/2}\Phi_0$ in view of the special normalization of Φ_0. Thus

$$J_{21} = 2^{1/2}(1 - |F_1|^2)^{1/2}(1 - F_2)^{1/2}$$

To calculate $J_{0,0}$, we specialize (1.70) to $n = 0$:

$$2^{1/2}(1 - |F_1|^2)^{1/2}(1 - F_2)^{1/2}(\Phi_1, \Phi_1)$$
$$= (1, 1)(w + F_1, 0)(1, 1)$$

Or

$$2^{1/2}(1 - |F_1|^2)^{1/2}(1 - F_2)^{1/2}\Phi_1 = (w + w^*)\Phi_0 + 2F_1\Phi_0$$

Thus

$$J_{0,0} = 2F_1$$

□

1.20. Further applications

1.21. Integrable systems

The linear Fourier transform takes partial differential operators into multiplication operators by coordinate functions. Hence the linear Fourier transform takes simple partial differential equations such as linear equations with constant coefficients into algebraic equations. The latter can often be solved by explicit expressions modulo the task of taking the linear Fourier transform and inverting it. In this section we discuss that the nonlinear Fourier transform can be used similarly to obtain explicit solutions to certain nonlinear partial differential equations, again modulo the task of taking the nonlinear Fourier transform and inverting it. We will present the calculations on a purely formal and thus expository level. They can be made rigorous in appropriate function spaces, e.g., by the precise calculus in [**3**]. The formal calculations for the particular example of the modified Korteweg de Vries equation that we choose for the exposition can be found in [**30**]. A more general discussion can be found in [**15**].

We need the nonlinear Fourier transform of functions on $\mathbf{R}$ and we shall briefly introduce it.

Recall that the linear Fourier transform of sequences is defined by

$$\widehat{F}(\theta) = \sum_{n\in\mathbf{Z}} F_n e^{-in\theta}$$

Here θ lives on the interval $[-\pi, \pi] \subset \mathbf{R}$. To pass to the continuous Fourier transform on $\mathbf{R}$, one can do a limiting process by letting $\theta \in [-\pi/\epsilon, \pi/\epsilon]$ and $n \in \epsilon\mathbf{Z}$ with ϵ approaching 0. Taking an appropriate limit, one obtains the Fourier transform of a function F on $\mathbf{R}$

$$\widehat{F(k)} = \int_{\mathbf{R}} F(x)e^{2ikx}\, dx \tag{1.72}$$

Here we have used a special normalization of $2ikx$ in the exponent of the exponential function, which is maybe unusual but convenient for the discussions to follow.

Now consider $F \in l^2(\mathbf{Z}, D)$, its truncations $F_{\leq n}$ and their nonlinear Fourier transforms $\overbrace{F_{\leq} n} = (a_n, b_n)$. Then we have the recursion equation

$$(a_n(z), b_n(z)) = (a_{n-1}(z), b_{n-1}(z))\frac{1}{1-|F_n|^2}(1, F_n z^n)$$

Subtracting (a_{n-1}, b_{n-1}) on both sides we obtain

$$\begin{aligned} &(a_n(z), b_n(z)) - (a_{n-1}(z), b_{n-1}(z)) \\ &= (a_{n-1}(z), b_{n-1}(z))\left[\frac{1}{1-|F_n|^2}(1, F_n z^n) - (1,0)\right] \end{aligned}$$

A similar type of limiting process as in the linear case, leads to an expression for the nonlinear Fourier transform on $\mathbf{R}$. The discrete variable n becomes a continuous variable $x \in \mathbf{R}$, and the variable z becomes e^{2ik} for some real k, and we obtain

$$\frac{\partial}{\partial x}(a(k,x), b(k,x)) = (a(k,x), b(k,x))(0, F(x)e^{2ikx})$$

In case of compactly supported F, solutions (a, b) to this ordinary differential equation are constant to the left and to the right of the support of F. We denote these constant values by $(a(k,-\infty), b(k,-\infty))$ and $(a(k,\infty), b(k,\infty))$. To obtain the nonlinear Fourier transform, we set the initial value condition

$$(a(k,-\infty), b(k,-\infty)) = (1, 0)$$

and then define

$$\overbrace{F}(k) = (a(k,\infty), b(k,\infty))$$

We review how the linear Fourier transform is used to solve linear constant coefficient PDE. Our example is the Cauchy problem for the Airy equation. Thus the problem is to find a solution $F(x,t)$ to the Airy equation

$$F_t = F_{xxx}$$

(where a variable in the index denotes a partial derivative) with the initial condition

$$F(0, x) = F_0(x)$$

for some given function F_0. Taking formally the linear Fourier transform of F in the x variable one obtains a function $\widehat{F}(t,k)$ satisfying

$$\widehat{F}_t = (-2ik)^3\widehat{F} = 8ik^3\widehat{F}$$

$$\widehat{F}(0,k) = \widehat{F_0}(k)$$

For fixed k this is an ordinary differential equation in t which has the solution

$$\widehat{F}(t,k) = e^{8ik^3t}\widehat{F_0}(k)$$

Taking the inverse Fourier transform, one obtains

$$F(t,x) = (e^{8ik^3t}\widehat{F_0}(k))^{\check{}}$$

Aanlogously, the nonlinear Fourier transform can be used to solve certain nonlinear partial differential equations. As an example we discuss the Cauchy problem for the modified Korteweg-de Vries (mKdV)

$$F_t = F_{xxx} + 6F^2F_x \tag{1.73}$$
$$F(0,x) = F_0(x)$$

Observe that the mKdV equation is a perturbation of the Airy equation by the nonlinear term $6F^2F_x$.

We take the nonlinear Fourier transform of the initial data:

$$\overbrace{F_0}(k) = (a(k), b(k))$$

Then the solution to the mKdV equation is formally given by

$$\overbrace{F}(t,k) = (a(k), e^{8ik^3t}b(k)) \tag{1.74}$$

Therefore, safe for the task of taking a nonlinear Fourier transform and an inverse nonlinear Fourier transform, this is an explicit solution.

We outline a proof of (1.74) on a formal level. The argument can be made rigorous in approriate function spaces.

A Lax pair is a pair of time dependent differential operators $L(t), P(t)$ in spatial variables such that L is selfadjoint, P is anti-selfadjoint, and

$$\frac{d}{dt}L(t) = [P(t), L(t)] = P(t)L(t) - L(t)P(t)$$

This Lax pair equation implies that eigenvectors of L are preserved under the flow of P. More precisely, this means that if we have a solution ϕ to the evolution equation

$$\frac{d}{dt}\phi(t) = P(t)\phi(t)$$

and at time $t = t_0$ we have

$$L(t_0)\phi(t_0) = \lambda\phi(t_0)$$

then we also have

$$L(t)\phi(t) = \lambda\phi(t)$$

for all t. Namely,

$$\begin{aligned}\frac{d}{dt}[L\phi - \lambda\phi] &= [P, L]\phi + LP\phi - \lambda P\phi \\ &= P[L\phi - \lambda\phi]\end{aligned}$$

Thus if $L\phi - \lambda\phi$ vanishes for some t_0, then it vanishes for all time.

We introduce the Lax pair which is useful for the mKdV equation. The operators L and P are two by two matrices of differential operators. For some real function $F(t,x)$ define the selfadjoint operator

$$L(t) = \begin{pmatrix} 0 & -i(\frac{\partial}{\partial x} + F(t,x) \\ i(-\frac{\partial}{\partial x} + F(t,x) & 0 \end{pmatrix}$$

where F denotes the operator of multiplication by F. We remark that this operator is called a Dirac operator, since it is a square root of a Schrödinger operator:

$$L^2(t) = \begin{pmatrix} -\frac{\partial^2}{\partial x^2} + F_x + F^2 & 0 \\ 0 & -\frac{\partial^2}{\partial x^2} - F_x + F^2 \end{pmatrix}$$

Thus L^2 separates into two operators that are of Schrödinger type.

We consider the eigenfunction equation for L:

$$L\phi = k\phi$$

If we make the ansatz

$$\phi(t,k,x) = \begin{pmatrix} a(t,k,x)e^{ikx} + b(t,k,x)e^{-ikx} \\ a(t,k,x)e^{ikx} - b(t,k,x)e^{-ikx} \end{pmatrix} \tag{1.75}$$

then the eigenfunction equation for L turns into the ordinary differential equation

$$\frac{\partial}{\partial x}(a,b) = (a,b)(0, Fe^{2ikx})$$

used to define the nonlinear Fourier transform.

Define the anti-selfadjoint operator $P = P(t)$ by

$$P = \begin{pmatrix} 4\frac{\partial^3}{\partial x^3} + 3\left\{\frac{\partial}{dx}, F_x + F^2\right\} - 4(ik)^3 & 0 \\ 0 & 4\frac{d^3}{dx^3} + 3\left\{\frac{\partial}{\partial x}, -F_x + F^2\right\} - 4(ik)^3 \end{pmatrix}$$

where $\{A,B\} = AB + BA$ denotes the anti-commutator of A and B. The operator $P(t)$ depends on the parameter k, but only through an additive multiple of the identity matrix which vanishes upon taking a commutator.

Using some elementary algebraic manipulations, the Lax pair equation

$$\frac{d}{dt}L = [P, L]$$

turns into the mKdV equation (1.73) for F. Thus L and P as above are a Lax pair precisely when F satisfies the mKdV equation.

With the ansatz (1.75), the evolution equation for ϕ,

$$\frac{d}{dt}\phi = P\phi$$

becomes a partial differential equation for a and b.

We shall now assume that $F(t,x)$ is compactly supported in x and remains in a fixed compact support as t evolves. This assumption is only good for the purpose of an exposition of the main ideas. In reality no solution to the mKdV equation remains supported in a compact set under the time evolution, so in a rigorous argument one needs to discuss asymptotic behaviour of the solutions for large $|x|$.

To the right and to the left of the support of F, the functions a and b are constant and we write $a(t,k,\pm\infty)$ and $b(t,k,\pm\infty)$ for the values to the right and left of the support of F. Outside the support of F, the partial differential equations for for $a(t,k,\pm\infty)$ and $b(t,k,\pm\infty)$ become

$$\frac{d}{dt}a(t,k,\pm\infty)e^{ikx} = \left[4\frac{d^3}{dx^3} - 4(ik)^3\right] a(t,k,\pm\infty)e^{ikx}$$

$$\frac{d}{dt}b(t,k,\pm\infty)e^{-ikx} = \left[4\frac{d^3}{dx^3} - 4(ik)^3\right] b(t,k,\pm\infty)e^{-ikx}$$

The right-hand side of the equation for a vanishes and thus we obtain that a remains constant:

$$a(t,k,\pm\infty) = a(0,k,\pm\infty)$$

The equation for b reduces to

$$\frac{d}{dt}b(t,k,\pm\infty) = 8ik^3 b(t,k,\pm\infty)$$

which has the solution

$$b(t,k,\pm\infty) = b(0,k,\pm\infty)e^{8ik^3 t}$$

We now specialize to the solutions to the eigenfunction equation for L such that

$$a(t,k,-\infty) = 1$$

$$b(t,k,-\infty) = 0$$

This is consistent with the evolution calculated above. Thus we have

$$\widehat{F}(t,k) = (a(t,k,\infty), b(t,k,\infty)) = (a(0,k,\infty), b(0,k,\infty)e^{8ik^3 t})$$

This is the explicit form of the solution F to the mKdV equation that we claimed.

1.22. Gaussian processes

We shall very briefly discuss the link between the nonlinear Fourier transform and stationary Gaussian processes. A detailed account on the subject of Gaussian processes can be found in [**12**].

Probability theory in the view of an analyst is a theory of measure and integration where the underlying measure spaces are hidden as much as possible in language and notation. For probabilists' intuition, the underlying measure spaces are uninteresting. Many statements in probability theory are fairly independent of the special structure of the underlying measure space.

A random variable f is a measurable function on some measure space of total measure 1. Analysts would write $f(x)$ referring to an element x of the underlying measure space. The integral of this function over the measure space is called the mean or the expectation $E(f)$ of f. Analysts would write $\int f(x)d\mu(x)$ referring to the measure μ. A collection of random variables living on the same measure space is called a family of random variables. The measure of a set in the measure space is called the probability P of the set. Inded, since the set is usually discribed by conditions on one or several random variables, the set is called the "event" that the random variables satisfy these conditions.

For example, one writes

$$P(f > \lambda)$$

for the measure $\mu(\{x : f(x) > \lambda\})$ and calls it the probability of the event $f > \lambda$. The function $P(f > \lambda)$ in λ is called the distribution function of the random variable f.

A real valued random variable is called Gaussian, if

$$P(f > \lambda) = c\int_\lambda^\infty e^{-(s-s_0)^2/2\sigma}\, ds$$

where c is normalized so that $P(f > -\infty) = 1$. Observe that for a Gaussian variable, the distribution function is determined by the mean value $E(f) = s_0$ and the variance

$$E((f - E(f))^2) = E(f^2) - E(f)^2 = \sigma$$

A Gaussian family indexed by $\mathbf{Z}$ consists of a family of random variables

$$f_n, \quad n \in \mathbf{Z}$$

such that each f_n is Gaussian distributed with mean zero and also each finite linear combination

$$f = \sum_{n \in \mathbf{Z}} \gamma_n f_n$$

is Gaussian distributed.

In particular, $E(f_n) = 0$ for all n. By linearity of the expectation, $E(f) = 0$ for all finite linear combinations f as above. Therefore, the distribution of each linear combination f is determined by the variance $E(f^2)$. By linearity of the expectation, we have the formula

$$E(f^2) = \sum_{n,m \in \mathbf{Z}} \gamma_n \gamma_m E(f_n f_m)$$

On the space of finite sequences (γ_n), identified as random variables f as above, we can define an inner product of two elements f and f' by $E(ff')$. It is positive definite since $E(f^2) > 0$ for all f. Let H be the Hilbert space closure of this inner product space. The elements of this Hilbert space are again random variables.

Orthogonality in this space translates to the probabilistic notion of independence. Independence means a factorization of the distribution functions:

$$P(f > \lambda, f' > \lambda') = P(f > \lambda) P(f' > \lambda')$$

If two random variables are orthogonal in H, then

$$E(\eta_1 \eta_2) = 0 = E(\eta_1) E(\eta_2)$$

and $E(\eta_1 \eta_2) = E(\eta_1) E(\eta_2)$ is a necessary condition for independence. In the setting of Gaussian variables, it is also sufficient for independence.

If H' is a subspace of H, then every $f \in H$ can be split as

$$f = f' + f''$$

where f' is in H and f'' is in the orthogonal complement of H'. The probabilistic interpretation is that f' is known if all elements in H are known, while f'' is independent of any knowledge about H'.

A stationary Gaussian process is one for which

$$E(f_n f_m) = Q(n - m)$$

for some sequence of numbers $Q \in l^\infty(\mathbf{Z})$. Equivalently, a Gaussian process is stationary if it is equal to the shifted proces $\tilde{f}_n = f_{n-1}$. In particular, the length of the vectors f_n in a stationary Gaussian process is independent of n.

We observe that $Q(n) = E(f_n, f_0) \leq E(f_0, f_0) = Q(0)$ for all n. This is a special case of the property

$$\sum_{n,m} Q(n - m) \gamma_n \gamma_m \geq 0$$

for all finite sequences γ_n. Indeed, the left-hand side has the meaning of $E(f^2)$ for some f and is therefore non-negative. This property is called positive definiteness of the sequence Q. The following theorem by Bochner characterizes all positive definite sequences.

THEOREM 1.43. *A nonzero sequence Q_n satisfies*

$$\sum_{n,m} Q(n-m)\gamma_n\gamma_m \geq 0$$

for all real valued finite sequences γ_n if and only if it is the Fourier series of a positive measure on $\mathbf{T}$:

$$Q_n = \int_{\mathbf{T}} z^n \, d\mu$$

We only prove one direction of the theorem. Assume Q_n is the Fourier series of a positive measure. Then

$$\sum_{n,m} Q(n-m)\gamma_n\gamma_m = \sum_{m,n} \int z^{n-m}\gamma_n\gamma_m \, d\mu$$

$$= \int |\sum_n \gamma_n z^n|^2 \, d\mu \geq 0$$

This proves one direction of Bochner's theorem.

Now there is an evident isometric isomorphism from H to $L^2(\mu)$. It maps the elements f_n to the function z^n. Isometry is seen as follows:

$$E(f_n, f_m) = Q(n-m) = \int z^{n-m} \, d\mu = \langle z^n, z^m \rangle_{L^2(\mu)}$$

Surjectivity follows from the definition of $L^2(\mu)$ as the closure of the linear span of the monomials z^n.

The space $L^2(\mu)$ takes us into the setting of orthogonal polynomials on the circle $\mathbf{T}$. Indeed, by reflection we consider polynomials in z^{-1} and have the following immediate consequence of Szegö's theorem:

COROLLARY 1.44. *The past, namely the span of $f_{-1}, f_{-2}, \dots$ determines the present f_0, i.e., f_0 is in the closed span of $f_{-1}, f_{-2}, \dots$ if and only if*

$$\int_{\mathbf{T}} \log|w| = -\infty$$

where w is the absolutely continuous part of μ.

1.23. Appendix: Some Background material

This lecture has two sections. The first section gives some background material on boundary regularity of harmonic and holomorphic functions on the unit disc. This section contains several theorems that are important for a rigorous development of the nonlinear Fourier transform.

The second section recalls some facts about the group $Sl_2(\mathbf{R})$ and the isomorphic group $SU(1,1)$. This section is meant to help understand some of the algebraic manipulations done in this lecture series on a group theoretical level, but otherwise is somewhat irrelevant for the overall understanding of the lecture series.

1.24. The boundary behaviour of holomorphic functions

We shall study classes of holomorphic functions on the unit disc defined by some size control.

For example, for any monotone increasing function $\phi : \mathbf{R}_{\geq 0} \to \mathbf{R}_{\geq 0}$ we can consider the space of holomorphic functions

$$\{f : \sup_{r<1} \int_{\mathbf{T}} \phi(|f(r\cdot)|) < \infty\} \tag{1.76}$$

Here $\int_{\mathbf{T}}$ denotes the integral over the circle $\mathbf{T}$, i.e., the set of all $z \in \mathbf{C}$ with $|z| = 1$, with the usual Lebesgue measure on $\mathbf{T}$ normalized to have total mass 1.

$$\int_{\mathbf{T}} f := \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta})\, d\theta$$

Thus for fixed r the integral in (1.76) has the meaning of an average over the circle of radius r about the origin.

We shall mainly be interested in $\phi(x) = x^p$ (producing Hardy spaces) and $\phi(x) = \log_+(x) = \max(0, \log |x|)$ (producing Nevanlinna class).

The largest space we consider and the space containing all functions we shall be concerned with is the Nevanlinna class:

$$N = \{f : \sup_{r<1} \int_{\mathbf{T}} \log_+ |f(r\cdot)| < \infty\}$$

This space can be identified with a space of almost everywhere defined functions on the circle, as the following theorem describes:

Theorem 1.45. *If $f \in N$, then f has radial limits*

$$\lim_{r\to 1} f(rz)$$

for almost every z in $\mathbf{T}$. (Here r is real and less than 1.) If the radial limits of two functions $f_1, f_2 \in N$ coincide on a subset of $\mathbf{T}$ of positive measure, then the functions are equal.

The proof of this theorem will be discussed below.

The uniqueness result is not special to Nevanlinna class, provided one passes to the notion of nontangential limits. A precise version of this statement is given in the theorem below. The previous theorem also holds with "nontangential" in place of "radial", but we shall not need this.

Theorem 1.46. *Assume that two holomorphic functions on the disc each have nontangential limits on sets of positive measure, and the limits coincide on a set of positive measure, then the two functions are equal.*

This statement is false if "nontangential" is replaced by "radial".

For a proof of this theorem and a discussion of nontangential limits see Garnett's book [**16**].

Thus we can talk about "the holomorphic extension" of a function defined on a positive set on $\mathbf{T}$, provided such an extension exists.

One of the reasons for us to try to identify holomorphic functions with their limits on the boundary is that on the Riemann sphere, the circle $\mathbf{T}$ is the boundary of the unit disc about 0 and the boundary of the unit disc about ∞. We would like to study simultaneously and compare the spaces of holomorphic functions on both discs. The boundary values of the functions on $\mathbf{T}$ are the only link between the two spaces.

While knowledge of the real part of a holomorphic function on the disc is sufficient to determine the imaginary part up to an additive constant, the above uniqueness result heavily relies on the fact that we know the limits of both the real and imaginary part. In particular, the analogue statement fails for harmonic functions, as the example $\mathrm{Re}(\frac{z+1}{z-1})$ shows, which has limit 0 almost everywhere on the circle $\mathbf{T}$ but clearly is not zero.

Thus the notion of harmonic extension cannot be used as freely as that of holomorphic extension, indeed it cannot be used easily in the context of functions defined almost everywhere. However, harmonic extensions of measures are well defined. This is described in the following theorem:

THEOREM 1.47. *Given a complex Borel measure μ on $\mathbf{T}$ (an element in the dual space of the space of continuous functions with the supremum norm), the function*

$$f(z) = \int P_z \, d\mu$$

where P_z is the Poisson kernel,

$$P_z(\zeta) = \mathrm{Re}\left(\frac{\zeta + z}{\zeta - z}\right)$$

is a harmonic function on the disc. Radial limits of this function exist almost everywhere and coincide almost everywhere with the density $f \in L^1$ of the absolutely continuous part of μ, i.e.

$$\mu = \frac{1}{2\pi} f d\theta + \mu_s$$

where μ_s is singular with respect to Lebesgue measure $d\theta$

PROOF. The kernel P_z is harmonic in $z \in D$, and thus the superposition $\int P_z \, d\mu$ of harmonic functions is again harmonic (use the mean value characterization of harmonicity and Fubini's theorem). To study the existence and behaviour of radial limits, it suffices to consider separately the cases of μ absolutely continuous and μ singular with respect to Lebesgue measure. If μ is absolutely continuous, $\mu = \frac{1}{2\pi} f d\theta$, we estimate

$$|f(z) - \int_{\mathbf{T}} P_{rz} f| \leq |f(z) - g(z)| + |g(z) - \int_{\mathbf{T}} P_{rz} g| + |\int_{\mathbf{T}} P_{rz}(f-g)|$$

$$\leq |f(z) - g(z)| + |g(z) - \int_{\mathbf{T}} P_{rz} g| + C|M(f-g)(z)|$$

where M is the Hardy - Littlewood maximal function, which up to a constant factor C dominates the integration against the Poisson kernel independently of r, and g is some appropriate smooth function. If g is sufficiently close to f in L^1 norm, then outside a small set the difference $|f - g|$ is small. Outside a possibly different small set, $M(f - g)$ is small by the Hardy - Littlewood maximal theorem. The term $g(z) - \int_{\mathbf{T}} P_{rz} g$ can be made small by choosing r close to 1 depending on the choice of the smooth function g. Making this argument rigorous using the correct quantifyers one proves convergence of $\int_{\mathbf{T}} P_{rz} f$ to $f(z)$ outside a set of arbitrarily small measure, i.e., almost everywhere.

If μ is singular, one proves using a Vitali - type covering lemma that μ has vanishing density almost everywhere,

$$\lim_{h\to 0} \mu([t-h, t+h])/2h = 0$$

for almost every t. For points t of zero density one then observes that $P_r * f(t)$ tends to 0. □

Observe that if μ is absolutely continuous with a continuous density function, then one can prove that the harmonic function f defined in the above theorem has continuous extension to $D \cup \mathbf{T}$. On $\mathbf{T}$ the function f coincides with the density of μ. Moreover, by the maximum principle, this extension is the unique harmonic function which has continuous extension to $\mathbf{T}$ coinciding with the density function of μ.

The following is a variant of the above theorem:

THEOREM 1.48. *Given a real measure μ on $\mathbf{T}$, the function*

$$f(z) = \int \mathrm{Im}(\frac{\zeta + z}{\zeta - z})\, d\mu(\zeta)$$

is a harmonic conjugate to the function defined in the previous theorem in the unit disc. Its radial limits exist almost everywhere and are equal almost everywhere to the Hilbert transform of μ.

PROOF. Harmonicity (holomorphicity) follows again by characterizing harmonic (holomorphic) functions by mean value (Cauchy) integrals and then using Fubini's theorem and the fact that the kernel $\frac{\zeta+z}{\zeta-z}$ is holomorphic. To see that radial limits exist, one has to study convergence of the conjugate Poisson kernels, which amounts to estimating maximal truncated singular integrals. We leave details as an exercise. □

There is a nice intrinsic characterization of those harmonic functions which are extensions of positive measures, due to Herglotz [**17**].

THEOREM 1.49. *Any positive harmonic function is the harmonic extension of a unique positive measure.*

We shall sometimes call a holomorphic function whose real or imaginary part is positive a Herglotz function.

A real harmonic function is the extension of a measure if it can be written as $f_1 - f_2$ with two positive harmonic functions.

PROOF. If f is a positive harmonic function, then for each radius r there is a positive measure μ_r on T with density $f(rz)$. Let μ be a weak-$*$ limit as $r \to 1$ of an appropriate subsequence of this collection of measures (each of them has total mass equal to $f(0)$ by the mean value property). By weak convergence, the Poisson extension of μ is equal to the pointwise limit of the Poisson extensions of μ_r, and thus equal to f. This proves existence of a measure whose Poisson extension is f. If μ_1 and μ_2 are two measures with the same harmonic extension, then the difference measure has vanishing Fourier coefficients (this follows from calculating the Taylor coefficients of the harmonic extension at 0), and thus is zero. □

Just as the real and imaginary parts of boundary values of a holomorphic function do not individually determine the function, neither does the absolute value of the boundary values. Indeed, the absolute value of the boundary value function a.e. does not even in general determine membership in a space. For Nevanlinna class, observe that Fatou's theorem implies (let f also denote the boundary value function on the circle)

$$\int_{\mathbf{T}} \log_+ |f| \leq \sup_{r<1} \int_{\mathbf{T}} \log_+ |f(r\cdot)|$$

Thus

$$\int_{\mathbf{T}} \log_+ |f| < \infty$$

is a necessary condition for f to belong to N. But it is not a sufficient condition since Fatou's inequality can be strict and finiteness of

$$\int \log_+ |f|$$

does not imply membership in Nevanlinna class. An example is the function $\exp(\frac{1+z}{1-z})$, which has absolute value 1 almost everywhere on the circle $\mathbf{T}$ but is not Nevanlinna class.

Similar statements hold for most function spaces, in particular the Hardy spaces.

Let us note another application of Fatou's theorem to Nevanlinna class functions.

LEMMA 1.50. *If $f \in N$ and $f \neq 0$, then $\log |f|$ is absolutely integrable on $\mathbf{T}$.*

The point of the lemma is that originally we control only the positive part of $\log |f|$, but via the lemma we also control the negative part.

PROOF. By dividing by a power of z if necessary, we may assume $f(0) \neq 0$ (here we use $f \neq 0$). As $\log |f(z)|$ is subharmonic in the disc, we have

$$\log |f(0)| \leq \int \log |f(r\cdot)| = \int \log_+ |f(r\cdot)| + \int \log_- |f(r\cdot)|$$

The first inequality is a consequence of subharmonicity and can be shown by Green's theorem (analoguously to the proof of the mean value property of harmonic functions), using the fact that the distributional Laplacian of $\log |f|$ is a positive measure (instead of zero as for harmonic functions).

Using the Nevanlinna class assumption we observe

$$\log |f(0)| - C \leq \int \log_- |f(r\cdot)|$$

and so Fatou's lemma implies

$$\log |f(0)| - C \leq \int \log_- |f|$$

□

Observe that this lemma also proves the uniqueness part of Theorem 1.45. Since the difference of two Nevanlinna functions is Nevanlinna again, it suffices to prove that if the limit function vanishes on a set of positive measure, then it is constant 0. However, if the limit function vanishes on a set of positive measure, then $\log |f|$ is not integrable on the circle, so by the lemma f is identically equal to 0.

When estimating the size of holomorphic functions, it is natural to consider the logarithm of $|f|$, which produces a harmonic function if f is zero free and a subharmonic function if f has zeros. Thus Nevanlinna class functions are related to harmonic extensions because for zero free Nevanlinna functions $\log |f|$ is the harmonic extension of a measure.

LEMMA 1.51. *If $f \in N$ then f can be factored as $f_1 f_2$ where f_1 is an analytic function in D bounded by 1 and $\log |f_2|$ is the harmonic harmonic extension of a measure. If f has no zeros, we may pick $f_1 = 1$.*

PROOF. As before, we consider the measures μ_r defined by

$$\int_{\mathbf{T}} g d\mu_r = \int_{\mathbf{T}} g(\cdot) \log_+ |f(r\cdot)|$$

By the Nevanlinna property of f, these measures have bounded total mass (independent of r), and thus there is a weak-$*$ limit, μ, of a subsequence of these measures. Let f_2 be a function in N such that $\log |f_2|$ is the harmonic extension of μ. Using subharmonicity of f and a limiting process, one can show that f_2 dominates f, thereby proving the above theorem.

If f has no zeros, then f/f_2 has no zeros, and $\log(f/f_2)$ is a holomoprhic function with negative real part. Thus its real part is the harmonic extension of a negative measure. Thus $\log(f)$ is the harmonic extension of a measure. □

We can now prove existence of radial limits almost everywhere for any Nevanlinna function f. It suffices to prove existence for f_1 and f_2 where $f = f_1 f_2$ is a splitting as in the last lemma. By adding a constant to f_1, we may assume that f_1 has positive real part. For such functions we proved existence of the radial limits almost everywhere before. The function f_2 has no zeros, and thus $\log(f_2)$ is holomorphic and its real part is positive. Thus radial limits of f_2 exist.

The bounded function f_1 cannot be omitted from the last theorem, because f may have zeros. Using Blaschke products, one may choose f_1 to be a possibly infinite Blaschke product. This can be deduced from the following lemma:

LEMMA 1.52. *Let $f \in N$ and let z_n be the zeros of f (multiple zeros appear in the sequence according to multiplicity). Then*

$$\sum_n (1 - |z_n|) < \infty$$

Conversely, if z_n is any such sequence with 0 appearing m times, then the (Blaschke) product

$$B(z) = z^m \prod_{z_n \neq 0} \frac{-\overline{z_n}}{|z_n|} \frac{z - z_n}{1 - z z_n}$$

converges uniformly on compact subsets of D to a bounded analytic function with exactly the zeros z_n. The radial limits of B on $\mathbf{T}$ have modulus 1 almost everywhere on $\mathbf{T}$.

PROOF. See Garnett's book [**16**]. □

We define an outer function to be a Nevanlinna function without zeros such that $\log|f|$ is the harmonic extension of an absolutely continuous measure. Thus $\log|f|$ is the Poisson integral of its a.e. radial limits.

Outer functions are very special functions: they are determined (up to a constant phase factor) by the modulus of their limits almost everywhere.

This relates to the following lemma, which is a form of an "inverse Fatou" for outer functions.

LEMMA 1.53. *If the boundary value functions of an outer function is in L^p, then the outer function is in H^p.*

PROOF. By Poisson extension we have

$$\log|f(z)| = \int P_z(.)\log|f(.)|$$

By convexity of the function e^{rx} and Jensen's inequality we have

$$|f(z)|^p \leq \int P_z(.)|f(.)|^p$$

This implies that the restrictions of f to smaller circles are uniformly in L^p. $\square$

We will use the following criterion for outerness.

LEMMA 1.54. *Let f and $1/f$ be in $H^p(D)$ for some $p > 0$. Then f is outer.*

PROOF. Since f and $1/f$ are hoplomoprhic in D, we can choose a branch of $\log(f)$ which is holomorphic in D. The L^p estimates for f and f^{-1} can be used to obtain L^2 estimates of $\log(f)$ on circles of radius r about 0, uniformly in r. Thus $\log(f)$ in $H^2(D)$, which implies it is the Poisson extension of its boundary values. This implies that f is outer. $\square$

1.25. The group $Sl_2(\mathbf{R})$ and friends

The general linear group $Gl_2(\mathbf{C})$ consists of all 2×2 invertible complex matrices with the usual matrix product as group multiplication.

This group acts on the complex vector space $\mathbf{C}^2$ by linear transformations

$$\begin{pmatrix} u \\ v \end{pmatrix} \to \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}$$

Indeed, $Gl_2(\mathbf{C})$ can be identified as the group of linear automorphisms of $\mathbf{C}^2$. The projective (complex) line P^1 is the set of all complex lines in $\mathbf{C}^2$ of the form

$$L_{\alpha,\beta} = \{(\alpha z, \beta z), z \in \mathbf{C}\}$$

with parameters (α, β) and $\alpha\beta \neq 0$.

The projective line P_1 is a complex manifold with two charts $\mathbf{C} \to P^1$ given by $z \to L_{z,1}$ and $z' \to L_{1,z'}$. The first chart misses only the line $L_{1,0}$ and the second chart misses only the line $L_{0,1}$. The transition between the two charts is given by $z = 1/z'$.

Thus P^1 is isomorphic to the Riemann sphere. We shall mainly use the first chart described above and write $z = \infty$ for the line $L_{1,0}$.

The action of $Gl_2(\mathbf{C})$ on P_1 is then given by the linear fractional transformation

$$z \to \frac{az + b}{cz + d}$$

Thus the action is by biholomorphic maps (Möbius transforms) of the Riemann sphere. Indeed, every biholomorphic self map of the Riemann sphere has to be a fractional linear transformation (exercise), and thus be given by action of an element in $Gl_2(\mathbf{C})$. Thus we have a surjection of $Gl_2(\mathbf{C})$ onto the set of Möbius transforms of the Riemann sphere.

Two elements in $Gl_2(\mathbf{C})$ give the same Möbius transform, if and only if they are scalar multiples of each other. (The trivial action only comes from the scalar matrices). The group of matrices in $Gl_2(\mathbf{C})$ with determinant one is called $Sl_2(\mathbf{C})$. Since every matrix can be normalized to have determinant one (if $\det(g) = \lambda$, then $\det(\nu g) = 1$ if $\nu^2 = \lambda^{-1}$, an equation that can always be solved for ν in the complex numbers), $Sl_2(\mathbf{C})$ still covers all Möbius transforms. However, the two elements id and $-$id of $Sl_2(\mathbf{C})$ both map onto the identity Möbius transform. Thus $Sl_2(\mathbf{C})$ is a double cover of the group of Möbius transforms. The quotient of $Sl_2(\mathbf{C})$ by the central subgroup with two elements id and $-$id is called $PSL_2(\mathbf{C})$. The group of Möbius transforms of the sphere is isomorphic to $PSL_2(\mathbf{C})$.

The group of Möbius transforms which leave the real line (a great circle) on the Riemann sphere invariant, has to come from an automorphism of $\mathbf{C}^2$ which maps real vectors in $\mathbf{C}^2$ to real vectors, and thus has to come from a real linear automorphism of $\mathbf{R}^2$.

These maps precisely give the matrices in $Gl_2(\mathbf{C})$ with real entries. The group of these matrices is $Gl_2(\mathbf{R})$. This group has two connected components, the component of elements with positive determinant and the component of elements with negative determinant. The elements of the first component map the upper half plane (positive imaginary part) to the upper half plane, the elements of the other component map the upper half plane to the lower half plane. The group of matrices in $Gl_2(\mathbf{R})$ with determinant 1 is called $Sl_2(\mathbf{R})$. The group can be identified as a double cover of all biholomorphic self maps of the upper half plane. The quotient by the central subgroup of two elements is called $PSl_2(\mathbf{R})$.

By conformal equivalence, more precisely by a rotation of the Riemann sphere, the Möbius transforms of the upper half plane correspond to the Möbius transforms on the unit disc $D = \{z : |z| < 1\}$. The latter are matrices in $Sl_2(\mathbf{C})$ which leave the set of vectors of the form $(e^{i\phi}z, z)$ with $\phi \in \mathbf{R}$ invariant. These vectors are exactly the null vectors of the quadratic form

$$B(u, v) = |u|^2 - |v|^2$$

A linear map preserving the null vectors of this form has to leave the whole quadratic form invariant up to a scalar multiple. By the determinant constraint, this multiple has to be 1 or -1, again corresponding to the maps which map inside of the unit circle to inside, or inside to outside respectively. The matrices in $Sl_2(\mathbf{C})$ which leave B invariant are precisely those that can be written in the form

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix}$$

with $|a|^2 - |b|^2 = 1$. These matrices form the group $SU(1, 1)$. It is a double cover of $PSU(1, 1)$, which is $SU(1, 1)/\{\mathrm{id}, -\mathrm{id}\}$. By the above discussion this group is isomorphic to $Sl_2(\mathbf{R})$, and an explicit isomorphism can be given by conjugating with a Möbius transform mapping the upper half plane to the disc. An explicit isomorphism is given by

$$Sl_2(\mathbf{R}) \to SU(1, 1)$$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \to \begin{pmatrix} (a+d)/2 + i(b-c)/2 & (a-d)/2 + i(b+c)/2 \\ (a-d)/2 - i(b+c)/2 & (a+d)/2 - i(b-c)/2 \end{pmatrix}$$

Let us discuss eigenvalues of matrices in $Sl_2(\mathbf{R})$. The eigenvalues of a matrix G in $Sl_2(\mathbf{R})$ satisfy the equation

$$\lambda^2 - \mathrm{Tr}(G)\lambda + 1 = 0$$

and the trace, $\mathrm{Tr}(G)$, is a real number. For $|\mathrm{Tr}(G)| < 2$, the two solutions are conjugates; they are distinct and have modulus one. In this case we call the group element elliptic. For $|\mathrm{Tr}(G)| = 2$, there is a double root 1 or -1. Such group elements (in general Jordan blocks) are called parabolic. For $|\mathrm{Tr}(G)| > 2$, we have two distinct real roots. Such group elements are called hyperbolic.

The Möbius transformation associated to a matrix in $Sl_2(\mathbf{C})$ will have two fixed points on the Riemann sphere unless it is the identity transformation (which fixes all points), or is a Jordan block and so has only one fixed point. For matrices in $SU(1,1)$ or $Sl_2(\mathbf{R})$ the classification into elliptic, parabolic, or hyperbolic points can be understood by the location of these fixed points relative to the domain (disc, half plane).

The elliptic elements have a fixed point in the interior of the domain (disc, half plane). Rotations of the unit disc are easy examples, but any fixed point is possible. Parabolic Möbius transforms have a single fixed point on the boundary of the domain. Horizontal translation of the upper half plane is an example. Hyperbolic elements have two fixed points on the boundary. Multiplication of the upper half plane by a positive scalar is such an example.

Intuitive geometric coordinates on $SU(1,1)$ are

$$(b/|a|, \arg(a)) \in \mathcal{D} \times \mathbf{R}/2\pi\mathbf{Z}$$

Observe that the modulus of $b/|a|$ determines the modulus of a and b via the constraint $|a|^2 - |b|^2 = 1$. The argument of b is equal to the argument of $b/|a|$, and the argument of a is given. Thus the above coordinates indeed determinant a and b. Since $b/|a|$ lies in the open unit disc, the group $SU(1,1)$ can be visualized as a solid open torus or equivalently an infinite cylinder in $\mathbf{C} \times \mathbf{R}$ where the last coordinate is taken modulo 2π. The main axis $b/|a| = 0$ consists of the elements in the compact subgroup of elements

$$\begin{pmatrix} e^{i\phi} & 0 \\ 0 & e^{-i\phi} \end{pmatrix}$$

In $Sl_2(\mathbf{R})$ these elements correspond to the rotations

$$\begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix}$$

Rotating about the main axis by an angle 2ϕ corresponds to multiplying b by a phase factor $e^{2i\phi}$. This can be achieved by conjugating with the previously displayed diagonal element of $SU(1,1)$. Thus rotating the torus about the main axis is an inner automorphism.

Reflecting across the plane determined by requiring b to be real corresponds to replacing b by its complex conjugate. This corresponds to transposing the matrix, which is an anti-automorphism (it changes the order of multiplication). Using the previous rotation symmetry, reflecting across any plane through the main axis is an anti-automorphism.

General Lie theory puts the tangent vectors at the identity element in one-to-one correspondence with one parameter subgroups. We claim that any such subgroup lies in the plane spanned by the tangent vector and the main axis. For the compact subgroup along the main axis this is trivially evident. For any other subgroup, observe that reflecting across this plane gives another one parameter subgroup (these groups are commutative) with the same tangent vector, and thus the reflected subgroup has to be the same as the original subgroup. Thus the subgroup lies inside the plane.

Moreover, we claim that the subgroups are contained in traces of the form

$$\sin(\arg(a))|a|/|b| = const \tag{1.77}$$

Which is to be read projectively.

Indeed it suffices to prove this for those subgroups with real b. Consider two matrices in $SU(1,1)$ with parameters a, b and a', b' and assume b, b' real. If the two elements are in the same subgroup, than the off diagonal element of the product is again real,

$$0 = \mathrm{Im}(ab' + b\overline{a'}) = |a|\sin(\arg(a)b' - |a|'\sin(\arg(a'))b$$

This proves the claim.

If the modulus of the constant in (1.77) is less than one, the entire trace determined by the above equation is contained in the solid torus. The entire trace is a subgroup consisting of elliptic elements. The subgroup is compact. We remark that the special subgroup consisting of the main axis of the infinite cylinder is only special in the chosen coordinates. It is by inner automorphisms equivalent to any of the other elliptic subgroups.

If the constant in (1.77) has modulus equal to one, the solution set of the equation (1.77) meets the boundary of the solid torus. Intersecting with the open torus, we obtain two connected components. The component containing the identity element is a non-compact subgroup, the other component is -1 times this subgroup. All elements in these groups and remainder class are parabolic.

If the constant has modulus larger than one, then the trace of the equation intersected with the torus has again two components. On component is a subgroup consisting of hyperbolic elements, the other component is -1 times the subgroup and also consists of hyperbolic elements. (The two components are mapped onto the same group in $PSU(1,1)$.)

If the constant is infinite, which means $\sin(\arg(a)) = 0$, then the trace consists of two lines, one through the origin is a hyperbolic subgroup, the other one through -1 is -1 times this subgroup.

Note that all automorphisms of the group leave the infinitesimal cone of parabolic elements near the origin invariant. The group acts naturally on its Lie algebra, which is $\mathbf{R}^3$. Since the group leaves a cone invariant, it is easy to see that it acts as $SO(2,1)$. Thus there is a map from $Sl_2(\mathbf{R})$ to $SO(2,1)$. The kernel of this map consists of the identity matrix and minus the identity matrix, therefore there is an embedding of $PSL_2(\mathbf{R})$ into $SO(2,1)$.

Exercise: If K is the compact subgroup of $SU(1,1)$ of diagonal elements, calculate the residue classes $SU(1,1)/K$ and $K\backslash SU(1,1)$. They are spirals in the solid torus representing $SU(1,1)$.

LEMMA 1.55. *If* $G \in SU(1,1)$*, then*

$$\|G\|_{op} = |a| + |b|$$
$$\log \|G\|_{op} = \operatorname{arccosh}(|a|) = \operatorname{arcsinh}(|b|)$$

PROOF. The operator norm of G does not change if we multiply from the left or from the right by an element in the subgroup of diagonal elements. Thus we may assume that a and b are real and positive. Now the matrix is real and symmetric and thus its operator norm is equal to the maximal eigenvalue. However, a basis of eigenvectors is $(1,1)$ and $(1,-1)$ so the eigenvalues are $|a| + |b|$ and $|a| - |b|$. This proves the lemma. $\square$

CHAPTER 2

The Dirac scattering transform

2.1. Introduction

The linear Fourier transform on the integers can be defined as follows. If $F_n \in l^2(\mathbf{Z})$ is a square-summable sequence of complex numbers, then we can (formally, at least) define the linear Fourier transform $\hat{F}(z)$ on the unit circle $\mathbf{T} := \{z \in \mathbf{C} : |z| = 1\}$ by the formula

$$\hat{F}(z) = \sum_{n \in \mathbf{Z}} F_n z^n.$$

Strictly speaking, this summation is absolutely convergent only when F_n is in $l^1(\mathbf{Z})$, but of course the Plancherel theorem guarantees that the map $F \mapsto \hat{F}$ can be continuously extended from $l^1(\mathbf{Z})$ to $l^2(\mathbf{Z})$, and is in fact a unitary transformation from $l^2(\mathbf{Z})$ onto $L^2(\mathbf{T})$ (where we normalize the measure on $\mathbf{T}$ to have total mass one). We also remark that if F is supported on the discrete positive half-line $[0, \infty) := \{n \in \mathbf{Z} : n \geq 0\}$, then $\hat{F}$ is in fact contained in the Hardy space $H^2(\mathcal{D})$ of the unit disk $\mathcal{D} := \{z \in \mathbf{C} : |z| < 1\}$, and conversely. Similarly if F is supported on the discrete negative half-line $(-\infty, -1]$ then $\hat{F}$ is contained in the Hardy space $H_0^2(\mathcal{D}^*)$ of the exterior disk $\mathcal{D}^* := \{z \in \mathbf{C} : |z| > 1\} \cup \{\infty\}$, where $H_0^2(\mathcal{D}^*)$ denotes those elements of $H^2(\mathcal{D}^*)$ whose extension to $\mathcal{D}^*$ vanishes at infinity.

The linear Fourier transform is of course useful for many tasks. We just mention one of them here: if one wishes to solve the linear discrete Schrödinger equation[1]

$$\partial_t F_n(t) = i(F_{n+1}(t) + F_{n-1}(t))$$

then one can easily verify the formula

$$\hat{F}(t, z) = \exp(i(z + z^{-1})t)\hat{F}(0, z). \tag{2.1}$$

Thus, if one knows how to compute Fourier transforms and inverse Fourier transforms, one can solve the Cauchy problem for the linear discrete Schrödinger equation (or indeed for any linear discrete equation) explicitly, say for initial data $F_n(0)$ in $l^2(\mathbf{Z})$.

The purpose of this paper is to generalize the above (very well-known) results to the *(Dirac) non-linear Fourier transform* (NLFT) on the integers. The NLFT will be defined on potentials $F_n \in l^2(\mathbf{Z})$ which obey the additional constraint that $|F_n| < 1$ for all n; such potentials will be called *admissible*. Formally, the NLFT $\overbrace{F}(z)$ of such a potential is defined for $z \in \mathbf{T}$ by the formula

$$\overbrace{F}(z) = \prod_{n \in \mathbf{Z}} \frac{1}{\sqrt{1 - |F_n|^2}} \begin{pmatrix} 1 & \overline{F_n} z^{-n} \\ F_n z^n & 1 \end{pmatrix}, \tag{2.2}$$

where the (non-commutative) product is interpreted from left-to-right, thus formally we have

$$\prod_{n \in \mathbf{Z}} M_n = \ldots M_{-2} M_{-1} M_0 M_1 M_2 \ldots.$$

This infinite product is only absolutely convergent when F is in $l^1(\mathbf{Z})$, but we will show (in analogy with the linear situation) that we may extend the non-linear

---

[1] It is traditional to also place a factor of $-2iF_n(t)$ on the right-hand side, but we have elected not to do so here (or later on in (2.7)) in order to simplify our formulae slightly. In any event this factor of $-2iF_n(t)$ can be restored simply by multiplying $F_n(t)$ by e^{-2it}.

Fourier transform[2] from $l^1(\mathbf{Z})$ to $l^2(\mathbf{Z})$ by developing a non-linear version of the Plancherel theorem.

Thus the NLFT $\overbrace{F}(z)$ is a multiplicative analogue of the linear Fourier transform $\hat{F}(z)$, but takes values in the space of 2×2 complex matrices instead of the complex numbers; the factor $\frac{1}{\sqrt{1-|F_n|^2}}$ ensures that this matrix has determinant 1. In fact it must take the form

$$\overbrace{F}(z) = \begin{pmatrix} a(z) & b(z) \\ \overline{b(z)} & \overline{a(z)} \end{pmatrix},$$

where $|a(z)|^2 - |b(z)|^2 = 1$; the functions $a(z)$ and $b(z)$ (or more precisely $1/a(z)$ and $b(z)/a(z)$) are sometimes known as *transmission* and *reflection* coefficients. In fact, as we shall see, these functions are closely connected with the scattering transform of a *discrete Dirac operator* $L = L[F]$, defined as follows. Let $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ be the space of pairs $\begin{pmatrix} \alpha(\zeta) \\ \beta(\zeta) \end{pmatrix}$ of square-integrable functions on the unit circle[3] $\mathbf{T}$, with inner product

$$\langle \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \alpha_2 \\ \beta_2 \end{pmatrix} \rangle_{L^2(\mathbf{T}) \oplus L^2(\mathbf{T})} := \int_{\mathbf{T}} \alpha_1 \overline{\alpha_2} + \beta_1 \overline{\beta_2}.$$

We endow $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ with the orthonormal Fourier basis

$$v_n := \begin{pmatrix} \zeta^n \\ 0 \end{pmatrix}; \quad w_n := \begin{pmatrix} 0 \\ \zeta^{-n} \end{pmatrix} \tag{2.3}$$

and then define the linear operator $L : L^2(\mathbf{T}) \oplus L^2(\mathbf{T}) \to L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ on the Fourier basis vectors by the formulae

$$\begin{aligned} Lv_n &:= \sqrt{1 - |F_n|^2} v_{n+1} + F_n w_n \\ Lw_{n+1} &:= -F_n^* v_{n+1} + \sqrt{1 - |F_n|^2} w_n. \end{aligned} \tag{2.4}$$

It is easy to check that L is well-defined and in fact extends to a unitary operator on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, with inverse $L^{-1} = L^*$ given by

$$\begin{aligned} L^{-1} v_{n+1} &:= \sqrt{1 - |F_n|^2} v_n - F_n w_{n+1} \\ L^{-1} w_n &:= F_n^* v_n + \sqrt{1 - |F_n|^2} w_{n+1}. \end{aligned} \tag{2.5}$$

In particular, the spectrum of L resides entirely on the unit circle. This fact will ensure, for instance, that L has no bound states away from the unit circle, which makes the analysis of these operators somewhat simpler than similar operators such as Jacobi matrices. However, when F is slowly decaying (e.g. if it is admissible but no better) it is still possible for L to have embedded bound states or singular continuous spectrum on the circle; we shall see that this will cause some difficulties with analyzing the NLFT.

---

[2]We do not however resolve the very interesting question of whether this infinite product converges in a pointwise sense for almost every z when F is in $l^2(\mathbf{Z})$. This would be a non-linear version of a famous theorem of Carleson [Terry: Supply reference! Maybe also discuss NL Walsh case, Christ-Kiselev].

[3]We parameterize the circle here by ζ to distinguish it from the complex parameter z. The relation will be given by $z = \zeta^{\pm 2}$.

The NLFT shares many features in common with the linear Fourier transform, except of course that it is non-linear. When the potential F is very small, we in fact have the *Born approximation*

$$\widehat{F}(z) \approx \begin{pmatrix} 1 + \frac{1}{2}P_{[0,+\infty)}|\hat{F}|^2(z) & \hat{F}(z) \\ \overline{\hat{F}(z)} & 1 + \frac{1}{2}P_{(-\infty,0]}|\hat{F}|^2(z) \end{pmatrix}, \tag{2.6}$$

which is easily obtained by discarding all terms of cubic and higher order in F. Here $P_{[0,+\infty)}$ is the Riesz projection that maps $\sum_{n\in\mathbf{Z}} c_n z^n$ to $\sum_{n\in[0,+\infty)} c_n z^n$ (i.e. the orthogonal projection from $L^2(\mathbf{T})$ to $H^2(\mathcal{D})$, and similarly for $P_{(-\infty,0]}$. Later on we shall see other common features, for instance the non-linear Fourier transform enjoys many of the same symmetries as the linear Fourier transform and also has a Plancherel identity. Moreover, just as the linear Fourier transform can be used to solve the discrete linear Schrödinger equation, the non-linear Fourier transform can be used to solve[4] the discrete non-linear Schrödinger equation (or *Ablowitz-Ladik* equation)

$$\partial_t F_n = i(1 - |F_n|^2)(F_{n-1} + F_{n+1}). \tag{2.7}$$

The factor $(1-|F_n|^2)$ will ensure that the property of being admissible (in particular, that $|F_n| < 1$ for all n) is preserved by the flow (2.7). As is well known, this equation is completely integrable, and in fact can be placed in Lax pair formulation with the operator L defined above in (2.4); we review this fact in ???. As a consequence, we can show (in analogy with (2.1), see also (2.6)) that

$$b(t,z) = \exp(i(z + z^{-1})t)b(0,z); \quad a(t,z) = a(0,z) \tag{2.8}$$

whenever F solves the Ablowitz-Ladik equation (2.7). Initially we will only be able to derive this for sufficiently fast decaying potentials F, but later on we will use a continuity argument to extend this fact to all admissible solutions to (2.7).

In view of (2.8), one sees that one could solve (2.7) explicitly (for $l^2(\mathbf{Z})$ potentials, for instance) if one knew how to compute the NLFT and its inverse for $l^2(\mathbf{Z})$. The forward NLFT is straightforward, being given more or less explicitly by (2.2), but the inverse NLFT is more difficult. When the potential F has sufficient decay (e.g. is integrable), then the non-linear Fourier transform is a bounded function of z, and one can use the inverse scattering methods of Gelfand-Levitan (based on inverting Hankel operators) and Marcenko (based on solving a discrete integral equation, basically a linear Fourier transform of the Gelfand-Levitan approach) to invert the NLFT uniquely; for ease of reference we reproduce that result here. In the special case when F is supported on a half-line one can also use a "layer stripping" (or Schur algorithm) approach to recover the potential; this method works even when F is admissible and no better. [Terry: insert references here!]

However when F is supported on the full line, and is admissible and no better, then inversion becomes more difficult. In fact we show that it is possible to have several potentials with the same NLFT. This phenomenon is partly due to the presence of singular spectrum of the Dirac operator L (a fact which is only partially detected by the scattering data $a(z)$, $b(z)$), however even if one restricts to those class of potentials F for which L has purely absolutely continuous spectrum, there is still breakdown of uniqueness, in that it is still possible to find two distinct

[4]In fact, the NLFT can solve a number of discrete evolution equations, including a discrete version of the modified KdV equation; see [Terry: insert Ablowitz-Ladik reference here].

admissible potentials with the same NLFT. For Jacobi matrices, this phenomenon was first observed by Yuditskii and Volberg [**34**].

Nevertheless, we are still able to obtain some results regarding the inverse NLFT. First, by modifying the Hilbert space method introduced by Yuditskii and Volberg, we are able to show that every set of scattering data $a(z)$ and $b(z)$ which is "admissible" (specifically, that $a(z)$ is an outer function on $\mathcal{D}$, is positive real-valued at the origin and $|a(z)|^2 - |b(z)|^2 = 1$ almost everywhere on $\mathbf{T}$) is the NLFT of at least one admissible potential on the line. In fact we supply two methods for giving such a potential, and the two potentials thus produced coincide if and only if the solution to the inverse NLFT is unique. Indeed, these two potentials form the two extreme solutions to the inverse NLFT; we will show in a very specific sense that all other solutions lie "in between" these two extremes.

[Emphasize that this work is mostly self-contained despite touching on a number of different fields.]

[Thank Camil, John Garnett, Barry Simon, anyone else?]

2.2. Functions on the circle, disk, and exterior disk

In this section we set out some standard notation for complex functions on the disk and recall some known facts concerning these functions. A standard reference for these facts is [**16**].

We will be considering various complex-valued or matrix-valued measurable functions from the circle $\mathbf{T} := \{z \in \mathbf{C} : |z| = 1\}$, endowed with normalized circle measure $\frac{|dz|}{2\pi}$, thus

$$\int_{\mathbf{T}} f := \frac{1}{2\pi} \int_0^{2\pi} f(e^{i\theta})\ d\theta.$$

Many of these functions will have holomorphic or meromorphic extensions to the disk $\mathcal{D} := \{z \in \mathbf{C} : |z| < 1\}$ or the exterior disk $\mathcal{D}^* := \{z \in \mathbf{C} : |z| > 1\} \cup \{\infty\}$. We abuse notation and use z to denote the identity function $z \mapsto z$, z^{-1} to denote its reciprocal $z \mapsto z^{-1}$, etc. Note from the Plancherel theorem that the functions $\{z^n\}_{n\in\mathbf{Z}}$ form an orthonormal basis of $L^2(\mathbf{T})$. We define a *(scalar) Laurent polynomial* to be a finite linear combination $\hat{c}$ of integer powers of z, thus $\hat{c} = \sum_{n\in\mathbf{Z}} c_n z^n$ for some compactly supported sequence c_n; such functions extend of course to the punctured plane $\mathbf{C} - \{0\}$, and extend also to 0 if c is supported on $[0, +\infty)$ and to ∞ if c is supported on $(-\infty, 0]$.

We also shall need other circles $(1 \pm \varepsilon)\mathbf{T} := \{z \in \mathbf{C} : |z| = 1 \pm \varepsilon\}$ close to the unit circle $\mathbf{T}$. We always give these circles normalized arclength measure, so that the total mass of these circles is always 1. Observe from Cauchy's theorem that $\int_{r\mathbf{T}} f = \int_{r'\mathbf{T}} f$ if f is holomorphic on an open neighborhood of the annulus bordered by $r\mathbf{T}$ and $r'\mathbf{T}$.

We will identify a function f on the circle $\mathbf{T}$ with its holomorphic extensions (if they exist) to $\mathcal{D}$ or $\mathcal{D}^*$ (possibly with singularities at 0 or ∞), where these extensions converge non-tangentially a.e. back to f on $\mathbf{T}$. By the theorem of [??Riesz??] we know that such extensions, if they exist, are unique. We also adopt the usual convention of identifying two functions on $\mathbf{T}$ if they agree almost everywhere.

Given any (possibly infinite) interval $[p, q] := \{n \in \mathbf{Z} : p \le n \le q\}$ of the integers, we let $H^2_{[p,q]}$ be the closed subspace of $L^2(\mathbf{T})$ generated by the orthonormal basis $\{z^n : n \in [p, q]\}$; note that this is the image of $l^2([p, q])$ under the linear

Fourier transform, and $L^2(\mathbf{T}) = H^2_{(-\infty,\infty)}$. In particular we can define the usual Hardy spaces

$$H^2(\mathcal{D}) := H^2_{[0,+\infty)}; H^2(\mathcal{D}^*) := H^2_{(-\infty,0]}$$

together with their mean zero variants

$$H^2_0(\mathcal{D}) := H^2_{[1,+\infty)}; H^2_0(\mathcal{D}^*) := H^2_{(-\infty,-1]}.$$

As is well known $H^2(\mathcal{D})$ consists of boundary values of holomorphic functions in $\mathcal{D}$ which are uniformly in L^2 on circles slightly smaller than the unit circle, while $H^2_0(\mathcal{D})$ is the codimension one subspace of $H^2(\mathcal{D})$ consisting of functions which vanish at the origin; there are similar properties for $H^2(\mathcal{D}^*)$ and $H^2_0(\mathcal{D}^*)$. Also observe that $H^2_0(\mathcal{D})$ (resp. $H^2_0(\mathcal{D}^*)$) is the orthogonal complement of $H^2(\mathcal{D}^*)$ (resp. $H^2(\mathcal{D})$) in $L^2(\mathbf{T})$.

If f is a complex-valued function on $\mathbf{T}$, $\mathcal{D}$, or $\mathcal{D}^*$, we define the *conjugate* f^* to be the function

$$f^*(z) := \overline{f(\overline{z}^{-1})};$$

thus if f is defined on $\mathbf{T}$, $\mathcal{D}$, or $\mathcal{D}^*$, then f^* is defined on $\mathbf{T}$, $\mathcal{D}^*$, or $\mathcal{D}$ respectively; in particular, $f \in H^2(\mathcal{D})$ if and only if $f^* \in H^2(\mathcal{D}^*)$, etc. We note that the conjugation operation preserves holomorphicity, and is a skew-linear involution. On the circle $\mathbf{T}$ we of course have $f^*(z) = \overline{f(z)}$. We observe that the conjugate of identity function z is z^{-1}.

[Define Nevanlinna, Smirnov, outer]

A *Herglotz function on* $\mathcal{D}^*$ is any holomorphic function $f : \mathcal{D}^* \to \{z \in \mathbf{C} : \mathrm{Re}(z) \geq 0\}$ from the unit disk to the right half-plane, normalized so that $f(\infty) = 1$. The Herglotz representation theorem (see e.g. [**16**]) shows that one can associate to each Herglotz function f a unique probability measure μ on $\mathbf{T}$ such that

$$f(z) = \int_{\mathbf{T}} \frac{z + e^{i\theta}}{z - e^{i\theta}} d\mu(e^{i\theta})$$

for all $z \in \mathcal{D}^*$. Furthermore, f lies in the Hardy space $H^p(\mathcal{D}^*)$ for all $0 < p < 1$ (but not necessarily at $p = 1$!) and thus has non-tangential limits a.e. on $\mathbf{T}$; indeed, we have $\mathrm{Re} f(z) = \frac{d\mu_{ac}}{|dz|/2\pi}(z)$ for almost every $z \in \mathbf{T}$, where μ_{ac} is the absolutely continuous component of μ and $\frac{|dz|}{2\pi}$ is normalized measure on $\mathbf{T}$. Also, there exists a real-valued distribution on $\mathbf{T}$, which we shall call $H\mu$ (H denoting the Hilbert transform), such that $\mu + iH\mu$ is the weak limit of f; in other words

$$\lim_{\varepsilon \to 0} \int_{(1+\varepsilon)\mathbf{T}} c(z) f(z) = \int_{\mathbf{T}} c(z)\ d\mu + i \int_{\mathbf{T}} c(z) H\mu$$

for any smooth function $c(z)$ on a neighbourhood of $\mathbf{T}$. In particular if c vanishes on $\mathbf{T}$ then the left-hand side is zero, while if c is real-valued on $\mathbf{T}$ then

$$\lim_{\varepsilon \to 0} \mathrm{Re} \int_{(1+\varepsilon)\mathbf{T}} c(z) f(z) = \int_{\mathbf{T}} c(z)\ d\mu.$$

In particular we see that the defect

$$\mathrm{Re} \int_{\mathbf{T}} c(z) f(z) - \lim_{\varepsilon \to 0} \mathrm{Re} \int_{(1+\varepsilon)\mathbf{T}} c(z) f(z)$$

in the convergence of $\mathrm{Re} \int_{(1+\varepsilon)\mathbf{T}} c(z) f(z)$ is equal to $\int_{\mathbf{T}} c(z)\ d(\mu_{sc} + \mu_{pp})$, where $\mu_{sc} + \mu_{pp}$ is the singular component of μ. Thus we can use the defect of integrals on circles to detect the presence of singular measure.

2.3. Matrix-valued functions on the disk

For any two complex functions $a(z), b(z)$ defined a.e. on $\mathbf{T}$, define the 2×2 matrix $M[a, b]$ by

$$M[a,b] = \begin{pmatrix} a & b^* \\ b & a^* \end{pmatrix};$$

this is then a matrix-valued function defined almost everywhere on $\mathbf{T}$. The space of all such matrices is clearly a real vector space. It is also closed under multiplication:

$$M[a_-, b_-]M[a_+, b_+] = M[a_-a_+ + b_-^* b_+, a_-^* b_+ + b_- a_+]. \tag{2.9}$$

We say that $M[a, b]$ is $SU(1, 1)$*-valued* if the determinant is identically 1 almost everywhere on $\mathbf{T}$, i.e. $aa^* - bb^* = |a|^2 - |b|^2 = 1$ on $\mathbf{T}$. The space of $SU(1, 1)$-valued matrix functions is a group with identity $id = M[1, 0]$ and inverse

$$M[a,b]^{-1} = M[a^*, -b]. \tag{2.10}$$

Observe that in the event that $M[a, b]$ can be holomorphically extended to $\mathcal{D}$ and $\mathcal{D}^*$, then the identity $aa^* - bb^* = 1$ must then still hold (by uniqueness of extensions), but the identity $|a|^2 - |b|^2 = 1$ need not. We can a $SU(1, 1)$*-valued Laurent polynomial* any $SU(1, 1)$-valued matrix function whose coefficients are all Laurent polynomials.

For any $SU(1, 1)$-valued $M[a, b]$, we define the *reflection coefficients* $r = r[a, b]$, $s = s[a, b]$ by

$$r := \frac{b}{a}; \quad s := \frac{b^*}{a}.$$

Thus we have the pointwise estimate $|r| = |s| < 1$ almost everywhere on $\mathbf{T}$.

Observe that the adjoint of $M[a, b]$ is given by

$$M[a, b]^* = M[a^*, b],$$

in particular the space of $SU(1, 1)$-valued matrices is closed under adjoint. The space is similarly closed under transpose: $M[a, b]^t = M[a, b^*]$.

We now define matrix-valued non-linear analogues of the spaces $L^2(\mathbf{T})$, $H^2_{[p,q]}$, $H^2(\mathcal{D})$, $H^2(\mathcal{D}^*)$, $H^2_0(\mathcal{D})$, and $H^2_0(\mathcal{D}^*)$. Define $\mathcal{L}^2(\mathbf{T})$ to be the space of $SU(1, 1)$-valued functions $M[a, b]$ on the torus such that a extends to an outer function on $\mathcal{D}$ with $a(0)$ real and positive. Observe that since a is outer, the function $1/a$ is also outer; in particular a has no zeroes in $\mathcal{D}$. Since $|a|^2 = 1 + |b|^2 \geq 1$ on $\mathbf{T}$, we thus see from the maximum principle that $|1/a(z)| \leq 1$ for all $z \in \mathcal{D}$. In particular we have $1 \leq a(0) < +\infty$; we shall refer to $a(0)$ as the *energy* of $M[a, b]$, and denote it as $E(M[a, b]) := a(0)$. Since a is outer and has magnitude greater than or equal to 1, we see in particular that a is log-integrable in $\mathbf{T}$ with

$$0 \leq \int_{\mathbf{T}} \log |a| = \log a(\infty) < +\infty.$$

Also we see that the phase $\arg(a)$ of a is the imaginary part of $\log a$ and thus can be recovered from $\log |a|$ via the Hilbert transform. In particular, the phase of a can be determined explicitly from the magnitude. We refer to elements $M[a, b]$ in $\mathcal{L}^2(\mathbf{T})$ as *scattering data*; later we shall see that these elements are precisely the non-linear Fourier transform (or scattering transform) of admissible potentials.

If $M[a, b] \in \mathcal{L}^2(\mathbf{T})$ is a scattering datum, then the reflection coefficients $r = b/a$, $s = b^*/a$ are bounded functions on $\mathbf{T}$, and in particular lie in $L^2(\mathbf{T})$. We can then endow $\mathcal{L}^2(\mathbf{T})$ with a somewhat artificial metric, writing

$$d(M[a,b], M[a',b']) := \|\log|a| - \log|a'|\|_{L^1(\mathbf{T})} + \|r - r'\|_{L^2(\mathbf{T})} + \|s - s'\|_{L^2(\mathbf{T})}$$

where $r = b/a$, $s = b^*/a$ and similarly for r', s'. This is clearly a metric on $\mathcal{L}^2(\mathbf{T})$ (recall that if $|a| = |a'|$, then a and a' must have the same phase and are thus equal), and can easily be seen to turn $\mathcal{L}^2(\mathbf{T})$ into a complete metric space. Later on, we shall prove that the NLFT is a continuous surjection from $L^2(\mathbf{Z}; \mathcal{D})$ onto $\mathcal{L}^2(\mathbf{T})$, although it is not quite a bijection or a homeomorphism because there is failure of injectivity.

For any finite interval $[p,q]$ in the integers, we define the space $\mathcal{H}^2_{[p,q]}$ to be the space of all scattering data $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ such that $a \in H^2_{[0,q-p]}$ and $b \in H^2_{[p,q]}$. This definition has to be modified for the half-infinite case, as a and b will no longer lie in $L^2(\mathbf{T})$; instead we will use reflection coefficients. More precisely, we define $\mathcal{H}^2_{[p,+\infty)}$ to be the space of scattering data $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ where $r \in H^2_{[p,+\infty)}$, and $\mathcal{H}^2_{(-\infty,q]}$ to be the space of scattering data $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ where $s^* \in H^2_{(-\infty,q]}$. We warn the reader that while $\mathcal{H}^2_{[p,+\infty)} \cap \mathcal{H}^2_{(-\infty,q]}$ contains $\mathcal{H}^2_{[p,q]}$, the two spaces are not equal; this is in fact a major contributor to the failure of uniqueness of the NLFT on the integers.

We then define

$$\mathcal{H}^2(\mathcal{D}) := \mathcal{H}^2_{[0,+\infty)}; \quad \mathcal{H}^2(\mathcal{D}^*) := \mathcal{H}^2_{(-\infty,0]}; \quad \mathcal{H}^2_0(\mathcal{D}) := \mathcal{H}^2_{[1,+\infty)}; \quad \mathcal{H}^2_0(\mathcal{D}^*) := \mathcal{H}^2_{(-\infty,-1]}.$$

The spaces $\mathcal{H}^2_{[p,q]}$ can be easily verified to be closed subspaces of $\mathcal{L}^2(\mathbf{T})$ (mainly because their linear analogues $H^2_{[p,q]}$ are closed subspaces of $L^2(\mathbf{T})$). As we shall see, the NLFT will be a homeomorphism from $l^2([p,q])$ to $\mathcal{H}^2_{[p,q]}$ for any finite or half-infinite interval $[p,q]$, but not quite for the full interval $(-\infty, \infty)$.

2.4. The non-linear Fourier transform for compactly supported potentials

Define an *admissible potential* to be any sequence $F = (F_n)_{n\in\mathbf{Z}}^{\infty}$ in $l^2(\mathbf{Z})$ such that $|F_n| < 1$ for all n. We call $l^2(\mathbf{Z}; \mathcal{D})$ the space of all such potentials, which we endow with the topology induced by $l^2(\mathbf{Z})$. If $F \in l^2(\mathbf{Z}; \mathcal{D})$ is an an admissible potential, we define the *energy* $E(F)$ to be the quantity

$$E(F) := \prod_{n\in\mathbf{Z}} \frac{1}{\sqrt{1 - |F_n|^2}};$$

observe that this product is absolutely convergent and $1 \leq E(F) < +\infty$.

We also define the *transfer matrices*[5] $M_{n\leftarrow n+1}$ by

$$M_{n\leftarrow n+1} = \frac{1}{\sqrt{1 - |F_n|^2}} M[1, F_n z^n]; \tag{2.11}$$

these are $SU(1,1)$-valued Laurent polynomials on $\mathbf{T}$, which then of course extend to the punctured plane $\mathbf{C} - \{0\}$. More generally, for any integers $n' > n$, we define

$$M_{n\leftarrow n'} := \prod_{n\leq m<n'} M_{m\leftarrow m+1} = M_{n\leftarrow n+1} M_{n+1\leftarrow n+2} \dots M_{n'-1\leftarrow n'}; \tag{2.12}$$

[5]The reason for the arrow pointing left here is because of the convention that operators should lie to the left of their operands; this will prevent the groupoid law (2.13) from being reversed.

thus all the transfer matrices $M_{n \to n'}$ are $SU(1,1)$-valued Laurent polynomials. We observe the groupoid law

$$M_{n \leftarrow n'} M_{n' \leftarrow n''} = M_{n \leftarrow n''} \tag{2.13}$$

whenever $n < n' < n''$.

If F is compactly supported, then the transfer matrices $M_{-N \to N}$ eventually become constant as $N \to +\infty$, and so we can define $M_{-\infty \leftarrow +\infty}$ to equal this constant value; we refer to this as the *non-linear Fourier transform* $\overbrace{F}$ of F. In other words,

$$\overbrace{F} := M_{-\infty \leftarrow +\infty} := \prod_{-\infty < n < +\infty} M_{n \leftarrow n+1} = \prod_{-\infty < n < +\infty} \frac{1}{\sqrt{1 - |F_n|^2}} M[1, F_n z^n]. \tag{2.14}$$

Again, for compactly supported potentials we see that $\overbrace{F}$ is a $SU(1,1)$-valued Laurent polynomial.

Thus for instance, the non-linear Fourier transform of the zero potential $F_n \equiv 0$ is the identity function $\overbrace{0} = M[1,0]$, while the non-linear Fourier transform of a Dirac mass $F_n \delta_n$ (where δ_n is the Kronecker delta at n) is

$$\overbrace{F_n \delta_n} = \frac{1}{\sqrt{1 - |F_n|^2}} M[1, F_n z^n]. \tag{2.15}$$

From the groupoid law we observe the multiplicativity property

$$\overbrace{F_1 + F_2} = \overbrace{F_1}\, \overbrace{F_2} \tag{2.16}$$

whenever F_1 is supported to the left of F_2 (i.e. $F_1(n) F_2(n')$ is non-zero only when $n < n'$); this is a weak analogue of the additivity property $\widehat{F_1 + F_2} = \widehat{F_1} + \widehat{F_2}$ for the linear Fourier transform. Note in fact that the two properties (2.15), (2.16) could be used to define $\overbrace{F}$ for all compactly supported potentials.

We now record some useful symmetries of the non-linear Fourier transform, which are analogues of the corresponding symmetries for the linear Fourier transform.

LEMMA 2.1. *Let F be an admissible potential with compact support.*

- *(Phase rotation symmetry) If $e^{i\theta} \in \mathbf{T}$ and F' is the potential $F'_n := e^{i\theta} F_n$, then*
$$\overbrace{F'} = M[e^{-i\theta/2}, 0]\, \overbrace{F}\, M[e^{i\theta/2}, 0],$$
or in other words $b'(z) = e^{i\theta} b(z)$ and $a'(z) = a(z)$.
- *(Translation symmetry) If $k \in \mathbf{Z}$ and F' is the potential $F'_n := F_{n-k}$, then*
$$\overbrace{F'} = M[z^{-k/2}, 0]\, \overbrace{F}\, M[z^{k/2}, 0]$$
or in other words $b'(z) = z^k b(z)$ and $a'(z) = a(z)$.
- *(Modulation symmetry) If $e^{i\alpha} \in \mathbf{T}$ and F' is the potential $F'_n := e^{in\alpha} F_n$, then*
$$\overbrace{F'}(z) = \overbrace{F}(e^{i\alpha} z)$$
or in other words $b'(z) = b(e^{i\alpha} z)$ and $a'(z) = a(e^{i\alpha} z)$.

- *(Reflection symmetry) If F' is the potential $F'_n = F^*_{-n}$, then*
$$\overbrace{F'}(z) = \overbrace{F}(z)^t$$
or in other words $b'(z) = b(z)^*$ and $a'(z) = a(z)$.

The reader should verify that the above symmetries are consistent with the Born approximation (2.6) and the corresponding symmetries for the linear Fourier transform. Although we are currently only verifying these symmetries for compactly supported potentials, they will easily extend to all admissible potentials by a limiting argument once we show the continuity of the NLFT in the next section.

PROOF. The phase rotation symmetry follows from the observation that

$$M'_{n\leftarrow n+1} = M[e^{-i\theta/2}, 0]M_{n\leftarrow n+1}M[e^{i\theta/2}, 0]$$

for all $n \in \mathbf{Z}$. The translation symmetry is similar. The modulation symmetry follows by direct expansion of both sides, and the reflection symmetry can be obtained by observing that $M'_{n\leftarrow n+1} = M^t_{-n\leftarrow -n+1}$. □

We now give the nonlinear Plancherel theorem for compactly supported potentials.

PROPOSITION 2.2. *Let $[p, q]$ be any finite interval in the integers. Then the map $F \mapsto \overbrace{F}$ is a homeomorphism from $l^2([p, q]; \mathcal{D})$ to $\mathcal{H}^2_{[p,q]}$, where $\mathcal{H}^2_{[p,q]}$ was defined in the previous section. Furthermore for every $F \in l^2([p, q])$ we have the Plancherel identity*

$$E(\overbrace{F}) = E(F)$$

or in other words

$$\int_{\mathbf{T}} \log|a| = \log a(0) = \prod_{n\in\mathbf{Z}} \frac{1}{\sqrt{1-|F_n|^2}}. \tag{2.17}$$

This should be compared with the situation for the linear Fourier transform, which maps $l^2([p, q])$ unitarily to $H^2_{[p,q]}$; observe also that (2.17) is consistent with the Born approximation and with the linear Plancherel identity

$$\int_{\mathbf{T}} |\hat{F}|^2 = \sum_{n\in\mathbf{Z}} |F_n|^2.$$

One can also view (2.17) as a trace identity for the unitary operator L defined earlier, but we will not pursue this viewpoint.

PROOF. We induct on the cardinality of the interval $[p, q]$. When $[p, q]$ is empty the claims are vacuously true, and when $p = q$ the claims can be verified by direct inspection. Now suppose inductively that $[p, q]$ has cardinality greater than 1, and the claim has already been proven for all smaller intervals. The key is the following lemma.

LEMMA 2.3. *Let $[p_1, q_1]$, $[p_2, q_2]$ be two finite intervals such that $q_1 < p_2$ (i.e. $[p_1, q_1]$ lies to the left of $[p_2, q_2]$. Then for any $M[a_1, b_1] \in \mathcal{H}^2_{[p_1,q_1]}$ and $M[a_2, b_2] \in \mathcal{H}^2_{[p_2,q_2]}$, we have*

$$M[a_1, b_1]M[a_2, b_2] \in \mathcal{H}^2_{[p_1,q_2]}$$

and

$$E(M[a_1,b_1]M[a_2,b_2]) = E(M[a_1,b_1])E(M[a_2,b_2]). \tag{2.18}$$

PROOF. Using translation invariance we may take $p_2 = 0$. Write $M[a_1,b_1]M[a_2,b_2] = M[a,b]$. Since the two factors on the left are $SU(1,1)$-valued, so is the right factor, thus $|a|^2 - |b|^2 = 1$. Also by (2.9) we have

$$a = a_1a_2 + b_1^* b_2 = a_1a_2(1 + r_2 s_1)$$

where $r_2 = b_2/a_2$, $s_1 = b_1^*/a_1$. Since $M[a_2,b_2]$ lies in $\mathcal{H}^2_{[0,q_2]}$, we see that r_2 is holomorphic on $\mathcal{D}$ and has magnitude strictly less than 1 on this disk. Similarly s_1 is also holomorphic on $\mathcal{D}$, is strictly less than 1, and vanishes at the origin. Thus $1 + r_2 s_1$ is holomorphic on the disk, bounded and bounded away from zero, and equals one at the origin. Since a_1, a_2 are outer, this implies that a is also outer, and that $a(0) = a_1(0)a_2(0)$, which is (2.18). Finally, since a_j is in $H^2_{[0,q_j-p_j]}$ and b_j is in $H^2_{[p_j,q_j]}$ for $j = 1,2$, one can easily see from (2.9) that $a \in H^2_{[0,q_2-p_1]}$ and $b_j \in H^2_{[p_1,q_2]}$, and thus $M[a,b] \in \mathcal{H}^2_{[p_1,q_2]}$ as desired. $\square$

From this lemma, (2.16) and induction we thus see that the NLFT maps $l^2([p,q])$ to $\mathcal{H}^2([p,q])$; this map is also clearly continuous from (2.14). Now we prove that this map is injective on $l^2([p,q])$. By translation invariance we may take $p = 0$. Let F be any element of $l^2([0,q])$; we write $F = F_0\delta_0 + F'$ where δ_0 is the Kronecker delta at 0, and $F' \in l^2([1,q])$. By (2.16) we have

$$\overbrace{F} = \frac{1}{\sqrt{1-|F_0|^2}} M[1,F_0]\overbrace{F'}. \tag{2.19}$$

If we write $\overbrace{F} = M[a,b]$ and $\overbrace{F'} = M[a',b']$, we thus have

$$a = \frac{1}{\sqrt{1-|F_0|^2}}(a' + F_0^* b'); \quad b = \frac{1}{\sqrt{1-|F_0|^2}}(F_0 a' + b').$$

Since $a' \in H^2_{[0,q-1]}$ and $b' \in H^2_{[1,q]}$, we thus see that b extends holomorphically to $\mathcal{D}$ and

$$a(0) = \frac{1}{\sqrt{1-|F_0|^2}} a'(0); \quad b(0) = \frac{F_0}{\sqrt{1-|F_0|^2}} a'(0).$$

In particular we have

$$F_0 = \frac{b(0)}{a(0)} = r(0). \tag{2.20}$$

In particular we can reconstruct F_0 from $\overbrace{F} = M[a,b]$, and hence by (2.19) we can recover $\overbrace{F'}$ from $\overbrace{F}$. By the inductive hypothesis the NLFT is already injective on $l^2([1,q])$, and is hence also injective on $l^2([0,q])$. [Christoph: should probably mention layer stripping and/or the Szego algorithm here.]

Now we show the NLFT is also surjective from $l^2([p,q];\mathcal{D})$ to $\mathcal{H}^2([p,q])$, by reversing the above injectivity argument. Again, we may take $p = 0$. Let $M[a,b]$ be any element of $\mathcal{H}^2([0,q])$. We define the complex number F_0 by (2.20); note that

$$F_0 = \frac{b(0)}{a(0)} = \int_{\mathbf{T}} \frac{b}{a}$$

and in particular $|F_0| < 1$ (since $|a|^2 = 1 + |b|^2$, and hence $b/a| < 1$, on $\mathbf{T}$).

We then define a', b' by the analogue of (2.19), namely

$$M[a,b] =: \frac{1}{\sqrt{1-|F_0|^2}} M[1, F_0] M[a', b'],$$

thus

$$M[a', b'] := \frac{1}{\sqrt{1-|F_0|^2}} M[1, -F_0] M[a, b] \tag{2.21}$$

and hence by (2.9)

$$a' = \frac{1}{\sqrt{1-|F_0|^2}}(a - F_0^* b); \quad b' = \frac{1}{\sqrt{1-|F_0|^2}}(-F_0 a + b). \tag{2.22}$$

We know that $a \in H^2_{[0,q]}$ and $b \in H^2_{[0,q]}$; by (2.20) we thus see that $b' \in H^2_{[1,q]}$. At first glance it seems that we can only place a' in $H^2_{[0,q]}$, but (2.20) and an inspection of the z^q coefficient of the identity $aa^* - bb^* = 1$ shows that $a - F_0^* b$ has no z^q coefficient and thus $a' \in H^2_{[0,q-1]}$. Also we see that $M[a', b']$ is $SU(1,1)$-valued and hence $|a'|^2 - |b'|^2 = 1$ on $\mathbf{T}$. Evaluationg a' at zero and using (2.20) we see that

$$a'(0) = \sqrt{1-|F_0|^2} a(0)$$

and in particular $a'(0)$ is positive and real. Finally from the identity

$$a - F_0^* b = a(1 - F_0^* r)$$

and noting that r is bounded by 1 on $\mathcal{D}$ we see that $1 - F_0^* r$ is bounded and is bounded away from zero, and thus a' is outer. Thus $M[a', b'] \in \mathcal{H}^2_{[1,q]}$, and by inductive hypothesis arises as the non-linear Fourier transform of a potential F' in $l^2([1,q])$. By setting $F = F_0\delta_0 + F$ and using (2.16) we obtain $M[a,b] = \widehat{F}$, which proves surjectivity.

Note this argument also shows that the inverse of the NLFT from $\mathcal{H}^2([p,q])$ to $l^2([p,q]; D)$ is continuous; note that both spaces are finite-dimensional so the exact nature of the topology on $\mathcal{H}^2([p,q])$ is not a concern. This completes the proof of the Proposition. $\square$

2.5. The non-linear Fourier transform on half-line potentials

We now extend the non-linear Fourier transform, defined in the previous section for compactly supported potentials, to admissible potentials on the half-line $[0, +\infty)$. If the potential is absolutely summable (i.e. in $l^1([0,+\infty))$ then the formula (2.2) is absolutely convergent, but just as with the linear Fourier transform, there is no obvious reason why this series should converge for potentials that are merely admissible. However, the Plancherel identity (2.17) will allow us to make the NLFT well-defined for such potentials, just as the linear Plancherel identity does the same to the linear Fourier transform on $l^2(\mathbf{Z})$. Our main theorem here is

THEOREM 2.4. *Let $F \in l^2([0,+\infty), \mathcal{D})$ be an admissible potential on $[0, +\infty)$, and let $F_{\leq N}$ be the restriction of F to $[0, N]$. Then the non-linear Fourier transforms $\widehat{F_{\leq N}}$ form a Cauchy sequence in the complete metric space $\mathcal{H}^2_{[0,+\infty)} = \mathcal{H}^2(\mathcal{D})$. In particular we may define the non-linear Fourier transform of F by the formula $\widehat{F} := \lim_{N\to+\infty} \widehat{F_{\leq N}}$. Furthermore, the NLFT is a homeomorphism from $l^2([0,+\infty); \mathcal{D})$ to $\mathcal{H}^2(\mathcal{D})$.*

[Cite Sylvester here, mention Layer stripping.]

PROOF. Let $F \in l^2([0,+\infty);\mathcal{D})$ be an admissible potential on $[0,+\infty)$. From Proposition 2.2 we see that $\widehat{F_{\leq N}}$ lies in $\mathcal{H}^2_{[0,N]}$, and hence in $\mathcal{H}^2(\mathcal{D})$. Now we show that $\widehat{F_{\leq N}}$ is a Cauchy sequence. Let $0 \leq N < N'$. Then we can write $F_{\leq N'} = F_{\leq N} + F_{(N,N']}$, where $F_{(N,N']}$ is the restriction of F to $(N,N']$. If we write $\widehat{F_{\leq N'}} = M[a_{\leq N'}, b_{\leq N'}]$, etc., then by (2.16) we have

$$M[a_{\leq N'}, b_{\leq N'}] = M[a_{\leq N}, b_{\leq N}]M[a_{(N,N']}, b_{(N,N']}] \tag{2.23}$$

In particular we have by (2.9) we have

$$a_{\leq N'} = a_{\leq N}a_{(N,N']}(1 + s_{\leq N}r_{(N,N']}), \tag{2.24}$$

where $s_{\leq N} = b^*_{\leq N}/a_{\leq N}$, etc. Hence we have the pointwise estimate on $\mathbf{T}$

$$\big|\log|a_{\leq N'}| - \log|a_{\leq N}|\big| \leq \log|a_{(N,N']}| + |\log|1 + s_{\leq N}r_{(N,N']}||.$$

But since $|s_{\leq N}|$ is bounded by 1, and $r_{(N,N']}$ has magnitude $\sqrt{1 - |a_{(N,N']}|^{-2}}$, we thus have[6]

$$\big|\log|a_{\leq N'}| - \log|a_{\leq N}|\big| \leq C\log|a_{(N,N']}|.$$

By (2.17) we thus have

$$\|\log|a_{\leq N'}| - \log|a_{\leq N}|\|_{L^1(\mathbf{T})} \leq C\sum_{N<n\leq N'}|F_n|^2. \tag{2.25}$$

In particular we see that $\log|a_{\leq N}|$ is a Cauchy sequence in $L^1(\mathbf{T})$.

Now we consider the $L^2(\mathbf{T})$ convergence of $r_{\leq N}$. We have

$$r_{\leq N'} - r_{\leq N} = \frac{b_{\leq N'}a_{\leq N} - b_{\leq N}a_{\leq N'}}{a_{\leq N'}a_{\leq N}}.$$

But if we rearrange (2.23) as

$$M[a^*_{\leq N}, -b_{\leq N}]M[a_{\leq N'}, b_{\leq N'}] = M[a_{(N,N']}, b_{(N,N']}] \tag{2.26}$$

and apply (2.9) we see that

$$b_{(N,N']} = a_{\leq N}b_{\leq N'} - b_{\leq N}a_{\leq N'}$$

and thus

$$r_{\leq N'} - r_{\leq N} = \frac{b_{(N,N']}}{a_{\leq N'}a_{\leq N}}.$$

But taking operator norms of all terms in we obtain

$$|a_{(N,N']}| \leq C|a_{\leq N'}||a_{\leq N}|$$

and hence

$$|r_{\leq N'} - r_{\leq N}| \leq C\frac{|b_{(N,N']}|}{|a_{(N,N']}|}.$$

Since $1 + |b_{(N,N']}|^2 = |a_{(N,N']}|^2$, we see that

$$\frac{|b_{(N,N']}|^2}{|a_{(N,N']}|^2} \leq C\log|a_{(N,N']}|$$

[6]Here and in the sequel we use C to denote various absolute constants.

and hence by (2.17)

$$\|r_{\leq N'} - r_{\leq N}\|_{L^2}^2 \leq C \sum_{N<n\leq N'} |F_n|^2 \tag{2.27}$$

and so $r_{\leq N}$ is a Cauchy sequence in $L^2(\mathbf{T})$.

Finally, we consider the $L^2(\mathbf{T})$ convergence of $s_{\leq N}$. We begin by applying the Hilbert transform (which is of weak-type $(1,1)$) to (2.25) we see that

$$\|\arg a_{\leq N} - \arg a_{\leq N'}\|_{L^{1,\infty}} \leq C \sum_{N<n\leq N'} |F_n|^2;$$

exponentiating this we obtain

$$\|\frac{a^*_{\leq N}}{a_{\leq N}} - \frac{a^*_{\leq N'}}{a_{\leq N}}\|_{L^{1,\infty}(\mathbf{T})} \leq C \sum_{N<n\leq N'} |F_n|^2$$

and thus (since the expression inside the norm is bounded)

$$\|\frac{a^*_{\leq N}}{a_{\leq N}} - \frac{a^*_{\leq N'}}{a_{\leq N}}\|_{L^2(\mathbf{T})} \leq C \sum_{N<n\leq N'} |F_n|^2.$$

Since $s_{\leq N} = r^*_{\leq N}\frac{a^*_{\leq N}}{a_{\leq N}}$, and similarly for N', we thus see from this and (2.27) that

$$\|s_{\leq N'} - s_{\leq N}\|_{L^2}^2 \leq C \sum_{N<n\leq N'} |F_n|^2.$$

Thus $\widehat{F_{\leq N}}$ is a Cauchy sequence in $\mathcal{H}^2(\mathcal{D})$. Indeed note that we have proven the more precise bound

$$d(\widehat{F_{\leq N}}, \widehat{F_{\leq N'}}) \leq C \sum_{N<n\leq N'} |F_n|^2.$$

We can thus define a non-linear Fourier transform $\widehat{F}$ for any $F \in l^2([0,+\infty);\mathcal{D})$, and we have the convergence estimate

$$d(\widehat{F_{\leq N}}, \widehat{F}) \leq C \sum_{N<n} |F_n|^2. \tag{2.28}$$

Now we show continuity. Let $F^{(k)}$ be any sequence of admissible potentials in $l^2([0,+\infty);\mathcal{D})$ which converges to another potential F in $l^2([0,+\infty);\mathcal{D})$. We need to show that for every $\varepsilon > 0$ we have

$$d(\widehat{F^{(k)}}, \widehat{F}) \leq C\varepsilon$$

for all sufficiently large k. To show this we first choose N large enough so that

$$\sum_{N<n} |F_n|^2 \leq \varepsilon,$$

and then choose k so large so that

$$\sum_{N<n} |F_n^{(k)}|^2 \leq \varepsilon.$$

Thus from (2.28) we have

$$d(\widehat{F_{\leq N}}, \widehat{F}), d(\widehat{F^(k)_{\leq N}}, \widehat{F^{(k)}}) \leq C\varepsilon.$$

From the continuity on finite intervals $[0, N]$ (from Proposition 2.2) we can also choose k large enough so that

$$d(\widehat{F^{(k)}_{\leq N}}, \widehat{F_{\leq N}}) \leq \varepsilon.$$

The claim then follows from the triangle inequality.

Because of this continuity we now know that the symmetries in Lemma 2.1, as well as the groupoid law (2.16) and the Plancherel identity (2.17) will continue to hold for these class of potentials by a limiting argument.

Now we show injectivity of the NLFT from $l^2([0,\infty))$ to $\mathcal{H}^2(\mathcal{D})$. This will be a repetition of the proof of injectivity in Proposition 2.2. Let F be any admissible potential in $l^2([0,\infty))$, and let $F_{\leq N}$ be as before. Then we have by (2.20)

$$F_0 = r_{\leq N}(0) = \int_{\mathbf{T}} r_{\leq N}.$$

Taking limits as $N \to +\infty$ (since $r_{\leq N}$ is convergent to r in $H^2(\mathcal{D})$), we obtain

$$F_0 = r(0) = \int_{\mathbf{T}} r = \frac{b(0)}{a(0)}. \tag{2.29}$$

So we see as before that F_0 can be reconstructed from $\widehat{F}$. Now write $F = F_0\delta_0 + F'$ as before. Applying (2.16), we thus see that

$$\widehat{F} = \frac{1}{\sqrt{1-|F_0|^2}} M[1, F_0] \widehat{F'}, \tag{2.30}$$

and hence we can reconstruct F' from F. If we write $F' = F_1\delta_1 + F''$, where F'' is the restriction of F to $[2, +\infty)$, then by a similar argument to the above (using the translation invariance first to shift F' back to $[0,\infty)$) we can recover F_1 from $\widehat{F'}$. Continuing this "layer stripping" procedure indefinitely we can reconstruct all of the potential F from $\widehat{F}$, which shows the injectivity.

Next, we show surjectivity from $l^2([0,\infty))$ to $\mathcal{H}^2(\mathcal{D})$; again, this is analogous to the corresponding argument in Proposition 2.2. Let $M[a,b]$ lie in $\mathcal{H}^2(\mathcal{D})$; we need to find a potential F in $l^2([0,+\infty);\mathcal{D})$ whose nonlinear Fourier transform equals $M[a,b]$. Since $M[a,b]$ lies in $\mathcal{H}^2(\mathcal{D})$, r extends holomorphically into $\mathcal{D}$ and has magnitude strictly less than 1 on almost all of $\mathbf{T}$. Thus we can define F_0 by (2.29), and we have $|F_0| < 1$. Now we define $M[a',b']$ by (2.21) (or (2.22)). Since $|F_0| < 1$, we see that $1 - F_0^* r$ is holomorphic on $\mathcal{D}$ and is bounded away from zero. Thus by (2.22) a' is an outer function, and by (2.29) we see that $a'(0)$ is a positive real. Also, from (2.22) we see that

$$r' = \frac{b'}{a'} = \frac{-F_0 + r}{1 - F_0^* r};$$

since $1 - F_0^* r$ is bounded away from zero, we see that r' lies in $H^2(\mathcal{D})$; in fact it lies in $H^2_0(\mathcal{D})$ since $r'(0)$ vanishes thanks to (2.29). Finally, from (2.21) we see that $M[a',b']$ is $SU(1,1)$-valued on $\mathbf{T}$. Thus $M[a',b']$ lies in $\mathcal{H}^2_0(\mathcal{D}) = \mathcal{H}^2_{[1,\infty)}$.

We can then translate $M[a',b']$, to $M[z^{-1/2},0]M[a',b']M[z^{1/2},0]$, which lies in $\mathcal{H}^2(\mathcal{D})$, and apply the above procedure again. Undoing the translation, this gives us another complex number $|F_1| < 1$ such that

$$M[a', b'] = \frac{1}{\sqrt{1 - |F_1|^2}} M[1, F_1 z] M[a'', b'']$$

where $M[a'', b''] \in \mathcal{H}^2_{[2,\infty)}$. Continuing this process indefinitely, we obtain a sequence $(F_n)_{n \geq 0}$ of complex numbers with $|F_n| < 1$, such that for every $N \geq 0$ we have

$$M[a, b] = (\prod_{0 \leq n < N} \frac{1}{\sqrt{1 - |F_n|^2}} M[1, F_n z^n]) M[a_{\geq N}, b_{\geq N}]$$

where $M[a_{\geq N}, b_{\geq N}] \in \mathcal{H}^2_{[N,\infty)}$. In other words, if we use the transfer matrices defined in (2.11), (2.12), we have

$$M[a, b] = M_{0 \leftarrow N} M[a_{\geq N}, b_{\geq N}]. \tag{2.31}$$

Write $M_{0 \leftarrow N} =: M[a_{<N}, b_{<N}]$. By Proposition 2.2 we have $M[a_{<N}, b_{<N}] \in \mathcal{H}^2_{[0,N)}$, so $a_{<N}, b_{<N} \in H^2_{[0,N)}$. From (2.9) we have

$$a = a_{<N} a_{\geq N} + b^*_{<N} b_{\geq N} = a_{\geq N}(a_{<N} + b^*_{<N} r_{\geq N}).$$

Since $M[a_{\geq N}, b_{\geq N}] \in \mathcal{H}^2_{[N,\infty)}$, we have $r_{\geq N} \in H^2_{[N,+\infty)}$, and hence $b^*_{<N} r_{\geq N} \in H^2_{(0,+\infty)} = H^2_0(\mathcal{D})$. Thus $b^*_{<N} r_{\geq N}$ vanishes at the origin, and so we have

$$E(M[a, b]) = a(0) = a_{\geq N}(0) a_{<N}(0) = E(M_{0 \leftarrow N}) E(M[a_{\geq N}, b_{\geq N}]).$$

Since the energy of scattering data is always at least 1, we thus have

$$E(M_{0 \leftarrow N}) \leq E(M[a, b]).$$

By the Plancherel identity (2.17) we thus have

$$\sum_{0 \leq n \leq N} \frac{1}{\sqrt{1 - |F_n|^2}} \leq E(M[a, b])$$

and hence F (extended by zero on $(-\infty, 0)$) is an admissible potential with

$$E(F) = \sum_{n \in \mathbf{Z}} \frac{1}{\sqrt{1 - |F_n|^2}} \leq E(M[a, b]).$$

Since F is admissible, we can form the nonlinear Fourier transform $\widehat{F}$ of F. We now claim that $\widehat{F} = M[a, b]$, which would prove the surjectivity. To this end, we define the function $M[a_\infty, b_\infty]$ by

$$M[a_\infty, b_\infty] = M[a, b]^{-1} \widehat{F};$$

our task is to show that $M[a_\infty, b_\infty]$ is the identity $M[1, 0]$. For any $N \geq 0$, we see from (2.16) that

$$\widehat{F} = M_{0 \leftarrow N} \widehat{F_{\geq N}}$$

where $F_{\geq N}$ is the restriction of F to $[N, +\infty)$. From this and (2.31) we thus see that

$$M[a_\infty, b_\infty] = M[a_{\geq N}, b_{\geq N}]^{-1} \widehat{F_{\geq N}}.$$

Writing $\widehat{F_{\geq N}} =: M[a'_{\geq N}, b'_{\geq N}]$, we then see from (2.9) and (2.10) that

$$b_\infty = a_{\geq N} b'_{\geq N} - a'_{\geq N} b_{\geq N} = a_{\geq N} a'_{\geq N} (r'_{\geq N} - r_{\geq N}).$$

Since $a_{\geq N}$, $a'_{\geq N}$ is in the Nevanlinna class on $\mathcal{D}$ and $r'_{\geq N}, r_{\geq N}$ both live in $H^2_{[N,\infty)}$ (because $\widehat{F_{\geq N}}$ and $M[a_{\geq N}, b_{\geq N}]$ both live in $\mathcal{H}^2_{[N,\infty)}$), we thus see that b_∞ is Nevanlinna on $\mathcal{D}$ and vanishes to order N at the origin. But N is arbitrary, hence b_∞ must be identically zero. This implies that $r'_{\geq N} = r_{\geq N}$ for every N; taking magnitudes we thus see that $|a'_{\geq N}| = |a_{\geq N}|$ on $\mathbf{T}$. Since $a'_{\geq N}$ and $a_{\geq N}$ are both outer, this implies that $a'_{\geq N} = a_{\geq N}$, and thus $M[a_{\geq N}, b_{\geq N}] = M[a'_{\geq N}, b_{\geq N}]$. But this implies that $M[a_\infty, b_\infty] = M[1,0]$, and thus $\widehat{F} = M[a,b]$. This proves the surjectivity of the NLFT from $l^2([0,+\infty))$ to $\mathcal{H}^2(\mathcal{D})$.

Finally, it remains to verify that the inverse NLFT is continuous. First observe from (2.29) that the map from $\widehat{F}$ to F_0 is continuous on $\mathcal{H}^2(\mathcal{D})$. From (2.30) this implies that the map from $\widehat{F}$ from $\widehat{F'}$ is continuous from $\mathcal{H}^2(\mathcal{D})$ to $\mathcal{H}^2_0(\mathcal{D})$. Iterating this, we see that the map from $\widehat{F}$ to F_n is continuous on $\mathcal{H}^2(\mathcal{D})$ for all $n \geq 0$. Now let $F^{(k)}$ be any sequence of admissible potentials in $l^2([0,+\infty); \mathcal{D})$ such that $\widehat{F^{(k)}}$ converges to $\widehat{F}$ in $\mathcal{H}^2(\mathcal{D})$; by the above discussion we have the pointwise convergence result $\lim_{k\to\infty} F_n^{(k)} = F_n$ for all $n \in \mathbf{Z}$. In particular, we see that

$$\lim_{k\to\infty} \prod_{0\leq n\leq N} \frac{1}{\sqrt{1-|F_n^{(k)}|^2}} = \prod_{0\leq n\leq N} \frac{1}{\sqrt{1-|F_n|^2}}$$

for any $0 \leq N < +\infty$. Also, since $E(\widehat{F^{(k)}})$ converges to $E(\widehat{F})$, we see from (2.17) that $E(F^{(k)})$ converges to $E(F)$, so in particular

$$\lim_{k\to\infty} \prod_{0\leq n} \frac{1}{\sqrt{1-|F_n^{(k)}|^2}} = \prod_{0\leq n} \frac{1}{\sqrt{1-|F_n|^2}} = E(F).$$

Thus for any $\varepsilon > 0$, if we choose N large enough so that

$$1 \leq \prod_{n>N} \frac{1}{\sqrt{1-|F_n|^2}} < 1+\varepsilon$$

(using the admissibility of F to do so), we see that

$$1 \leq \prod_{n>N} \frac{1}{\sqrt{1-|F_n^{(k)}|^2}} < 1+2\varepsilon$$

for sufficiently large k; taking logarithms, this implies that

$$\sum_{n>N} |F_n^{(k)} - F_n|^2 < C \sum_{n>N} |F_n^{(k)}|^2 + |F_n|^2 < C\varepsilon.$$

But since $F_n^{(k)}$ converges pointwise to F_n, we also have

$$\sum_{0\leq n\leq N} |F_n^{(k)} - F_n|^2 < C\varepsilon$$

for k large enough. This shows that $F^{(k)}$ converges to F in the l^2 topology, which gives inverse continuity as desired. □

Using reflection symmetry (and also translation symmetry by one unit) we obtain the analogous result on the negative half-line:

COROLLARY 2.5. *Let $F \in l^2((-\infty,-1],\mathcal{D})$ be an admissible potential on $(-\infty,-1]$, and let $F_{\geq -N}$ be the restriction of F to $[-N,-1]$. Then the non-linear Fourier transforms $\overbrace{F_{\geq -N}}$ form a Cauchy sequence in the complete metric space $\mathcal{H}^2_{(-\infty,-1]} = \mathcal{H}^2_0(\mathcal{D}^*)$. In particular we may define the non-linear Fourier transform of F by the formula $\overbrace{F} := \lim_{N\to+\infty} \overbrace{F_{\geq -N}}$. Furthermore, the NLFT is a homeomorphism from $l^2((-\infty,-1];\mathcal{D})$ to $\mathcal{H}^2_0(\mathcal{D}^*)$.*

There are of course analogous results for other half-lines such as $(-\infty,q]$ and $[p,+\infty)$ by translation symmetry; the NLFT is a homeomorphism from $l^2((-\infty,q];\mathcal{D})$ onto $\mathcal{H}^2_{(-\infty,q]}$ and from $l^2([p,+\infty);\mathcal{D})$ onto $\mathcal{H}^2_{[p,+\infty)}$ (one can think of these facts as a nonlinear version of the Paley-Wiener theorem). Of course all these definitions of the NLFT are consistent with each other since they all agree on compactly supported potentials, which are dense in all the above spaces, and we have established continuity of the NLFT.

[Mention here the Muscalu-Tao-Thiele counterexample that shows the NLFT is not C^3. But mention also Christ-Kiselev.]

The uniqueness of the inverse NLFT on the half-line is closely related to the uniqueness of the inverse spectral problem for (admissible) Dirac operators on the discrete half-line; we will return to this point briefly when we discuss the eigenfunction problem for such operators. We remark here though that Killip and Simon [**19**] have recently solved the inverse spectral problem for (admissible) Jacobi matrices on the discrete half-line, which has the substantial additional difficulty of potentially having an infinite number of isolated eigenvalues in the spectrum.

2.6. The NLFT on the whole line Z

Having defined the NLFT on both the positive half-line and negative half-line, we now concatenate the two (using (2.16)) on the whole line. The key lemma is

LEMMA 2.6. *If $M[a_-,b_-] \in \mathcal{H}^2_0(\mathcal{D}^*)$ and $M[a_+,b_+] \in \mathcal{H}^2(\mathcal{D})$, then $M[a_-,b_-]M[a_+,b_+] \in \mathcal{L}^2(\mathbf{T})$, and*

$$E(M[a_-,b_-]M[a_+,b_+]) = E(M[a_-,b_-])E(M[a_+,b_+]). \tag{2.32}$$

Furthermore, this multiplication operation is continuous from $\mathcal{H}^2_0(\mathcal{D}^*) \times \mathcal{H}^2(\mathcal{D})$ to $\mathcal{L}^2(\mathbf{T})$.

Later on we shall show that this multiplication map is surjective, but not injective.

PROOF. Write $M[a,b] := M[a_-,b_-]M[a_+,b_+]$. Clearly $M[a,b]$ is $SU(1,1)$-valued. By (2.9) we have

$$a = a_-a_+b_-^*b_+ = a_-a_+(1+r_+s_-). \tag{2.33}$$

But $r_+ \in H^2(\mathcal{D})$ and $s_- \in H^2_0(\mathcal{D})$, with both functions strictly less than 1 a.e. on $\mathbf{T}$, and hence on $\mathcal{D}$ (by the maximum principle). Hence $1 + r_+s_-$ is a Herglotz function on $\mathcal{D}$ which equals 1 at the origin, and in particular is outer. Thus a is outer, and

$$a(0) = a_-(0)a_+(0)$$

which is (2.32). In particular $a(0)$ is real and positive, and so $M[a,b]$ lies in $\mathcal{L}^2(\mathbf{T})$ as claimed.

Now we consider the continuity. Let $M[a_-^{(k)}, b_-^{(k)}]$ be a sequence in $\mathcal{H}_0^2(\mathcal{D}^*)$ converging to $M[a_-, b_-] \in \mathcal{H}_0^2(\mathcal{D}^*)$, and let $M[a_+^{(k)}, b_+^{(k)}]$ be as sequence in $\mathcal{H}^2(\mathcal{D})$ converging to $M[a_+, b_+] \in \mathcal{H}^2(\mathcal{D})$. Write $M[a^{(k)}, b^{(k)}] := M[a_-^{(k)}, b_-^{(k)}]M[a_+^{(k)}, b_+^{(k)}]$ and $M[a, b] := M[a_-, b_-]M[a_+, b_+]$. We need to show that $M[a^{(k)}, b^{(k)}]$ converges to $M[a, b]$. Actually it suffices to do this for a subsequence, since the initial sequence was arbitrary. In particular we may pass to a subsequence for which $M[a_\pm^{(k)}, b_\pm^{(k)}]$ converge pointwise to $M[a_\pm, b_\pm]$, and hence $M[a^{(k)}, b^{(k)}]$ converges pointwise to $M[a, b]$.

The convergence of $r^{(k)}$ to r and $s^{(k)}$ to s then follows from the Lebesgue dominated convergence theorem, so it suffices to check $\log |a^{(k)}|$. From (2.33) we have

$$\log |a| = \log |a_-| + \log |a_+| + \log |1 + r_+ s_-|$$

and similarly

$$\log |a^{(k)}| = \log |a_-^{(k)}| + \log |a_+^{(k)}| + \log |1 + r_+^{(k)} s_-^{(k)}|.$$

Since $\log |a_\pm^{(k)}|$ converges to $\log |a_\pm|$ in $L^1(\mathbf{T})$, it thus suffices to show that

$$\int_{\mathbf{T}} \left| \log |1 + r_+^{(k)} s_-^{(k)}| - \log |1 + r_+ s_-| \right| \to 0.$$

But we can use the identity $|a_\pm^{(k)}|^2 = 1 + |b_\pm^{(k)}|^2$ to easily verify the pointwise estimate

$$|\log |1 + r_+^{(k)} s_-^{(k)}|| \leq C \log |a_-^{(k)}| + C \log |a_+^{(k)}|$$

and the right-hand side is convergent in $L^1(\mathbf{T})$,and so the claim follows from the generalized Lebesgue dominated convergence theorem. $\square$

From the above lemma we can define the NLFT of any admissible potential $F \in l^2(\mathbf{Z}; \mathcal{D})$ by the formula

$$\overbrace{F} = \overbrace{\overbrace{F_{(-\infty,0)}}\overbrace{F_{[0,+\infty}}}$$

where $F_{(-\infty,0]}$ is the restriction of F to $(-\infty, 0]$, and similarly for $F_{[0,+infty)}$. By (2.16) this definition is consistent with the prior definition of the NLFT for compactly supported potentials, and by the above Lemma, Theorem 2.4, and Corollary 2.5 we see that the NLFT is thus a continuous map from $l^2(\mathbf{Z}; \mathcal{D})$ to $\mathcal{L}^2(\mathbf{T})$. Furthermore we have the Plancherel identity (2.17) for all admissible potentials, as well as Lemma 2.1 and (2.16).

One consequence of the continuity of the NLFT on $l^2(\mathbf{Z}; \mathcal{D})$ is that we can now define transfer matrices $M_{n \leftarrow +\infty}$, $M_{-\infty \leftarrow n}$, and $M_{-\infty \leftarrow +\infty}$ by the obvious limiting procedure; note in particular that $M_{-\infty \leftarrow +\infty}$ is exactly the same as $\overbrace{F}$. Also observe that the convergence of $M_{-N \leftarrow N}$ (for instance) to $M_{-\infty \leftarrow +\infty}$ is in the topology of $\mathcal{L}^2(\mathbf{T})$, but this can be altered to pointwise convergence a.e. by passing to a subsequence in the usual manner.

It remains to consider the question of whether the NLFT, as a map from $l^2(\mathbf{Z}; \mathcal{D})$ to $\mathcal{L}^2(\mathbf{T})$, is surjective or injective. In light of the previous results, this is equivalent to asking whether the multiplication map from $\mathcal{H}_0^2(\mathcal{D}^*) \times \mathcal{H}^2(\mathcal{D})$ to $\mathcal{L}^2(\mathbf{T})$ is surjective or injective; in other words, whether for any given scattering

datum $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ there is existence and uniqueness of the *Riemann-Hilbert problem* (RHP)
(2.34)
$$M[a,b] = M[a_-,b_-]M[a_+,b_+]; \quad M[a_-,b_-] \in \mathcal{H}^2_0(\mathcal{D}^*), M[a_+,b_+] \in \mathcal{H}^2(\mathcal{D}).$$

Unfortunately the uniqueness to the RHP (and hence injectivity of the NLFT) can fail. The obstruction is that there are non-trivial elements $M[a_0,b_0]$ in the intersection space $\mathcal{H}^0$, defined by

$$\mathcal{H}^0 := \mathcal{H}^2_0(\mathcal{D}^*) \cap \mathcal{H}^2(\mathcal{D});$$

in other words, there exist $SU(1,1)$-valued functions $M[a_0,b_0]$ such that a_0 is an outer function on $\mathcal{D}$ with $a_0(0) > 0$, and such that b_0/a_0 and b_0^*/a_0 lie in $H^2(\mathcal{D})$ and $H^2_0(\mathcal{D})$ respectively. A simple example is given by

$$b_0 := \frac{2\sinh\theta}{z-1}; a_0 := \frac{e^\theta z - e^{-\theta} z}{z-1}$$

where $\theta > 0$ is an arbitrary parameter. More generally, one can construct examples $M[a_0,b_0]$ in $\mathcal{H}^0$ by choosing b_0 to be any function in both $\mathcal{N}^+(\mathcal{D})$ and$\mathcal{N}^+_0(\mathcal{D}^*)$ (e.g. a rational function vanishing at infinity whose only poles are at $\mathbf{T}$), and then set a_0 to be the unique outer function with magnitude $\log|a_0| = \log\sqrt{1+|b_0|^2}$ and which is real and positive at the origin. These examples clearly show the NLFT is not injective on the full line $\mathbf{Z}$, since Theorem 2.4 and Corollary 2.5 show that $M[a_0,b_0]$ is the non-linear Fourier transform of a non-trivial admissible potential on $[0,+\infty)$ and also a non-trivial admissible potential on $(-\infty,0]$. Equivalently, these examples show that failure of uniqueness for the RHP (2.34), as the identity $M[1,0]M[a_0,b_0] = M[a_0,b_0]M[1,0]$ clearly shows.

Thus the NLFT does not have a globally well-defined inverse. Nevertheless, it will turn out that in many situations the NLFT can be inverted; that there are many examples of scattering data $M[a,b]$ in $\mathcal{L}^2(\mathbf{T})$ which arise from only one admissible potential F in $l^2(\mathbf{Z};\mathcal{D})$; when this happens, we say that $M[a,b]$ has unique inverse NLFT. For instance, we will show in ??? that one has unique inverse NLFT whenever $M[a,b]$ is a bounded function of $\mathbf{T}$, or if $M[a,b]$ lies in $\mathcal{H}^2(\mathcal{D})$ or $\mathcal{H}^2(\mathcal{D}^*)$ and is an L^2 function of $\mathbf{T}$. Furthermore, we can show that the space $\mathcal{H}^0$ is in some sense the only obstruction to inverting the NLFT. More precisely, we will show the following factorization theorem. Let $\mathcal{H}^-$ denote the space of all potentials $M[a_-,b_-]$ in $\mathcal{H}^2_0(\mathcal{D}^*)$ which have unique inverse NLFT, and similarly define $\mathcal{H}^0$ to be the space of all potentials $M[a_+,b_+]$ in $\mathcal{H}^2(\mathcal{D})$ which have unique inverse NLFT.

THEOREM 2.7 (Triple factorization). *Let $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ be a scattering datum. Then there exists a unique factorization*

$$M[a,b] = M[a_{--},b_{--}]M[a_0,b_0]M[a_{++},b_{++}] \tag{2.35}$$

where $M[a_{--},b_{--}] \in \mathcal{H}^-$, $M[a_0,b_0] \in \mathcal{H}^0$, $M[a_{++},b_{++}] \in \mathcal{H}^+$, with

$$E(M[a,b]) = E(M[a_{--},b_{--}])E(M[a_0,b_0])E(M[a_{++},b_{++}]). \tag{2.36}$$

Furthermore, any solution to the Riemann-Hilbert problem (2.34) *can itself be factorized in the form*
(2.37)
$$M[a_-,b_-] = M[a_{--},b_{--}]M[a_{-0},b_{-0}]; \quad M[a_+,b_+] = M[a_{0+},b_{0+}]M[a_{++},b_{++}]$$

where $M[a_{-0}, b_{-0}]$, $M[a_{0+}, b_{0+}]$ *solve the reduced RHP*

$$M[a_{-0}, b_{-0}]M[a_{0+}, b_{0+}] = M[a_0, b_0]; \quad M[a_{-0}, b_{-0}], M[a_{0+}, b_{0+}] \in \mathcal{H}^0. \tag{2.38}$$

Conversely, every solution to the reduced RHP (2.38) *induces a solution to the original RHP* (2.34) *via* (2.37)*.*

Thus the problem of solving the RHP (2.34) is equivalent to that of solving the RHP (2.38); this explains our earlier comment that $\mathcal{H}^0$ was in some sense the only obstruction to inverting the NLFT. In particular we see that $M[a, b]$ has unique inverse NLFT if and only if its middle factor $M[a_0, b_0]$ in the canonical factorization is trivial (i.e. is equal to $M[1, 0]$). To put it another way, the set of functions $M[a, b]$ with unique inverse NLFT is equal to $\mathcal{H}^- \cdot \mathcal{H}^+$.

The rest of the paper is devoted to proving this theorem, which is non-trivial and requires quite a bit of machinery involving the scattering theory of the Dirac operator $L[F]$. While the theorem does give quite a bit of insight into the inversion problem for the NLFT, there are still several questions that we were unable to answer satisfactorily. While we do have an explicit construction for finding $M[a_{--}, b_{--}]$, $M[a_{++}, b_{++}]$, and $M[a_0, b_0]$ from $M[a, b]$, which in principle gives an explicit characterization of $\mathcal{H}^-$, $\mathcal{H}^+$, and of the unique inverse NLFT property, this construction relies on the Riesz representation theorem for Hilbert spaces (cf. the proof of Beurling's theorem) and so is not easy to work with in practice. For instance, we do not yet have a satisfactory necessary and sufficient condition on the potential F in order for $\overbrace{F}$ to have unique inverse NLFT, although we do have some partial results in this direction.

We close this section with a basic result which is consistent with Theorem 2.7.

LEMMA 2.8. *We have* $\mathcal{H}^2_0(\mathcal{D}^*) \cdot \mathcal{H}^0 \subseteq \mathcal{H}^2_0(\mathcal{D}^*)$ *and* $\mathcal{H}^0 \cdot \mathcal{H}^2(\mathcal{D}) \subseteq \mathcal{H}^2(\mathcal{D})$*. As an immediate corollary we have* $\mathcal{H}^0 \cdot \mathcal{H}^0 \subseteq \mathcal{H}^0$*.*

PROOF. It suffices to prove the first claim. Let $M[a_-, b_-] \in \mathcal{H}^2_0$ and $M[a_0, b_0] \in \mathcal{H}^0$, and $M[a, b] := M[a_-, b_-]M[a_+, b_+]$. By Lemma 2.6 we know that $M[a, b] \in \mathcal{L}^2(\mathbf{T})$. From the identity $a_0 a_0^* - b_0 b_0^* = 1$ we can write $a_0 = \frac{1+b_0 b_0^*}{a_0^*}$. Thus a_0 has a holomorphic extension to $\mathcal{D}^*$, as everything on the right- hand side does also and a_0^* is outer on $\mathcal{D}^*$. From the formula $b = a_-^* b_0 + b_- a_0$ (from (2.9)) we thus see that b has a holomorphic extension to $\mathcal{D}^*$ which vanishes at infinity. Since b/a^* was already in $L^2(\mathbf{T})$, it is now in $\mathcal{H}^2_0(\mathcal{D})$ also, and so $M[a, b] \in \mathcal{H}^2_0(\mathcal{D}^*)$ as desired. The second claim in the Lemma is similar, and the third follows immediately by intersecting the first two claims. □

Before we begin the proof of Theorem 2.7, we must return to the Lax operator L defined in (2.4) and relate it to the NLFT. This will be done next.

2.7. Connection between the NLFT and the Lax operator L

Let F be an admissible potential, and let $L = L[F]$ be the Lax operator acting on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ defined in (2.4). In this section we show how the NLFT is connected to the scattering and spectral data of $L[F]$.

We begin by considering the generalized eigenfunction equation

$$L[F]\Phi = \zeta\Phi \tag{2.39}$$

where ζ is a complex number, and Φ is a formal linear combination of the Fourier basis vectors v_n, w_n of $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ defined in the introduction:

$$\Phi = \sum_n \phi_n v_n + \psi_n w_n. \tag{2.40}$$

Our discussion will be purely algebraic for now, and we will not assume that the complex numbers ϕ_n or ψ_n have any decay properties as $n \to \pm\infty$; in particular, Φ need not actually belong to the Hilbert space $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, but may merely be a pair of formal Fourier series; note from (2.4) that L is well defined on such formal objects. Because $L[F]$ is unitary, one should usually think of the spectral parameter ζ as lying on the unit circle $\mathbf{T}$, although for our formal analysis here all we need is that ζ is finite and invertible.

To motivate the analysis let us first consider the trivial case $F \equiv 0$. In this case $L[0]$ is just the direct sum of a left shift and right shift,

$$L[0]v_n = v_{n+1}; \quad Lw_{n+1} = w_n$$

(or in the physical basis $L[0]\Phi = \zeta\Phi$) and the eigenfunction equation (2.39) becomes

$$\phi_{n+1} = \zeta\phi_n; \quad \psi_n = \zeta\psi_{n+1}$$

and hence the general solution in this case is given by[7]

$$\phi_n = a\zeta^n; \quad \psi_{n+1} = b\zeta^{-n} \tag{2.41}$$

for some complex constants a, b. In particular we see that the coefficients of Φ will grow exponentially if $\zeta \in \mathcal{D}$ or $zeta \in \mathcal{D}^*$, and stay bounded (but not in l^2) when $\zeta \in \mathbf{T}$. This is consistent with the well-known fact that the spectrum of $L[0]$ is purely absolutely continuous and is supported on $\mathbf{T}$ (indeed, in the physical space representation $L[0]$ is just the operation of multiplication by ζ on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$.

Now let us consider non-zero admissible potentials F_n. Using (2.4), the eigenfunction equation (2.39) becomes

$$\begin{aligned} \sqrt{1 - |F_n|^2}\phi_n - F_n^*\psi_{n+1} &= \zeta\phi_{n+1} \\ F_n\phi_n + \sqrt{1 - |F_n|^2}\psi_{n+1} &= \zeta\psi_n \end{aligned}$$

which after some algebraic manipulation can be rewritten as

$$\begin{aligned} \phi_n &= \frac{1}{\sqrt{1 - |F_n|^2}}(\zeta\phi_{n+1} + F_n^*\psi_{n+1}) \\ \psi_n &= \frac{1}{\sqrt{1 - |F_n|^2}}(F_n\phi_{n+1} + \zeta^{-1}\psi_{n+1}). \end{aligned}$$

Motivated by (2.41), we now introduce the change of variables

$$\phi_n =: a_n\zeta^n; \quad \psi_n =: b_n\zeta^{1-n} \tag{2.42}$$

[7] In the physical space representation, Φ does not belong to $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ but is instead a (vector-valued) Dirac mass at ζ. Thus one may think of the physical space representation as the spectral representation of $L[0]$.

so that the above equation becomes

$$a_n = \frac{1}{\sqrt{1-|F_n|^2}}(a_{n+1} + F_n^* \zeta^{-2n} b_{n+1})$$
$$b_n = \frac{1}{\sqrt{1-|F_n|^2}}(F_n \zeta^{2n} \phi_{n+1} + b_{n+1}).$$

But this can be written using the transfer matrices (2.11) as

$$M[a_n, b_n](\zeta) = M_{n \leftarrow n+1}(\zeta^2) M[a_{n+1}, b_{n+1}](\zeta).$$

Applying the groupoid law (2.13) we thus see that

$$M[a_n, b_n](\zeta) = M_{n \leftarrow n'}(\zeta^2) M[a_{n'}, b_{n'}](\zeta) \tag{2.43}$$

whenever $-\infty < n < n' < +\infty$. Thus the transfer matrices $M_{n\leftarrow n'}$ form (up to some trivial factors) the fundamental solution of the eigenfunction equation (2.39).

Observe the presence of ζ^2 in the above formula. This suggests that changing the spectral parameter from ζ to $-\zeta$ will not significantly affect the eigenfunction equation (2.39). Indeed, one can see this directly by introducing the parity operator $\sigma : L^2(\mathbf{T}) \oplus L^2(\mathbf{T}) \to L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ defined in the Fourier basis by

$$\sigma v_n := (-1)^n v_n; \quad \sigma w_n := (-1)^{n+1} w_n$$

or equivalently in the physical space representation as

$$\sigma \begin{pmatrix} \alpha(\zeta) \\ \beta(\zeta) \end{pmatrix} := \begin{pmatrix} \alpha(-\zeta) \\ -\beta(-\zeta) \end{pmatrix},$$

and then observing (using the Fourier basis) that $\sigma L[F] \sigma^{-1} = -L[F]$. Thus $L[F]$ is unitarily conjugate to $-L[F]$, which explains the equivalence of the spectral parameters ζ and $-\zeta$. One could exploit this parity property by working with L^2 (with spectral parameter $z := \zeta^2$) instead of $L[F]$; since $L[F]^2$ commutes with σ we can then split $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ into even and odd components[8]. (The analogue of this in the continuous case would be a Dirac operator on $L^2(\mathbf{R}) \oplus L^2(\mathbf{R})$, whose square was direct sum of two scalar Schrödinger operators on $L^2(\mathbf{R})$). The operator $L[F]^2$ is somewhat like a Jacobi matrix, however it does not have as clean a form as the original operator $L[F]$, and so we shall continue working with the Dirac operator directly; it will mean however that all our Hilbert spaces will be in some sense "twice as large" as the analogous objects in the Jacobi theory.

We also observe that $L[F]$ is skew-unitarily conjugate to $L[F]^{-1}$. More precisely, if we define the skew-linear involution $* : L^2(\mathbf{T}) \oplus L^2(\mathbf{T}) \to L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ in the Fourier basis by

$$*v_n := w_n; \quad *w_n := v_n$$

or equivalently in the physical space basis by

$$* \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \beta^* \\ \alpha^* \end{pmatrix},$$

then $*$ is skew-unitary (i.e. $\langle *v, *w \rangle = \overline{\langle v, w \rangle}$ for all $v, w \in L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$), and one can easily verify from (2.4), (2.5) that $*L[F]* = L[F]^{-1}$. Also we have the anti-commutation property $*\sigma = -\sigma *$.

[8]More precisely, one can split into one component where α is even and β is odd, and a component where α is odd and β is even.

Since $L[F]$ is unitary on $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$, the spectral theorem for unitary operators (reference? Reed-Simon, perhaps?) shows that there is a canonical projection-valued measure $d\mu = d\mu[F]$ on $\mathbf{T}$, such that $\mu(\mathbf{T}) = 1$ and

$$f(L[F]) = \int_{\mathbf{T}} f(z)\langle d\mu(z)$$

for any continuous function f on $\mathbf{T}$. This measure splits as usual as $d\mu = d\mu_{ac} + d\mu_{sc} + d\mu_{pp}$. The corresponding projection operators $1 = \mu_{ac}(\mathbf{T}) + \mu_{sc}(\mathbf{T}) + \mu_{pp}(\mathbf{T})$ induces a decomposition of the norm

$$\|v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \|v\|^2_{(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{ac}} + \|v\|^2_{(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{sc}} + \|v\|^2_{(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{pp}}$$

where the semi-norm $\|v\|_{(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{ac}}$ is defined by

$$\|v\|_{(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{ac}} := \|\mu_{ac}(\mathbf{T})v\|_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})},$$

etc. and a corresponding orthogonal decomposition

$$L^2(\mathbf{T}) \oplus L^2(\mathbf{T}) = (L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{ac} \oplus (L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{sc} \oplus (L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{pp}$$

of the Hilbert space $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, where $(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{ac}$ is the range of the orthogonal projection $\mu_{ac}(\mathbf{T})$ (or equivalently, the space of vectors v where $\|v\|_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \|v\|_{(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{ac}}$), etc. For instance, if $F = 0$, then $d\mu$ is purely absolutely continuous (so $d\mu_{sc} = d\mu_{pp} = 0$), and is given by

$$\begin{aligned}
\langle d\mu v_n, v_m\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} &= \zeta^{n-m}\ \frac{|d\zeta|}{2\pi} \\
\langle d\mu w_n, w_m\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} &= \zeta^{m-n}\ \frac{|d\zeta|}{2\pi} \\
\langle d\mu v_n, w_m\rangle_{\mathbf{H}} &= 0 \\
\langle d\mu w_n, v_m\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} &= 0
\end{aligned}$$

for all $n, m \in \mathbf{Z}$, where $\frac{|d\zeta|}{2\pi}$ is normalized arclength measure on $\mathbf{T}$. In the physical space representation $d\mu$ is equally simple:

$$\langle d\mu \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \alpha_2 \\ \beta_2 \end{pmatrix}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = (\alpha_1\alpha_2^* + \beta_1\beta_2^*)(\zeta)\frac{|d\zeta|}{2\pi}.$$

We have shown that the non-linear Fourier transform is connected with the eigenfunction equation for $L[F]$. It will thus be unsurprising that it is also connected with the spectral and scattering theory for $L[F]$. In fact, the reflection and transmission coefficients $1/a$, b/a, b^*/a will be closely related wave operators $\Omega_{0\leftarrow\pm\infty}$ for $L[F]$, which map onto the absolutely continuous component $(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{ac}$ of the Hilbert space $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$.

2.8. Scattering theory

Let $F \in l^2(\mathbf{Z}; \mathcal{D})$ be an admissible potential. We now investigate the scattering theory of the unitary operator $L = L[F]$, i.e. the asymptotic behavior of the evolution operators $L[F]^m$ as $m \to \pm\infty$; one may think of m as a discrete time parameter. We will compare this operator to the corresponding free operators $L[0]^m$, which are of course given explicitly by

$$L[0]^m v_n = v_{n+m}; \quad L[0]^m w_n = w_{n-m}.$$

Observe that $L[0]^m$ can also be given in the physical space representation as

$$L[0]^m(\alpha,\beta)(\zeta) = (\zeta^m\alpha(\zeta), \zeta^m\beta(\zeta)),$$

so we will sometimes write ζ^m instead of $L[0]^m$.

First, we need to understand the matrix coefficients of $L[F]^m$ in the Fourier basis $\{v_n\}_{n\in\mathbf{Z}} \cup \{w_n\}_{n\in\mathbf{Z}}$. We will focus mainly on the positive values of m, since $L[F]^{-m}$ is the adjoint of $L[F]^m$. We recall that the parity operator σ and conjugation operator $*$ are related to $L[F]^m$ by the relations

$$\sigma L[F]^m \sigma^{-1} = (-1)^m L[F]^m; \quad *L[F]^m* = L[F]^{-m}. \tag{2.44}$$

LEMMA 2.9. *Let n, n', m be integers.*

- *(Parity property) If $n+n'+m$ is odd, then*
  $$\langle L[F]^m v_n, v_{n'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \langle L[F]^m v_n, w_{n'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = 0,$$
  while if $n+n'+m$ is even, then
  $$\langle L[F]^m w_n, v_{n'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \langle L[F]^m w_n, w_{n'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = 0.$$
- *(Finite speed of propagation) If $m > 0$, then $\langle L[F]^m v_n, v_{n'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}$ and $\langle L[F]^m w_{n'}, w_n\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}$ vanishes unless $2-m \le n'-n \le m$, while $\langle L[F]^m v_n, w_{n'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}$ and $\langle L[F]^m w_n, v_{n'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}$ vanishes unless $|n'-n| \le m-1$. In other words, $L[F]^m v_n$ is a linear combination of $v_{n+2-m}, \ldots, v_{n+m}$ and $w_{n+1-m}, \ldots, w_{n+m-1}$, while $L[F]^m w_n$ is a linear combination of $v_{n+1-m}, \ldots, v_{n+m-1}$ and $w_{n-m}, \ldots, w_{n+m-2}$.*
- *(Boundary values) If $m > 0$, then*

$$\langle L[F]^m v_n, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \langle L[F]^m w_{n+m}, w_n\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \prod_{n\le n'<n+m} \sqrt{1-|F_{n'}|^2}.$$

PROOF. The parity property can be proven by using the parity property in (2.44). The finite speed of propagation property follows easily from (2.4) and induction. Now we check the boundary value estimate. It suffices to verify the estimate for $\langle L[F]^m v_n, v_{n+m}\rangle$, as the estimate for $\langle L[F]^m w_{n+m}, w_n\rangle$ follows using the conjugation property in (2.44). We use induction. For $m = 1$ the estimate follows directly from (2.4). For $m > 1$ we use (2.4) to compute

$$\langle L[F]^m v_n, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \langle L[F]^{m-1}(\sqrt{1-|F_n|^2}v_{n+1} + F_n w_n), v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}.$$

The contribution of w_n vanishes by finite speed of propagation, so we have

$$\langle L[F]^m v_n, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \sqrt{1-|F_n|^2}\langle L[F]^{m-1} v_{n+1}, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})},$$

and the claim follows by induction. □

As a corollary of the boundary value formula, and the unitarity of $L[F]^m$, we have

$$\begin{aligned}\langle L[F]^{-m}L[0]^m v_n, L[F]^{-m'}L[0]^{m'} v_n\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} &= \langle L[F]^{m'-m} v_{n+m}, v_{n+m'}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} \\ &= \prod_{n+m\le n'<n+m'} \sqrt{1-|F_{n'}|^2}\end{aligned}$$

when $m < m'$. In particular, since F is admissible, we have

$$\lim_{m,m'\to+\infty} \langle L[F]^{-m}L[0]^m v_n, L[F]^{-m'}L[0]^{m'} v_n\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = 1.$$

But since $L[F]^{-m}L[0]^m v_n$ and $L[F]^{-m'}L[0]^{m'}v_n$ are unit vectors, we thus see from the cosine rule that

$$\lim_{m,m'\to+\infty}\|L[F]^{+m}L[0]^m v_n - L[F]^{-m'}L[0]^{m'}v_n\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = 0.$$

In other words, $L[F]^{-m}L[0]^m v_n$ is a Cauchy sequence in $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$. A similar argument shows that $L[F]^{-m}L[0]^m w_n$ is also a Cauchy sequence. By the usual limiting argument (starting with finite linear combinations of basis vectors, and exploiting the uniform boundedness of $L[F]^{-m}L[0]^m$) we may then define the forward wave operator $\Omega_{0\leftarrow+\infty} = \Omega_{0\leftarrow+\infty}[F]$ on $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$ by

$$\Omega_{0\leftarrow+\infty}v := \lim_{m\to+\infty} L[F]^{-m}L[0]^m v$$

for all $v\in L^2(\mathbf{T})\oplus L^2(\mathbf{T})$, where the limit is in the strong sense in $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$. Thus $\Omega_{0\leftarrow+\infty}$ is an isometry on $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$, i.e. it is a unitary transformation from $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$ to the range $\Omega_{0\leftarrow+\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))$. We can similarly define a backward wave operator $\Omega_{0\leftarrow-\infty} = \Omega_{0\leftarrow-\infty}[F]$ on $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$ by

$$\Omega_{0\leftarrow-\infty}v := \lim_{m\to-\infty} L[F]^{-m}L[0]^m v.$$

It is clear from the definitions that we have the intertwining relationships

$$L[F]^k\Omega_{0\leftarrow\pm\infty} = \Omega_{0\leftarrow\pm\infty}L[0]^k = \Omega_{0\leftarrow\pm\infty}\zeta^k \tag{2.45}$$

for any integer k. In particular, we see that $L[F]$ preserves the space $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))$, and the action here is unitarily equivalent to that of $L[0]$. Since $L[0]$ has purely absolutely continuous spectrum, we thus see the range of $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))$ is contained[9] in $(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{ac}$. Later we will show that in fact $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T})) = (L^2(\mathbf{T})\oplus L^2(\mathbf{T}))_{ac}$; in other words, we have asymptotic completeness on the absolutely continuous portion of the spectrum. This was already known for potentials with some decay (see for instance the work of [**35**] for the continuous Schrödinger model), but appears to be new for potentials which are merely in $l^2(\mathbf{Z};\mathcal{D})$, at least in the case of discrete Dirac operators on the line.

From (2.44) we obtain the intertwining property

$$*\Omega_{0\leftarrow+\infty}* = \Omega_{0\leftarrow-\infty}. \tag{2.46}$$

and the parity preservation property

$$\Omega_{0\leftarrow\pm\infty}\sigma = \sigma\Omega_{0\leftarrow\pm\infty}. \tag{2.47}$$

Now define the adjoint wave operators

$$\Omega_{\pm\infty\leftarrow0} := \Omega^*_{0\leftarrow\pm\infty}.$$

These adjoints are unitary from the Hilbert space $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))$ onto $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$ and vanish on the orthogonal complement of this space. Observe that $\Omega_{\pm\infty\leftarrow0}\Omega_{0\leftarrow\pm\infty} = 1$, while $\Omega_{0\leftarrow\pm\infty}\Omega_{\pm\infty\leftarrow0}$ is the orthogonal projection onto $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))$.

We define the linear operators $\alpha_{\pm\infty\leftarrow0}:\mathbf{H}\to L^2(\mathbf{T})$ and $\beta_{\pm\infty\leftarrow0}:\mathbf{H}\to L^2(\mathbf{T})$ by

$$\Omega_{\pm\infty\leftarrow0}(v) =: \begin{pmatrix}\alpha_{\pm\infty\leftarrow0}(v)\\ \beta_{\pm\infty\leftarrow0}(v)\end{pmatrix}; \tag{2.48}$$

[9] In particular this shows that $L[F]$ has absolutely continuous spectrum at almost every point in $\mathbf{T}$; this is the analogue of the well-known result of Deift-Killip (ref?) for Dirac operators on the discrete line.

thus $\alpha_{\pm\infty\leftarrow 0}$ and $\beta_{\pm\infty\leftarrow 0}$ are simply the components of $\Omega_{\pm\infty\leftarrow 0}$. Clearly we have

$$\|\alpha_{\pm\infty\leftarrow 0}(v)\|^2_{L^2(\mathbf{T})} + \|\beta_{\pm\infty\leftarrow 0}(v)\|^2_{L^2(\mathbf{T})} = \|\Omega_{\pm\infty\leftarrow 0}v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}. \tag{2.49}$$

Also, the intertwining property (2.45) implies that

$$\begin{aligned} \alpha_{\pm\infty\leftarrow 0}(L[F]^k v) &= \zeta^k \alpha_{\pm\infty\leftarrow 0}(v) \\ \beta_{\pm\infty\leftarrow 0}(L[F]^k v) &= \zeta^k \beta_{\pm\infty\leftarrow 0}(v) \\ \Omega_{\pm\infty\leftarrow 0}(L[F]^k v) &= \zeta^k \Omega_{\pm\infty\leftarrow 0}(v) \end{aligned} \tag{2.50}$$

while from (2.46) we have

$$\begin{aligned} \alpha_{\pm\infty\leftarrow 0}(*v) &= \beta_{\mp}(v)^* \\ \beta_{\pm\infty\leftarrow 0}(*v) &= \alpha_{\mp}(v)^* \\ \Omega_{\pm\infty\leftarrow 0}(*v) &= *\Omega_{\pm\infty\leftarrow 0}(v) \end{aligned} \tag{2.51}$$

and from (2.47) we have

$$\begin{aligned} \alpha_{\pm\infty\leftarrow 0}(\sigma v)(\zeta) &= \alpha_{\pm\infty\leftarrow 0}(v)(-\zeta) \\ \beta_{\pm\infty\leftarrow 0}(\sigma v)(\zeta) &= -\beta_{\pm\infty\leftarrow 0}(v)(-\zeta) \\ \Omega_{\pm\infty\leftarrow 0}(\sigma v) &= \sigma\Omega_{\pm\infty\leftarrow 0}(v). \end{aligned} \tag{2.52}$$

If we thus apply $\alpha_{\pm\infty\leftarrow 0}$ and $\beta_{\pm\infty\leftarrow 0}$ to the second equation in (2.4), noting that $w_n = *v_n$, we thus obtain the identities

$$\begin{aligned} \zeta\beta_{\mp\infty\leftarrow 0}(v_{n+1})^* &= -F_n^*\alpha_{\pm\infty\leftarrow 0}(v_{n+1}) + \sqrt{1-|F_n|^2}\beta_{\mp}(v_n)^* \\ \zeta\alpha_{\mp\infty\leftarrow 0}(v_{n+1})^* &= -F_n^*\beta_{\pm\infty\leftarrow 0}(v_{n+1}) + \sqrt{1-|F_n|^2}\alpha_{\mp}(v_n)^*. \end{aligned}$$

After some algebra, we then obtain

$$\begin{aligned} \beta_{\mp\infty\leftarrow 0}(v_n)^* &= \frac{1}{\sqrt{1-|F_n|^2}}(\zeta\beta_{\mp\infty\leftarrow 0}(v_{n+1})^* + F_n^*\alpha_{\pm\infty\leftarrow 0}(v_{n+1})) \\ \alpha_{\pm\infty\leftarrow 0}(v_n) &= \frac{1}{\sqrt{1-|F_n|^2}}(F_n\beta_{\mp\infty\leftarrow 0}(v_{n+1})^* + \zeta^{-1}\alpha_{\pm\infty\leftarrow 0}(v_{n+1})). \end{aligned}$$

Using the transfer matrices in (2.11), this can be rewritten (cf. (2.42)) as

$$M[\zeta^{n-1}\beta_{\mp\infty\leftarrow 0}(v_n)^*, \zeta^{-n}\alpha_{\pm\infty\leftarrow 0}(v_n)] = M_{n\leftarrow n+1}(\zeta^{-2})M[\zeta^n\beta_{\mp\infty\leftarrow 0}(v_{n+1})^*, \zeta^{-n-1}\alpha_{\pm\infty\leftarrow 0}(v_{n+1})],$$

and thus by induction

$$M[\zeta^{n-1}\beta_{\mp\infty\leftarrow 0}(v_n)^*, \zeta^{-n}\alpha_{\pm\infty\leftarrow 0}(v_n)] = M_{n\leftarrow n'}(\zeta^{-2})M[\zeta^{n'-1}\beta_{\mp\infty\leftarrow 0}(v_{n'})^*, \zeta^{-n'}\alpha_{\pm\infty\leftarrow 0}(v_{n'})]$$

for all $n < n'$. By (2.50), (2.51) we can write this as

$$\begin{aligned} &M[\alpha_{\pm\infty\leftarrow 0}(L[F]^{n-1}w_n), \alpha_{\pm\infty\leftarrow 0}(L[F]^{-n}v_n)] \\ &\qquad = M_{n\leftarrow n'}(\zeta^{-2})M[\alpha_{\pm\infty\leftarrow 0}(L[F]^{n'-1}w_{n'}), \alpha_{\pm\infty\leftarrow 0}(L[F]^{-n'}v_{n'})]. \end{aligned}$$

Now let $n' \to +\infty$. We know that $L[F]^{-n'}v_{n'}$ converges in $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ to $\Omega_{0\leftarrow+\infty}v_0$ by definition; similarly $L[F]^{n'-1}w_{n'}$ converges to $\Omega_{0\leftarrow-\infty}w_1$. Thus $\alpha_{\pm\infty\leftarrow 0}(L[F]^{n'-1}w_{n'})$ and $\alpha_{\pm\infty\leftarrow 0}(L[F]^{-n'}v_{n'})$ converge in $L^2(\mathbf{T})$ to $\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}w_1)$ and $\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}v_0)$ respectively. By passing to a subsequence we may change this $L^2(\mathbf{T})$ convergence to pointwise a.e. convergence. Also, $M_{n\leftarrow n'}(\zeta^{-2})$ converges (in the $\mathcal{L}^2(\mathbf{T})$ topology in the variable $z = \zeta^{-2}$) to $M_{n\leftarrow+\infty}(\zeta^{-2})$; again we may pass to a subsequence to convert this convergence to pointwise a.e. convergence. We can then conclude that

(2.53)
$$M[\alpha_{\pm\infty\leftarrow 0}(L[F]^{n-1}w_n), \alpha_{\pm\infty\leftarrow 0}(L[F]^{-n}v_n)](\zeta) = M_{n\leftarrow+\infty}(\zeta^{-2})$$
$$M[\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}w_1), \alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}v_0)](\zeta)$$

for almost every $\zeta \in \mathbf{T}$. If we let $n \to -\infty$ and apply a similar argument, we then conclude that

$$M[\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}w_1),\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}v_0)](\zeta) = M_{-\infty\leftarrow+\infty}(\zeta^{-2})$$
$$M[\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}w_1), \alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}v_0)](\zeta)$$

for almost every $\zeta \in \mathbf{T}$.

Recall that $M_{-\infty\leftarrow+\infty} = \widehat{F} = M[a,b]$. Since $\Omega_{\pm\infty\leftarrow 0}\Omega_{0\leftarrow\pm\infty} = 1$), we see from (2.48) that $\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow\pm\infty}v_0) = 1$ and $\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow\pm\infty}w_1) = 0$. Thus we have

$$M[0, \alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}v_0)(\zeta)] = M[a(\zeta^{-2}), b(\zeta^{-2})]M[\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}w_1)(\zeta), 1]$$

and

$$M[\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}w_1)(\zeta), 1] = M[a(\zeta^{-2}), b(\zeta^{-2})]M[0, \alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}v_0)(\zeta)]$$

for a.e. $\zeta \in \mathbf{T}$. Some algebra using (2.9) and the identity $aa^* - bb^* = 1$ then allows us to solve for $\alpha_{\pm\infty\leftarrow 0}(\Omega_{\mp}v_0)$ and $\alpha_{\pm\infty\leftarrow 0}(\Omega_{\mp}w_1)$ in terms of a and b:

(2.54)
$$\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}v_0)(\zeta) = \frac{1}{a(\zeta^{-2})}; \quad \alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}w_1)(\zeta) = -\frac{b^*(\zeta^{-2})}{a(\zeta^{-2})}; \quad \alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}v_0)(\zeta) = \frac{1}{a^*(\zeta^{-2})}$$

Using (2.45), (2.50) we can thus compute $\alpha_{\pm\infty\leftarrow 0}(\Omega_{\mp}v)$ for all basis vectors v and almost every $\zeta \in \mathbf{T}$:

$$\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}v_n)(\zeta) = \zeta^n\frac{1}{a(\zeta^{-2})};$$
$$\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}w_n)(\zeta) = -\zeta^{1-n}\frac{b^*(\zeta^{-2})}{a(\zeta^{-2})};$$
$$\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}v_n)(\zeta) = \zeta^n\frac{1}{a^*(\zeta^{-2})};$$
$$\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}w_n)(\zeta) = \zeta^{1-n}\frac{b^*(\zeta^{-2})}{a^*(\zeta^{-2})}.$$

Using (2.46), (2.51) we can then compute the corresponding formulae for $\beta_{\pm\infty\leftarrow 0}$:

$$\beta_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}w_n)(\zeta) = \zeta^{-n}\frac{1}{a^*(\zeta^{-2})};$$
$$\beta_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}v_n)(\zeta) = -\zeta^{n-1}\frac{b(\zeta^{-2})}{a^*(\zeta^{-2})};$$
$$\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}w_n)(\zeta) = \zeta^{-n}\frac{1}{a(\zeta^{-2})};$$
$$\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}v_n)(\zeta) = \zeta^{n-1}\frac{b(\zeta^{-2})}{a(\zeta^{-2})}.$$

In particular, if we let v be any vector in the range of $\Omega_{0\leftarrow-\infty}$, e.g.

$$v = \Omega_{0\leftarrow-\infty} \sum_n c_n v_n + d_n w_n, \tag{2.55}$$

then we have

$$\alpha_{+\infty\leftarrow 0}(v)(\zeta) = \frac{1}{a(\zeta^{-2})}(\sum_n c_n\zeta^n) - \frac{\zeta b^*(\zeta^{-2})}{a(\zeta^{-2})}(\sum_n d_n\zeta^{-n})$$

and

$$\beta_{+\infty\leftarrow 0}(v)(\zeta) = \frac{\zeta^{-1}b(\zeta^{-2})}{a(\zeta^{-2})}(\sum_n c_n\zeta^n) + \frac{1}{a(\zeta^{-2})}(\sum_n d_n\zeta^{-n}).$$

(Such formulae are justified when c_n, d_n are compactly supported, and then one can use a limiting argument to obtain the general case, noting that b/a, $1/a$, etc. are bounded functions). In particular, since $aa^* - bb^* = |a|^2 - |b|^2 = 1$ on $\mathbf{T}$, we have the pointwise estimate

$$|\alpha_{+\infty\leftarrow 0}(v)(\zeta)|^2 + |\beta_{+\infty\leftarrow 0}(v)(\zeta)|^2 = |\sum_n c_n\zeta^n|^2 + |\sum_n d_n\zeta^{-n}|^2$$

for a.e. $\zeta \in \mathbf{T}$. But the right-hand side is just

$$|\alpha_{-\infty\leftarrow 0}(v)(\zeta)|^2 + |\beta_{-\infty\leftarrow 0}(v)(\zeta)|^2$$

by (2.55) and (2.48) (recall that $\Omega_{0\leftarrow-\infty}$ is an isometry). Integrating this on $\mathbf{T}$ using (2.49), we obtain

$$\|\Omega_{+\infty\leftarrow 0}v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \|\Omega_{-\infty\leftarrow 0}v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}.$$

Since v was an arbitrary element of $\Omega_{0\leftarrow-\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))$, this implies that $\Omega_{0\leftarrow-\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T})) \subseteq \Omega_{0\leftarrow+\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))$. A similar argument shows that $\Omega_{0\leftarrow+\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T})) \subseteq \Omega_{0\leftarrow-\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))$, and so the isometries $\Omega_{0\leftarrow+\infty}$, $\Omega_{0\leftarrow-\infty}$ have the same range:

$$\Omega_{0\leftarrow-\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T})) = \Omega_{0\leftarrow+\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T})).$$

Furthermore, the above argument shows that

$$\alpha_{+\infty\leftarrow 0}(v)(\zeta) = \frac{1}{a(\zeta^{-2})}\alpha_{-\infty\leftarrow 0}(v)(\zeta) - \frac{\zeta b^*(\zeta^{-2})}{a(\zeta^{-2})}\beta_{-\infty\leftarrow 0}(v)(\zeta) \tag{2.56}$$

and

$$\beta_{+\infty\leftarrow 0}(v)(\zeta) = \frac{\zeta^{-1}b(\zeta^{-2})}{a(\zeta^{-2})}\alpha_{-\infty\leftarrow 0}(v)(\zeta) + \frac{1}{a(\zeta^{-2})}\beta_{-\infty\leftarrow 0}(v)(\zeta) \tag{2.57}$$

for all $v \in \Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))$. We can write this in another way. Define the *scattering operator* $\Omega_{+\infty\leftarrow-\infty}$ be the operator on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ defined in the physical space representation as

$$\Omega_{+\infty\leftarrow-\infty}\begin{pmatrix} \alpha_{-\infty}(\zeta) \\ \beta_{-\infty}(\zeta) \end{pmatrix} = \begin{pmatrix} \frac{1}{a(\zeta^{-2})} & \frac{-\zeta b^*(\zeta^{-2})}{a(\zeta^{-2})} \\ \frac{\zeta^{-1}b(\zeta^{-2})}{a(\zeta^{-2})} & \frac{1}{a(\zeta^{-2})} \end{pmatrix}\begin{pmatrix} \alpha_{-\infty}(\zeta) \\ \beta_{-\infty}(\zeta) \end{pmatrix}. \tag{2.58}$$

Then (2.56), (2.57) can be rewritten as

$$\Omega_{+\infty\leftarrow-\infty}\Omega_{-\infty\leftarrow 0} = \Omega_{+\infty\leftarrow 0}. \tag{2.59}$$

Observe that the scattering operator $\Omega_{+\infty\leftarrow-\infty}$ is unitary on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, and its inverse $\Omega_{-\infty\leftarrow+\infty} := \Omega^{-1}_{+\infty\leftarrow-\infty}$ is given by

$$\Omega_{-\infty\leftarrow+\infty}\begin{pmatrix}\alpha_{+\infty}(\zeta)\\ \beta_{+\infty}(\zeta)\end{pmatrix} = \begin{pmatrix}\frac{1}{a^*(\zeta^{-2})} & \frac{\zeta b^*(\zeta^{-2})}{a^*(\zeta^{-2})}\\ \frac{-\zeta^{-1}b(\zeta^{-2})}{a^*(\zeta^{-2})} & \frac{1}{a^*(\zeta^{-2})}\end{pmatrix}\begin{pmatrix}\alpha_{+\infty}(\zeta)\\ \beta_{+\infty}(\zeta)\end{pmatrix}. \tag{2.60}$$

Note that the scattering operator $\Omega_{+\infty\leftarrow-\infty}$ is determined completely by the reflection and transmission coefficients $1/a$, b/a, b^*/a, and conversely one can recover these coefficients from the scattering operator. Thus we recover the well-known interpretation of the functions $a(z)$, $b(z)$ as scattering coefficients of the Dirac operator L; the novelty here is that we are able to handle potentials that lie in $l^2(\mathbf{Z})$ (as opposed to potentials with more decay, e.g. $l^1(\mathbf{Z})$ potentials).

We now show asymptotic completeness on the absolutely continuous portion of the spectrum. For any integer n, we introduce the half-line transfer matrices

$$M[a_{-\infty\leftarrow n}, b_{-\infty\leftarrow n}] := M_{-\infty\leftarrow n} = \overbrace{F_{(-\infty,n)}}$$

and

$$M[a_{n\leftarrow+\infty}, b_{n\leftarrow+\infty}] := M_{n\leftarrow+\infty} = \overbrace{F_{[n,+\infty}};$$

by (2.13) (or (2.16)) we thus have

$$M[a,b] = M[a_{-\infty\leftarrow n}, b_{-\infty\leftarrow n}]M[a_{n\leftarrow+\infty}, b_{n\leftarrow+\infty}]. \tag{2.61}$$

From (2.53) (and (2.45), (2.50)) we have

$$\begin{aligned}M[\zeta^{n-1}\alpha_{\pm\infty\leftarrow 0}(w_n), \zeta^{-n}\alpha_{\pm\infty\leftarrow 0}(v_n)] =& M[a_{n\leftarrow+\infty}(\zeta^{-2}), b_{n\leftarrow+\infty}(\zeta^{-2})]\\ & M[\alpha_{\pm\infty\leftarrow 0}(\Omega_{-\infty\leftarrow n}w_1), \alpha_{\pm\infty\leftarrow 0}(\Omega_{n\leftarrow+\infty}v_0)];\end{aligned}$$

from (2.54) we thus have

$$M[\zeta^{n-1}\alpha_{+\infty\leftarrow 0}(w_n), \zeta^{-n}\alpha_{+\infty\leftarrow 0}(v_n)] = M[a_{n\leftarrow+\infty}(\zeta^{-2}), b_{n\leftarrow+\infty}(\zeta^{-2})]M[-\frac{b^*(\zeta^{-2})}{a(\zeta^{-2})}, 1]$$

and

$$M[\zeta^{n-1}\alpha_{-\infty\leftarrow 0}(w_n), \zeta^{-n}\alpha_{-\infty\leftarrow 0}(v_n)] = M[a_{n\leftarrow+\infty}(\zeta^{-2}), b_{n\leftarrow+\infty}(\zeta^{-2})]M[0, \frac{1}{a^*(\zeta^{-2})}].$$

Thus we can compute the scattering data of v_n and w_n in terms of $a, b, a_{n\leftarrow+\infty}, b_{n\leftarrow+\infty}, a_{-\infty\leftarrow n}, b_{-\infty\leftarrow n}$. Indeed, from the identity

$$\begin{aligned}M[a_{-\infty\leftarrow n}, b_{-\infty\leftarrow n}] &= M[a,b]M[a_{n\leftarrow+\infty}, b_{n\leftarrow+\infty}]^{-1}\\ &= M[a,b]M[a^*_{n\leftarrow+\infty}, -b_{n\leftarrow+\infty}]\\ &= M[aa^*_{n\leftarrow+\infty} - b^*b_{n\leftarrow+\infty}, ba^*_{n\leftarrow+\infty} - b_{n\leftarrow+\infty}a^*]\end{aligned} \tag{2.62}$$

we see that

(2.63)

$$\begin{aligned}
\beta_{-\infty\leftarrow 0}(w_n)^*(\zeta) = \alpha_{+\infty\leftarrow 0}(v_n)(\zeta) &= \zeta^n (a^*_{n\leftarrow+\infty}(\zeta^{-2}) - \frac{b_{n\leftarrow+\infty}(\zeta^{-2})b^*(\zeta^{-2})}{a(\zeta^{-2})}) \\
&= \zeta^n \frac{a_{-\infty\leftarrow n}(\zeta^{-2})}{a(\zeta^{-2})} \\
\beta_{-\infty\leftarrow 0}(v_n)^*(\zeta) = \alpha_{+\infty\leftarrow 0}(w_n)(\zeta) &= \zeta^{1-n} (b^*_{n\leftarrow+\infty}(\zeta^{-2}) - \frac{a_{n\leftarrow+\infty}(\zeta^{-2})b^*(\zeta^{-2})}{a(\zeta^{-2})}) \\
&= -\zeta^{1-n} \frac{b^*_{-\infty\leftarrow n}(\zeta^{-2})}{a(\zeta^{-2})} \\
\beta_{+\infty\leftarrow 0}(w_n)^*(\zeta) = \alpha_{-\infty\leftarrow 0}(v_n)(\zeta) &= \zeta^n \frac{a^*_{n\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})} \\
\beta_{+\infty\leftarrow 0}(v_n)^*(\zeta) = \alpha_{-\infty\leftarrow 0}(w_n)(\zeta) &= \zeta^{1-n} \frac{b^*_{n\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}.
\end{aligned}$$

The reader may verify that these formulae are consistent with (2.56), (2.57).

Now for any integer N, consider the function

$$f_N(\zeta) := \frac{a_{-\infty\leftarrow N}a_{N\leftarrow+\infty} - b^*_{-\infty\leftarrow N}b_{N\leftarrow+\infty}}{a_{-\infty\leftarrow N}a_{N\leftarrow+\infty} + b^*_{-\infty\leftarrow N}b_{N\leftarrow+\infty}}(\zeta^{-2}) = \frac{1 - r_{N\leftarrow+\infty}s_{-\infty\leftarrow N}}{1 + r_{N\leftarrow+\infty}s_{-\infty\leftarrow N}}(\zeta^{-2}).$$

Since $M[a_{-\infty\leftarrow N}, b_{-\infty\leftarrow N}] \in \mathcal{H}^2_0(\mathcal{D}^*)$ and $M[a_{N\leftarrow+\infty}, b_{N\leftarrow+\infty}] \in \mathcal{H}^2(\mathcal{D})$, we see that $r_{N\leftarrow+\infty}$ and $s_{-\infty\leftarrow N}$ extend holomorphically to $\mathcal{D}$ and are strictly less than 1 in magnitude on $\mathcal{D}$, with $s_{-\infty\leftarrow N}$ vanishing at the origin; thus f_N is a Herglotz function on $\mathcal{D}^*$ which equals 1 at infinity. By the Herglotz representation theorem there thus exists a probability measure μ_N on $\mathbf{T}$ such that

$$f_N(\zeta) = \int_{\mathbf{T}} \frac{\zeta + e^{i\theta}}{\zeta - e^{i\theta}} d\mu_N(e^{i\theta})$$

for all $\zeta \in \mathcal{D}^*$. This measure turns out to have the same support properties[10] as the spectral measure μ, and is also useful to establish asymptotic completeness of the wave operators $\Omega_{0\leftarrow\pm\infty}$ on $\mathbf{H}_{ac}$:

PROPOSITION 2.10. *We have* $\Omega_{0\leftarrow-\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T})) = \Omega_{0\leftarrow+\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T})) = \mathbf{H}_{ac}$*; in particular, for every* $v \in \mathbf{H}_{ac}$*, there exists vectors* $v_+, v_- \in L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ *such that*

$$\lim_{m\to\pm\infty} \|L[F]^m v - L[0]^m v_\pm\|_{\mathbf{H}} = 0.$$

Furthermore, for any N*, the support of* μ_{sc} *is the same as that of* $\mu_{N,sc}$*, and the support of* μ_{pp} *is the same as that of* $\mu_{N,pp}$*, while the supports of* μ_{ac} *and* $\mu_{N,ac}$ *are both equal to* $\mathbf{T}$.

The second part of this proposition can also be deduced by evaluating the spectral measure μ at v_n, w_n by means of resolvents (which can be calculated explicitly by modifying the analysis of the eigenfunction equation (2.39)), but we will choose a different argument which will also show the asymptotic completeness of $\Omega_{0\leftarrow\pm\infty}$.

[10]In fact, the μ_N are mutually absolutely continuous with respect to each other as N varies; this is basically because the transfer matrix $M_{N\leftarrow N'}$ and its inverse are bounded on $\mathbf{T}$ for finite $N < N'$.

PROOF. Without loss of generality we may take $N = 0$, since translating F by a finite amount does not change the spectral or scattering properties of $L[F]$ (it merely conjugates $L[F]$ by a translation operator). From (2.45) and (2.46) we know that the space $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))$ is invariant under $L[F]$ and under $*$; thus the orthogonal complement $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))^\perp$ of this space in $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$ is similarly invariant under $L[F]$ and $*$. Since we already know that $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))$ is contained in $\mathbf{H}_{ac}$, the first claim of this proposition will then follow if we can show that the spectrum of $L[F]$ is purely singular on $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))^\perp$. In fact we will show that the spectrum of $L[F]$ on $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))^\perp$ is exactly equal to the support of $\tilde{\mu}_{sc} + \tilde{\mu}_{pp}$.

Let $X \subset L^2(\mathbf{T})\oplus L^2(\mathbf{T})$ be the set of finite linear combinations of basis vectors v_n, w_n; this is a dense subspace of $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$, closed under both $L[F]$ and $*$. We observe that elements of this space are also finite linear combinations of the vectors $\{L[F]^m v_0\}_{m\in\mathbf{Z}} \cup \{L[F]^m w_0\}_{m\in\mathbf{Z}}$ (thus the pair of vectors v_0, w_0 are a cyclic pair of vectors for $L[F]$). To see this, observe from (2.4) that we may write v_1 and w_1 in terms of v_0, w_0, and various powers of $L[F]$; iterating this we can obtain all the basis vectors v_n, w_n for positive n. The negative n are handled by the same argument. Thus we have two quite different bases for generating X; on the one hand we have the orthonormal Fourier basis $\{v_n\}_{n\in\mathbf{Z}} \cup \{w_n\}_{n\in\mathbf{Z}}$, and on the other hand we have[11] a "dynamic basis" $\{L[F]^m v_0\}_{m\in\mathbf{Z}} \cup \{L[F]^m w_0\}_{m\in\mathbf{Z}}$, which is not orthonormal in general (except when $F = 0$, when the dynamic basis coincides with the Fourier basis) but allows us to represent $L[F]$ as a shift operator. This will be very useful, as several facts about X will be easy to check in one basis but quite non-trivial in the other.

We shall begin by rewriting the inner product on $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$, restricted to X, in another way. First observe that for elements $v \in L^2(\mathbf{T})\oplus L^2(\mathbf{T})$, the functions $\alpha_{+\infty\leftarrow 0}(v)$, $\beta_{+\infty\leftarrow 0}(v)$, $\alpha_{-\infty\leftarrow 0}(v)^*$, and $\beta_{-\infty\leftarrow 0}(v)^*$ extend holomorphically to the punctured disk $\mathcal{D}\backslash\{0\}$ (with the singularity at 0 being either removable or a pole). To see this, we see from the above discussion that it suffices to verify it when $v = L[F]^m v_0$ or $v = L[F]^m w_0$. But by (2.50) and (2.51) it suffices to do this when $v = v_0$. But this follows from (2.63), since $M[a_{-\infty\leftarrow 0}, b_{-\infty\leftarrow 0}] \in \mathcal{H}^2_0(\mathcal{D}^*)$, $M[a_{0\leftarrow+\infty}, b_{0\leftarrow+\infty}] \in \mathcal{H}^2(\mathcal{D})$, and $M[a,b] \in \mathcal{L}^2(\mathbf{T})$.

We now analyze the inner product $\langle v, w\rangle_{\mathbf{H}}$ restricted X, as well as the slightly smaller (semi-definite) inner product

$$\langle v, w\rangle_{\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))} := \langle \Omega_{0\leftarrow\pm\infty} v, \Omega_{0\leftarrow\pm\infty} w\rangle_{\mathbf{H}};$$

note from (2.59) and the unitarity of the scattering operators $\Omega_{\mp\infty\leftarrow\pm\infty}$ that this inner product is independent of the choice of sign. By Plancherel (or Parseval) and (2.48) we may write

$$\langle v, w\rangle_{\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))} = \int_{\mathbf{T}} \alpha_{\pm\infty\leftarrow 0}(v)\alpha^*_{\pm\infty\leftarrow 0}(w) + \beta_{\pm\infty\leftarrow 0}(v)\beta^*_{\pm\infty\leftarrow 0}(w);$$

using (2.56), (2.57) and the fact that $|a|^2 - |b|^2 = 1$, we may rewrite this as

$$\begin{aligned}\langle v, w\rangle_{\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))} = \int_{\mathbf{T}} &\alpha_{+\infty\leftarrow 0}(v)(\zeta)\alpha_{-\infty\leftarrow 0}(w)^*(\zeta)a(\zeta^{-2}) \\ &+ \beta_{-\infty\leftarrow 0}(v)(\zeta)\beta_{+\infty\leftarrow 0}(w)^*(\zeta)a^*(\zeta^{-2}).\end{aligned} \tag{2.64}$$

[11] We have not proven the linear independence of this basis, as it is not necessary for our argument, but it can be easily checked using (2.50) and (2.63).

By our previous discussion, for $v, w \in X$ we know that the first term in the integrand extends to the punctured exterior disk $\mathcal{D}^*\backslash\{\infty\}$, while the second term extends to the punctured disk $\mathcal{D}\backslash\{0\}$. Inspired by this, we define the bilinear form $\langle v, w\rangle_X$ on X by the formula

$$\langle v, w\rangle_X := \int_{\mathbf{T}_+} \alpha_{+\infty\leftarrow 0}(v)(\zeta)\alpha_{-\infty\leftarrow 0}(w)^*(\zeta)a(\zeta^{-2}) + \int_{\mathbf{T}_-} \beta_{-\infty\leftarrow 0}(v)(\zeta)\beta_{+\infty\leftarrow 0}(w)^*(\zeta)a^*(\zeta^{-2})$$

where $\mathbf{T}_-$ denotes a circle $\{z \in \mathbf{C} : |z| = \frac{1}{1+\varepsilon}\}$ for some $0 < \varepsilon$, normalized to have total mass 1, and $\mathbf{T}_+$ is similarly defined as $\{z \in \mathbf{C} : |z| = 1+\varepsilon\}$ with total mass 1. Note the Cauchy integral formula ensures that this definition is independent of ε, but as we shall see, we cannot quite take $\varepsilon = 0$, because of the presence of "singular measure" on $\mathbf{T}$; the inner product $\langle v, w\rangle_{\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T})\oplus L^2(\mathbf{T}))}$ will turn out to be just the "absolutely continuous" portion of $\langle v, w\rangle_X$.

We now make the above heuristic discussion rigorous. The form $\langle v, w\rangle_X$ is clearly complex linear in v and skew-linear in w. From (2.50) we see that

$$\langle L[F]v, L[F]w\rangle_X = \langle v, w\rangle_X \tag{2.65}$$

for all $v, w \in X$, while from (2.51) we see that

$$\langle *v, *w\rangle_X = \langle w, v\rangle_X \tag{2.66}$$

for all $v, w \in X$.

We now make the (not entirely obvious) claim that this bilinear form has the self-adjointness property

$$\langle w, v\rangle_X = \overline{\langle v, w\rangle_X}. \tag{2.67}$$

It suffices to verify this when v and w lie in the dynamic basis $\{L[F]^m v_0\}_{m\in\mathbf{Z}} \cup \{L[F]^m w_0\}_{m\in\mathbf{Z}}$, since as mentioned earlier these vectors finitely generate X. Let us first consider the case when $v = L[F]^n v_0$ and $w = L[F]^m w_0$ for some $n, m \in \mathbf{Z}$. Then from (2.63) (with $n = 0$) and (2.50) we have

$$\langle v, w\rangle_X = \int_{\mathbf{T}_+} \frac{\zeta^{n-m-1}a_{-\infty\leftarrow 0}(\zeta^{-2})b_{0\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} + \int_{\mathbf{T}_-} \frac{-\zeta^{n-m-1}b_{-\infty\leftarrow 0}(\zeta^{-2})a^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}$$

and

$$\langle w, v\rangle_X = \int_{\mathbf{T}_+} \frac{-\zeta^{m-n+1}b^*_{-\infty\leftarrow 0}(\zeta^{-2})a_{0\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} + \int_{\mathbf{T}_-} \frac{\zeta^{m-n+1}a^*_{-\infty\leftarrow 0}(\zeta^{-2})b^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}$$

and the claim follows by inspection (notice that the change of variables $\zeta \mapsto \overline{\zeta}^{-1}$ effectively maps $\mathbf{T}_-$ to $\mathbf{T}_+$ and vice versa). By (2.66) it thus remains to verify the case when $v = L[F]^n v_0$ and $w = L[F]^m v_0$; by (2.65) we may take $n = 0$ and $m \geq 0$. First consider the case $m = 0$; our task is to show that $\langle v_0, v_0\rangle_X$ is real. In fact it is equal to 1. To see this, we expand

$$\langle v_0, v_0\rangle_X = \int_{\mathbf{T}_+} \frac{a_{-\infty\leftarrow 0}(\zeta^{-2})a_{0\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} + \int_{\mathbf{T}_-} \frac{-b_{-\infty\leftarrow 0}(\zeta^{-2})b^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}.$$

But the first integrand extends holomorphically to $\mathcal{D}^*$ and has value $\frac{a_{-\infty\leftarrow 0}(0)a_{0\leftarrow+\infty}(0)}{a(0)} = 1$ at infinity, while the second integral extends holomorphically to $\mathcal{D}$ and has value $\frac{-b_{-\infty\leftarrow 0}(\infty)b_{0\leftarrow+\infty}(0)}{a(0)} = 0$ at the origin. Thus $\langle v_0, v_0\rangle_X = 1$ as claimed.

Now consider the case $m > 0$. We expand
(2.68)
$$\langle v_0, L[F]^m v_0\rangle_X = \int_{\mathbf{T}_+} \frac{\zeta^{-m} a_{-\infty\leftarrow 0}(\zeta^{-2}) a_{0\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} + \int_{\mathbf{T}_-} \frac{-\zeta^{-m} b_{-\infty\leftarrow 0}(\zeta^{-2}) b^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}$$
and
(2.69)
$$\langle L[F]^m v_0, v_0\rangle_X = \int_{\mathbf{T}_+} \frac{\zeta^{m} a_{-\infty\leftarrow 0}(\zeta^{-2}) a_{0\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} + \int_{\mathbf{T}_-} \frac{-\zeta^{m} b_{-\infty\leftarrow 0}(\zeta^{-2}) b^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}.$$
The first integrand in (2.68) is holomorphic on $\mathcal{D}^*$ and vanishes at infinity, so the integral is zero. Similarly the second integrand in (2.69) is holomorphic on $\mathcal{D}$ and vanishes at the origin, so the integral is also zero. Conjugating the remaining term in (2.69), it thus suffices to show that
$$\int_{\mathbf{T}_-} \frac{-\zeta^{-m} b_{-\infty\leftarrow 0}(\zeta^{-2}) b^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})} = \int_{\mathbf{T}_-} \frac{\zeta^{-m} a^*_{-\infty\leftarrow 0}(\zeta^{-2}) a^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}.$$
But this follows since $a^* = a^*_{-\infty\leftarrow 0} a^*_{0\leftarrow+\infty} + b_{-\infty\leftarrow 0} b^*_{0\leftarrow+\infty}$, and hence
$$\int_{\mathbf{T}_-} \zeta^{-m} \frac{a^*_{-\infty\leftarrow 0}(\zeta^{-2}) a^*_{0\leftarrow+\infty}(\zeta^{-2}) + b_{-\infty\leftarrow 0}(\zeta^{-2}) b^*_{0\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} = \int_{\mathbf{T}_-} \zeta^{-m} = 0.$$
We can now equate this bilinear form with the inner product on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$.

LEMMA 2.11. *For all $v, w \in X$, we have $\langle v, w\rangle_X = \langle v, w\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}$.*

We remark that this lemma can be used to compute the matrix coefficients of $L[F]^m$ explicitly in terms of transfer matrices, but we will not do so here.

PROOF. It suffices to verify this when v and w lie in the orthonormal basis $\{v_n\}_{n\in\mathbf{Z}} \cup \{w_n\}_{n\in\mathbf{Z}}$. We have already shown that $\langle v_0, v_0\rangle_X = \langle v_0, v_0\rangle_{\mathbf{H}} = 1$; a similar argument shows that $\langle v_n, v_n\rangle_X = \langle v_n, v_n\rangle_{\mathbf{H}} = 1$ for all $n \in \mathbf{Z}$. Applying (2.66) we see a similar statement for the w_n.

Now we check the remaining inner products of basis vectors. By (2.66) it suffices to show that
$$\langle v_n, w_m\rangle_X = 0 \text{ for all } n, m$$
and
$$\langle v_n, v_m\rangle_X = 0 \text{ for all } n \geq m.$$
By (2.67) we may take[12] $m \geq n$. By (2.63) we thus have
$$\langle v_n, w_m\rangle_X = \int_{\mathbf{T}_+} \zeta^{n+m-1} \frac{a_{-\infty\leftarrow n}(\zeta^{-2}) b_{m\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} - \int_{\mathbf{T}_-} \zeta^{n+m-1} \frac{b_{-\infty\leftarrow n}(\zeta^{-2}) a^*_{m\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}.$$
But since $r_{m\leftarrow+\infty} \in H^2_{[m,+\infty)}$ and all the "a" functions are outer on $\mathcal{D}$, we see that the first integrand is holomorphic on $\mathcal{D}^*$ and has a zero of order $2m-(n+m-1) > 0$ at infinity, so the first integral vanishes. Similarly, since $s_{-\infty\leftarrow n} \in H^2_{[n+1,+\infty)}$ and all the "a" functions are outer on $\mathcal{D}$, we see that the second integrand is holomorphic on $\mathcal{D}$ and has a zero of order $n+m-1-2(n-1) > 0$ at the origin, so the second integral also vanishes.

[12]This use of the non-trivial self-adjointness property (2.67) is crucial, as the computation is far messier for $m < n$!

Now we expand $\langle v_n, v_m\rangle_X$ similarly as

$$\langle v_n, v_m\rangle_X = \int_{\mathbf{T}_+} \zeta^{n-m}\frac{a_{-\infty\leftarrow n}(\zeta^{-2})a_{m\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})} + \int_{\mathbf{T}_-} \zeta^{n-m}\frac{b_{-\infty\leftarrow n}(\zeta^{-2})b^*_{m\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}.$$

The first integrand is holomorphic on $\mathcal{D}^*$ with a zero of order $m-n$ at infinity (since all the "a" functions are outer on $\mathcal{D}$), and so the first integral vanishes. Since $s_{-\infty\leftarrow n} \in H^2_{[1-n,+\infty)}$ and $r_{m\leftarrow+\infty} \in H^2_{[m,+\infty)}$, and all the "$a$" functions are outer on $\mathcal{D}$, we see that the second integrand is holomorphic on $\mathcal{D}$ with a zero of order $m-n+2$ at the origin, so the second integral also vanishes. This proves the Lemma. □

As a corollary of the above lemma we see that

$$\|v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \int_{\mathbf{T}_+} \alpha_{+\infty\leftarrow 0}(v)(\zeta)\alpha_{-\infty\leftarrow 0}(v)^*(\zeta)a(\zeta^{-2}) + \int_{\mathbf{T}_-} \beta_{-\infty\leftarrow 0}(v)(\zeta)\beta_{+\infty\leftarrow 0}(v)^*(\zeta)a^*(\zeta^{-2})$$

for all $v \in X$, while from (2.64) we have

$$\|\Omega_{\pm\infty\leftarrow 0}v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \int_{\mathbf{T}} \alpha_{+\infty\leftarrow 0}(v)(\zeta)\alpha_{-\infty\leftarrow 0}(v)^*(\zeta)a(\zeta^{-2})\beta_{-\infty\leftarrow 0}(v)(\zeta)\beta_{+\infty\leftarrow 0}(v)^*(\zeta)a^*(\zeta^{-2}).$$

Since the left-hand sides are real, we can conjugate the second factor of each to obtain

$$\|v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \mathrm{Re}\int_{\mathbf{T}_+} (\alpha_{+\infty\leftarrow 0}(v)(\zeta)\alpha_{-\infty\leftarrow 0}(v)^*(\zeta) + \beta_{-\infty\leftarrow 0}(v)^*(\zeta)\beta_{+\infty\leftarrow 0}(v)(\zeta))a(\zeta^{-2})$$
$$\|\Omega_{\pm\infty\leftarrow 0}v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \mathrm{Re}\int_{\mathbf{T}} (\alpha_{+\infty\leftarrow 0}(v)(\zeta)\alpha_{-\infty\leftarrow 0}(v)^*(\zeta) + \beta_{-\infty\leftarrow 0}(v)^*(\zeta)\beta_{+\infty\leftarrow 0}(v)(\zeta))a(\zeta^{-2}).$$

To compare these expressions, we use the dynamic basis to write v in terms of the vectors $L[F]^m v_0$, $L[F]^m w_0$ for various $m \in \mathbf{Z}$. In other words, we may write

$$v = c(L[F])v_0 + d(L[F])w_0$$

for some Laurent polynomials c, d. By (2.50) we thus have

$$\alpha_{\pm\infty\leftarrow 0}(v) = c(\zeta)\alpha_{\pm\infty\leftarrow 0}(v_0) + d(\zeta)\alpha_{\pm\infty\leftarrow 0}(w_0); \quad \beta_{\pm\infty\leftarrow 0}(v) = c(\zeta)\beta_{\pm\infty\leftarrow 0}(v_0) + d(\zeta)\beta_{\pm\infty\leftarrow 0}(w_0);$$

using (2.63) we thus have

$$\alpha_{+\infty\leftarrow 0}(v) = \frac{c(\zeta)a_{-\infty\leftarrow 0}(\zeta^{-2}) - \zeta d(\zeta)b^*_{-\infty\leftarrow 0}(\zeta^{-2})}{a(\zeta^{-2})}$$
$$\beta_{+\infty\leftarrow 0}(v) = \frac{\zeta^{-1}c(\zeta)b_{0\leftarrow+\infty}(\zeta^{-2}) + d(\zeta)a_{0\leftarrow+\infty}(\zeta^{-2})}{a(\zeta^{-2})}$$
$$\alpha_{-\infty\leftarrow 0}(v) = \frac{c(\zeta)a^*_{0\leftarrow+\infty}(\zeta^{-2}) + \zeta d(\zeta)b^*_{0\leftarrow+\infty}(\zeta^{-2})}{a^*(\zeta^{-2})}$$
$$\beta_{-\infty\leftarrow 0}(v) = \frac{-\zeta^{-1}c(\zeta)b_{-\infty\leftarrow 0}(\zeta^{-2}) + d(\zeta)a^*_{-\infty\leftarrow 0}(\zeta^{-2})}{a^*(\zeta^{-2}}.$$

Substituting this into our previous formulae for $\|v\|_{\mathbf{H}}$ and $\|\Omega_{0\leftarrow\pm\infty}v\|_{\mathbf{H}}$, we obtain

$$\|v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \mathrm{Re}\int_{\mathbf{T}_+} (cc^* + dd^*)f_0 + 2\zeta^{-1}cd^*g + 2\zeta dc^*h$$
$$\|\Omega_{\pm\infty\leftarrow 0}v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \mathrm{Re}\int_{\mathbf{T}} (cc^* + dd^*)f_0 + 2\zeta^{-1}cd^*g + 2\zeta dc^*h, \tag{2.70}$$

where

$$f_0(\zeta) := \frac{a_{-\infty\leftarrow 0}a_{0\leftarrow +\infty} - b^*_{-\infty\leftarrow N}b_{0\leftarrow +\infty}}{a}(\zeta^{-2}) = \frac{1 - r_{0\leftarrow +\infty}s_{-\infty\leftarrow 0}}{1 + r_{0\leftarrow +\infty}s_{-\infty\leftarrow 0}}(\zeta^{-2})$$

is the function defined earlier, and g, h are the auxiliary functions

$$g(\zeta) := \frac{a_{-\infty\leftarrow 0}b_{0\leftarrow +\infty}}{a}(\zeta^{-2}) = \frac{r_{0\leftarrow +\infty}}{1 + r_{0\leftarrow +\infty}s_{-\infty\leftarrow 0}}(\zeta^{-2})$$

and

$$h(\zeta) := -\frac{b^*_{-\infty\leftarrow 0}a_{0\leftarrow +\infty}}{a}(\zeta^{-2}) = -\frac{s_{-\infty\leftarrow 0}}{1 + r_{0\leftarrow +\infty}s_{-\infty\leftarrow 0}}(\zeta^{-2}).$$

We observed earlier that f_0 is Herglotz on $\mathcal{D}^*$., with associated measure μ_0 on $\mathbf{T}$. In fact we have the more general statement that $f_0 + \omega g + \overline{\omega} h$ is Herglotz on $\mathcal{D}^*$ for all complex numbers ω in the closed unit disk $\overline{D} = \{z \in \mathbf{C} : |z| \leq 1\}$. Indeed, it is clear that $f_0 + \omega g + \overline{\omega} h$ is holomorphic on $\mathcal{D}^*$ and equals 0 at infinity; it suffices then to check that it has positive real part. By convexity it suffices to verify the case when $|\omega| = 1$. But then we can factorize

$$f_0 + \omega g + \overline{\omega} h = \frac{(1 + \omega r_{0\leftarrow +\infty})(1 - \overline{\omega} s_{-\infty\leftarrow 0})}{1 + r_{0\leftarrow +\infty}s_{-\infty\leftarrow 0}}(\zeta^{-2}).$$

To show that this has positive real part, it suffices to show that its reciprocal does, i.e. that

$$\mathrm{Re}\frac{1 + r_{0\leftarrow +\infty}s_{-\infty\leftarrow 0}}{(1 + \omega r_{0\leftarrow +\infty})(1 - \overline{\omega} s_{-\infty\leftarrow 0})}(\zeta^{-2}) > 0.$$

But we can split the left-hand side as

$$\frac{1}{2}\mathrm{Re}\frac{1 - \omega r_{0\leftarrow +\infty}}{1 + \omega r_{0\leftarrow +\infty}}(\zeta^{-2}) + \frac{1}{2}\mathrm{Re}\frac{1 + \overline{\omega} s_{-\infty\leftarrow 0}}{1 - \overline{\omega} s_{-\infty\leftarrow 0}}(\zeta^{-2})$$

and both terms are clearly positive (since $r_{0\leftarrow +\infty}$ and $s_{-\infty\leftarrow 0}$ are bounded by 1 on $\mathcal{D}$).

By the Herglotz representation theorem, we can thus associate to each $\omega \in \overline{\mathcal{D}}$ a probability measure μ_ω on $\mathbf{T}$ such that

$$(f_0 + \omega g + \overline{\omega} h)(\zeta) = \int_{\mathbf{T}} \frac{\zeta + e^{i\theta}}{\zeta - e^{i\theta}} d\mu_\omega(e^{i\theta}) \tag{2.71}$$

for all $\zeta \in \mathcal{D}^*$. By taking various linear combinations of this identity, we thus see that μ_ω must depend on ω linearly, in the sense that

$$\mu_\omega = \mu_0 + \omega\nu + \overline{\omega\nu}$$

for some complex measures ν, $\overline{\nu}$; since μ_ω is always real we see that $\overline{\nu}$ must indeed be the conjugate of ν, as the notation suggests. Since the μ_ω are always positive for all $\omega \in \overline{\mathcal{D}}$, we easily see from this identity (and the [?Hahn?] decomposition theorem for measures) that ν is absolutely continuous with respect to μ_0, and in fact must take the form $\nu = \rho\mu_0$ for some function $\rho \in L^\infty(\mu_0)$ with $\|\rho\|_{L^\infty(\mu_0)} \leq 1/2$. Thus

$$\mu_\omega = (1 + \omega\rho + \overline{\omega\rho})\mu_0,$$

which then implies from (2.71) that

$$f_0(\zeta) = \int_{\mathbf{T}} \frac{\zeta + e^{i\theta}}{\zeta - e^{i\theta}} d\mu_0(e^{i\theta})$$
$$g(\zeta) = \int_{\mathbf{T}} \frac{\zeta + e^{i\theta}}{\zeta - e^{i\theta}} \rho d\mu_0(e^{i\theta})$$
$$g(\zeta) = \int_{\mathbf{T}} \frac{\zeta + e^{i\theta}}{\zeta - e^{i\theta}} \rho^* d\mu_0(e^{i\theta}).$$

Applying this to (2.70), and observing that f_0, g, h are all finite linear combinations of Herglotz functions, we obtain

$$\|v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \mathrm{Re} \int_{\mathbf{T}} (cc^* + dd^*) d\mu_0 + 2\zeta^{-1} cd^* \rho d\mu_0 + 2\zeta dc^* \rho^* d\mu_0$$
$$\|\Omega_{\pm\infty\leftarrow 0} v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \mathrm{Re} \int_{\mathbf{T}} (cc^* + dd^*) d\mu_{0,ac} + 2\zeta^{-1} cd^* \rho d\mu_{0,ac} + 2\zeta dc^* \rho^* d\mu_{0,ac}.$$

Actually the Re can be dropped since the integrands are already manifestly real. In particular we see that

(2.72)
$$\|v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} - \|\Omega_{\pm\infty\leftarrow 0} v\|^2_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \int_{\mathbf{T}} (|c|^2 + |d|^2 + 4\mathrm{Re}(\zeta^{-1} cd^* \rho))(d\mu_{0,sc} + d\mu_{0,pp})$$

for all $v \in X$. Now recall that X is dense in $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, and that $L[F]$ on X corresponds to multiplying c and d by ζ. Thus the action of $L[F]$ on $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))^\perp$ is unitarily equivalent to the action of multiplication by ζ on the space formed by the space $L^2_\rho(\mu_{0,sc} + \mu_{0,pp})$, defined as the closure of the space $\{(c(z), d(z)) : c, d \textit{ Laurent polynomials}\}$ under the semi-norm

$$\|(c,d)\|^2 := \int_{\mathbf{T}} (|c|^2 + |d|^2 + 4\mathrm{Re}(\zeta^{-1} cd^* \rho))(d\mu_{0,sc} + d\mu_{0,pp})$$

and with any null vectors (which can occur when $|\rho| = 1/2$; see below) quotiented out. This shows $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))^\perp$ has no absolutely continuous portion of the spectrum of $L[F]$, and hence $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T})) = \mathbf{H}_{ac}$ as desired. The above discussion also shows that the singular continuous and pure point components of μ coincide with that of μ_0, as claimed. The claim about μ_0 and μ both having absolutely continuous spectrum equal to $\mathbf{T}$ is also clear; the former follows from inspection (since f_0 is non-zero a.e. on $\mathbf{T}$, in fact it is log-integrable), and the latter follows since the spectrum of $L[F]$ on $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))$ is (by (2.45)) the same as that of $L[0]$ on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, which is absolutely continuous on all of $\mathbf{T}$. $\square$

Remark. Morally speaking, when singular spectrum occurs, $1+s_{-\infty\to 0} r_{0\to+\infty}(\zeta^{-2})$ must vanish, and ρ should equal $\frac{1}{2} r_{0\to+\infty}(\zeta^{-2}) = -\frac{1}{2} s_{-\infty\to 0}(\zeta^{-2})$. This heuristic can be made rigorous; indeed, one to show that $|\rho| = 1/2$ almost everywhere with respect to $\mu_{0,sc} + \mu_{0,pp}$, which implies that $L^2_\rho(\mu_{0,sc})$, and hence $\mathbf{H}_{sc}$ is unitarily equivalent to $L^2(\mu_{0,sc})$, and similarly for the pure point portion of the spectrum; this can be thought of as a generalization of the well-known fact (shown for instance via Wronskians) that an eigenvalue equation such as (2.39) can have at most one linearly independent solution in $l^2(\mathbf{Z})$. We sketch the argument as follows. The idea is to use the fact that $L[F]^{-m} v_m$ converges to $\Omega_{0\leftarrow+\infty} v_0$ as $m \to +\infty$, and so its projection to $\Omega_{0\leftarrow\pm\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))^\perp$ goes to zero. However, one can show that

$L[F]^{-m}v_m = c_m(L[F])v_0 + d_m(L[F])w_0$ for some Laurent polynomials $c_m(\zeta), d_m(\zeta)$ satisfying $c_m c_m^* - d_m d_m^* = 1$ (in fact $M[c_m, \zeta^{-1}d_m] = M_{0\leftarrow m}^{-1}(\zeta^2)$, as can be seen by induction). These facts are only compatible with (2.72) if $|\rho| = 1/2$ a.e. on $\mu_{0,sc} + \mu_{0,pp}$. One can then use ρ to define $s_{-\infty\to 0}(\zeta^{-2})$ and $r_{0\to+\infty}(\zeta^{-2})$ on the singular spectrum a.e. on $\mu_{0,sc} + \mu_{0,pp}$.

Proposition 2.10 gives us a criterion to test when a Dirac operator $L[F]$ with admissible potential F has purely absolutely continuous spectrum; this occurs if and only if the function f_0 (or any other f_N) is a Herglotz function $\mathcal{D}^*$ generated by an absolutely continuous measure. (Equivalently, e^{f_0} must be an outer function on $\mathcal{D}^*$). Note that this criterion depends on half-line transfer matrices $M_{N\to+\infty}$, $M_{-\infty\to N}$ as well as the nonlinear Fourier transform $\widehat{F}$. In fact the nonlinear Fourier transform $\widehat{F}$ is not always, by itself, enough to determine whether $L[F]$ has any singular spectrum or not. For instance, consider the example $M[a,b] \in \mathcal{H}_0 \subset \mathbf{L}^2(\mathbf{T})$ given by

$$b := \frac{2\sinh\theta_0}{z-1}; a := \frac{e^{\theta_0}z - e^{-\theta_0}z}{z-1}$$

which was considered earlier; here $\theta_0 > 0$ is an arbitrary parameter. From Theorem 2.4, Corollary 2.5, Lemma 2.6 we know that every solution of the Riemann-Hilbert problem (2.34) gives rise to a potential F with $\widehat{F} = M[a,b]$, and with $M_{0\to+\infty} = M[a_+, b_+]$ and $M_{-\infty\to 0} = M[a_-, b_-]$. We will have purely absolutely continuous spectrum if and only if the function

$$\frac{1 - s_- r_+}{1 + s_- r_+} = \frac{a_- a_+ + b_-^* b_+}{a}, \tag{2.73}$$

which is necessarily Herglotz on $\mathcal{D}$, arises from an absolutely continuous measure (we now work in $\mathcal{D}$ instead of $\mathcal{D}^*$ by means of the change of variables $z = \zeta^{-2}$. For instance, in the two extreme factorizations

$$M[a,b] = M[1,0]M[a,b]$$

and

$$M[a,b] = M[a,b]M[1,0]$$

which give rise to potentials supported on $[0,+\infty)$ and $(-\infty, 0]$ respectively, the function (2.73) is identically 1 and so there is no singular spectrum; as observed in [**34**], this shows that the problem of non-uniqueness for the inverse NLFT cannot be solved simply by restricting one's attention to Dirac operators with purely absolutely continous spectrum. In fact half-line admissible potentials can never have any singular continuous or pure point spectrum[13]. However, there are intermediate

[13]The reader may object that there are many constructions on the half-line that give embedded eigenvalues, etc. for admissible potentials, but in those constructions a non-zero boundary condition is imposed, e.g. in the notation of (2.39), (2.40) one might impose that $(\phi_0, \psi_0) = (1, 0)$ or $(\phi_0, \psi_0) = (0, 1)$; this would correspond to the singular nature of a function such as $\frac{1\mp r_+}{1\pm r_+}$ rather than $\frac{1-s_-r_+}{1+s_-r_+}$. In the full line setting, the "boundary condition" is that the eigenfunction Φ must decay as an l^2 function in *both* directions, which in the case of a half-line potential forces $(\phi_0, \psi_0) = (0, 0)$, which then quickly forces all of Φ to vanish. It is however an interesting question to see how the singular spectrum of two half-line potentials (with appropriate boundary conditions) relate to the singular spectrum of the concatenated full line potential; the compatibility of boundary conditions of the half-line potentials is of course crucial.

factorizations

$$M[a,b] = M[a_\theta, b_\theta]M[a_{\theta_0-\theta}, b_{\theta_0-\theta}] \tag{2.74}$$

for $0 \leq \theta \leq \theta_0$, where

$$b_\theta := \frac{2\sinh\theta}{z-1}; a_\theta := \frac{e^\theta z - e^{-\theta}}{z-1};$$

of course the two extremes $\theta = 0$, $\theta = \theta_0$ correspond to the two extreme factorizations given earlier. (For $\theta < 0$ or $\theta > \theta_0$, either a_θ or $a_{\theta_0-\theta}$ ceases to be an outer function on $\mathcal{D}$). But for the intermediate factorizations $0 < \theta < \theta_0$, the function (2.73) can be seen to have a simple pole at 1, and so the corresponding measure $d\mu_0$ will acquire a point mass at 1 (although the mass of this point will go to zero as $\theta \to 0$ or $\theta \to \theta_0$). Thus the Dirac operators corresponding the intermediate factorizations have an embedded eigenvalue at 1, i.e. an eigenfunction in $l^2(\mathbf{Z})$ with eigenvalue 1.

One can in fact show that the factorizations in (2.74) are the only solutions to the Riemann-Hilbert problem with the specified $M[a,b]$; we will tackle this (and the more general problem of solving Riemann-Hilbert problems for rational function data) in a future paper. Thus there is a continuum of solutions here to the RHP, with the two extremes admitting no singular spectrum and the intermediate cases exhibiting some singular spectrum. This turns out to be the typical[14] behavior, as we shall see when we derive triple factorization.

We close this section with an alternate characterization of the presence of singular spectrum.

Lemma 2.12. *Let F be an admissible potential with non-linear Fourier transform $M[a,b]$, let N be an arbitrary integer, and let $M[a_{-\infty\leftarrow N}, b_{-\infty\leftarrow N}] = M_{-\infty\leftarrow N}$ and $M[a_{N\leftarrow+\infty}, b_{N\leftarrow+\infty}] = M_{N\leftarrow+\infty}$ be the half-line transfer matrices. Then we have*

$$\int_{\mathbf{T}} \frac{|a_{-\infty\leftarrow N}|^2 + |b_{N\leftarrow+\infty}|^2}{|a|^2} = \int_{\mathbf{T}} \frac{|b_{-\infty\leftarrow N}|^2 + |a_{N\leftarrow+\infty}|^2}{|a|^2} \leq 1$$

and equality occurs if and only if $L[F]$ has purely absolutely continuous spectrum. In particular, we see that $a_{-\infty\leftarrow N}/a$, $b_{-\infty\leftarrow N}/a$, $a_{N\leftarrow+\infty}/a$, and $b_{N\leftarrow+\infty}/a$ all lie in $L^2(\mathbf{T})$.

Proof. We recall that the function f_N defined by

$$f_N(\zeta) = \frac{a_{-\infty\leftarrow N}a_{N\leftarrow+\infty} - b^*_{-\infty\leftarrow N}b_{N\leftarrow+\infty}}{a_{-\infty\leftarrow N}a_{N\leftarrow+\infty} + b^*_{-\infty\leftarrow N}b_{N\leftarrow+\infty}}(\zeta^{-2})$$

was a Herglotz function on $\mathcal{D}^*$. In particular, $\mathrm{Re} f_N(\zeta)$ is equal a.e. to the absolutely continuous portion of the measure μ_N associated to this function, which means that

$$\int_{\mathbf{T}} \mathrm{Re} f_N \leq 1$$

[14] Actually, the picture is slightly more complicated than this. For instance, if $M[a,b]$ had a double pole at 1 instead of a single pole, what happens is that in addition to the two extreme solutions to the RHP with no singular spectrum, there is also an intermediate solution to the RHP where $M[a_-, b_-]$ and $M[a_+, b_+]$ each have a simple pole at 1, and again there is no singular spectrum. Then there are two line segments of solutions, each of which connects this intermediate solution to one of the two extreme solutions, and the solutions in these line segments each have an embedded eigenvalue at 1. We will investigate these phenomena more thorougly in a future paper.

with equality occurring if and only if this measure μ_N has purely absolutely continuous spectrum, which by Proposition 2.10 is equivalent to $L[F]$ having purely absolutely continuous spectrum. But by (2.61) the denominator of f_N is equal to a, while from (2.62) we have

$$\begin{aligned} a_{-\infty\leftarrow n} &= aa^*_{n\leftarrow+\infty} - b^* b_{n\leftarrow+\infty} \\ b_{-\infty\leftarrow n} &= ba^*_{n\leftarrow+\infty} - b_{n\leftarrow+\infty} a^* \end{aligned}$$

which after some algebra becomes

$$\begin{aligned} a_{n\leftarrow+\infty} &= \frac{a^*_{-\infty\leftarrow n}}{a^*} + \frac{bb^*_{n\leftarrow+\infty}}{a^*} \\ b_{-\infty\leftarrow n} &= \frac{ba_{-\infty\leftarrow n}}{a} - \frac{b_{n\leftarrow+\infty}}{a}. \end{aligned}$$

Substituting this into the definition of f_N, we obtain after some more algebra

$$\mathrm{Re} f_N(\zeta) = \frac{|a_{-\infty\leftarrow N}|^2 + |b_{N\leftarrow+\infty}|^2}{|a|^2}(\zeta^{-2});$$

this is also equal to $\frac{|b_{-\infty\leftarrow N}|^2+|a_{N\leftarrow+\infty}|^2}{|a|^2}(\zeta^{-2})$ since

$$|a_{-\infty\leftarrow N}|^2 - |b_{-\infty\leftarrow N}|^2 = |a_{N\leftarrow+\infty}|^2 - |b_{N\leftarrow+\infty}|^2 = 1.$$

The claim follows. $\square$

2.9. A flag of Hilbert spaces

Let F be an admissible potential. In the last section we determined a fair amount about the structure of the Dirac operator $L = L[F]$ on $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$. For instance, we isolated a subspace $(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{ac}$ of $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ which was invariant under $L[F]$, $*$, and σ, and such that the system $(L[F], *, \sigma, (L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{ac})$ was unitarily equivalent (via either of the two adjoint wave operators $\Omega_{\pm\infty\leftarrow 0}$, which are related by (2.59)) to the system $(L[0], *, \sigma, L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))$. In particular $\mathbf{H}_{ac}$ contains a number of interesting vectors such as $\Omega_{0\leftarrow\pm\infty}(v_n)$ and $\Omega_{0\leftarrow\pm\infty}(w_n)$.

We now see how the original orthonormal basis vectors v_n, w_n relate to these scattering basis vectors $\Omega_{0\leftarrow\pm\infty}(v_n)$, $\Omega_{0\leftarrow\pm\infty}(w_n)$, or more precisely how various Hilbert spaces generated by one space correspond to the other. In the vacuum case $F = 0$, $\Omega_{0\leftarrow\pm\infty}$ is the identity and so these basis vectors co-incide, but of course for non-zero F these vectors will be different.

For any integer N, define the vector space $V_{<N} \subset L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ to be the Hilbert space spanned by the orthonormal basis vectors $\{v_n : n < N\} \cup \{w_n : n < N\}$. Similarly define $V_{\geq N}$ to be the Hilbert space spanned by $\{v_n : n \geq N\} \cup \{w_n : n \geq N\}$. Clearly these spaces are orthogonal complements of each other: $V_{\geq N} = V_{<N}^\perp$.

Now let $V^{min}_{<N}$ be the smallest Hilbert space which contains the vectors

$$\{\Omega_{0\leftarrow+\infty}(w_n) : n < N\} \cup \{\Omega_{0\leftarrow-\infty}(v_n) : n < N\}.$$

Note that the set of vectors $\{\Omega_{0\leftarrow+\infty}(w_n) : n < N\}$ and $\{\Omega_{0\leftarrow-\infty}(v_n) : n < N\}$ are separately orthonormal, but their union is not necessarily so. Note that this space necessarily lies in $(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{ac}$, thanks to Proposition 2.10. Similarly define $V^{min}_{\geq N} \subset (L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{ac}$ to be the smallest Hilbert space which contains the vectors

$$\{\Omega_{0\leftarrow+\infty}(v_n) : n \geq N\} \cup \{\Omega_{0\leftarrow-\infty}(w_n) : n \geq N\}.$$

Then define $V_{<N}^{max} := (V_{\geq N}^{min})^{\perp}$ and $V_{\geq N}^{max} := (V_{<N}^{min})^{\perp}$; these spaces necessarily contain $(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{sc} \oplus (L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))_{pp}$.

LEMMA 2.13. *For any N, we have*

$$V_{<N}^{min} \subseteq V_{<N} \subseteq V_{<N}^{max}$$

or equivalently that

$$V_{\geq N}^{min} \subseteq V_{\geq N} \subseteq V_{\geq N}^{max}.$$

PROOF. It suffices to prove that $V_{<N}^{min} \subseteq V_{<N}$ and $V_{\geq N}^{min} \subseteq V_{\geq N}$, as the other two inclusions follow by taking orthogonal complements. We shall just prove the first, as the second is similar. It suffices to show that $\Omega_{0\leftarrow+\infty}(w_n), \Omega_{0\leftarrow-\infty}(v_n) \in V_{<N}$ for all $n < N$. By the definition of $\Omega_{0\leftarrow\pm\infty}$, it in fact suffices to show that $L[F]^{-m}w_{n-m}, L[F]^m v_{n-m} \in V_{<N}$ for all $n < N$ and $m > 0$. But this follows from the finite speed of propagation property in Lemma 2.9, and the definition of $V_{<N}$. The corresponding claim that $V_{\geq N}^{min} \subseteq V_{\geq N}$ is proven similarly. □

In particular, if $V_{<N}^{min} = V_{<N}^{max}$, then $V_{<N}$ is completely determined from the scattering data $\Omega_{0\leftarrow\pm\infty}(v_n)$, $\Omega_{0\leftarrow\pm\infty}(v_n)$; later we will show that this is a necessary and sufficient condition for $\widehat{F}$ to have unique inverse NLFT.

We also observe the obvious invariances $*V_{<N} = \sigma V_{<N} = V_{<N}$ and $*V_{\geq N} = \sigma V_{\geq N} = V_{\geq N}$. Clearly the spaces $V_{<N}$ are increasing in N, and $V_{\geq N}$ is decreasing in N, with

$$\bigcap_N V_{<N} = \bigcap_N V_{\geq N} = \{0\} \text{ and } \sum_N V_{<N} = \sum_N V_{\geq N} = L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$$

(here we use $\sum_\alpha V_\alpha$ to denote the smallest Hilbert space containing all of the V_α; it is conjugate to $\bigcup$ under orthogonal complement). Also, while $V_{<N}$ is not quite invariant under L, we do have $LV_{<N} \subset V_{<N+1}$, as can be seen directly from (2.4). We also observe the following more precise inclusions.

LEMMA 2.14. *For any N, we have the formulae*

$$V_{<N+1} = V_{<N} + \mathbf{C}\Omega_{0\leftarrow+\infty}(w_N) + \mathbf{C}\Omega_{0\leftarrow-\infty}(v_N)$$
$$V_{<N} = V_{<N+1} \cap (\mathbf{C}\Omega_{0\leftarrow-\infty}(w_N))^{\perp} \cap (\mathbf{C}\Omega_{0\leftarrow+\infty}(v_N))^{\perp}.$$

PROOF. We just prove the first claim, as the second is similar. Since $V_{<N+1}$ contains $V_{<N+1}^{min}$, which in turn contains $\Omega_{0\leftarrow+\infty}(w_N)$ and $\Omega_{0\leftarrow-\infty}(v_N)$, we see that the left-hand side certainly contains the right-hand side. To show the other containment, observe that $V_{<N}$ has codimension two inside $V_{<N+1}$, so it will suffice to show that the vectors $\Omega_{0\leftarrow+\infty}(w_N)$ and $\Omega_{0\leftarrow-\infty}(v_N)$ are linearly independent modulo $V_{<N}$. Actually since these two vectors lie in different eigenspaces of the parity operator σ, and $V_{<N}$ is σ-invariant, it suffices to show that these vectors lie outside of $V_{<N}$. We shall just do this for $\Omega_{0\leftarrow+\infty}(w_N)$ as the other is similar (and follows from conjugation-invariance). Since $V_{<N}$ is contained in $V_{<N}^{max}$, it will suffice to show that $\Omega_{0\leftarrow+\infty}(w_N)$ is not orthogonal to $\Omega_{0\leftarrow-\infty}(w_N)$. But by (2.59) we have

$$\langle \Omega_{0\leftarrow+\infty}(w_N), \Omega_{0\leftarrow-\infty}(w_N)\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \langle w_N, \Omega_{+\infty\leftarrow-\infty}(w_N)\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})};$$

applying (2.58) we see that this is equal to

$$\int_{\mathbf{T}} \frac{1}{a(\zeta^{-2})} = \frac{1}{a(0)}$$

which is non-zero as desired. □

We now show a sort of converse to the above results, in that any family of Hilbert spaces $V_{<N}$ in an abstract Hilbert space $\mathbf{H}$ (equipped with the operators L, $*$, σ) which obey the above properties generate a Dirac operator with specified scattering data.

THEOREM 2.15. *Let $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ be a scattering datum, and let $\mathbf{H}$ be a Hilbert space with a unitary operator $L : \mathbf{H} \to \mathbf{H}$, a skew-unitary involution $* : \mathbf{H} \to \mathbf{H}$, and a unitary involution $\sigma : \mathbf{H} \to \mathbf{H}$ such that we have the commutation relations*

$$*L* = L^{-1}, L\sigma = -\sigma L \text{ and } *\sigma = -\sigma * . \tag{2.75}$$

Suppose also that we have unitary operators $\Omega_{0\leftarrow\pm\infty} : L^2(\mathbf{T}) \oplus L^2(\mathbf{T}) \to \mathbf{H}_{ac}$ for some closed subspace $\mathbf{H}_{ac}$ of $\mathbf{H}$, with the commutation properties

$$\Omega_{0\leftarrow\pm\infty}L[0] = L\Omega_{0\leftarrow\pm\infty}; \quad \Omega_{0\leftarrow\pm\infty}* = *\Omega_{\mp}; \quad \Omega_{0\leftarrow\pm\infty}\sigma = \sigma\Omega_{0\leftarrow\pm\infty}, \tag{2.76}$$

and such that the scattering operator $\Omega_{+\infty\leftarrow-\infty} := \Omega^*_{0\leftarrow+\infty}\Omega_{0\leftarrow-\infty}$ is given by (2.58). Suppose we also have a closed subspace $V_{<N}$ for every integer N such that

$$\begin{aligned}
V_{<N+1} &= V_{<N} + \mathbf{C}\Omega_{0\leftarrow+\infty}(w_N) + \mathbf{C}\Omega_{0\leftarrow-\infty}(v_N)\\
V_{<N} &= V_{<N+1} \cap (\mathbf{C}\Omega_{0\leftarrow-\infty}(w_N))^\perp \cap (\mathbf{C}\Omega_{0\leftarrow+\infty}(v_N))^\perp\\
*V_{<N} &= \sigma V_{<N} = V_{<N}\\
LV_{<N} &\subseteq V_{<N+1}\\
\bigcap_N V_{<N} &= 0\\
\sum_N V_{<N} &= \mathbf{H}.
\end{aligned} \tag{2.77}$$

Then, up to a unitary transformation of $\mathbf{H}$, the space $\mathbf{H}$ is equal to $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, and L is equal to $L[F]$ for some potential F such that $\widehat{F} = M[a,b]$. Furthermore, the operators $*$, σ, $\Omega_{0\leftarrow\pm\infty}$, $\alpha_{\pm\infty\leftarrow0}$, $\beta_{\pm\infty\leftarrow0}$, and spaces $V_{<N}$ correspond to the operators and spaces with the same name defined previously.

PROOF. Define the operators $\Omega_{\pm\infty\leftarrow0} : \mathbf{H} \to L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ to be the adjoints of $\Omega_{0\leftarrow\pm\infty}$, and define $\alpha_{\pm\infty\leftarrow0}$, $\beta_{\pm\infty\leftarrow0}$ by (2.48). Observe from (2.76) and (2.58) that these functions the identities (2.50), (2.51), (2.52), (2.56), (2.57), (2.49).

We first use the parity operator σ to decompose all our spaces into two pieces. Since σ is a unitary involution, we have $\sigma^2 = 1$, and so the projections $P_+ := \frac{1+\sigma}{2}$ and $P_- := \frac{1-\sigma}{2}$ are complementary orthogonal projections. Since $V_{<N}$ is σ-invariant, we have the decomposition $V_{<N} = P_+V_{<N} \oplus P_-V_{<N}$. Also, from (2.76) we know that $\Omega_{0\leftarrow\pm\infty}(v_N)$ lies in the range of P_+ if N is even and P_- if N is odd, and vice versa for $\Omega_{0\leftarrow\pm\infty}(w_N)$. Thus we have

$$P_{(-1)^N}V_{<N+1} = P_{(-1)^N}V_{<N} + \mathbf{C}\Omega_{0\leftarrow-\infty}(v_N); \quad P_{(-1)^{N+1}}V_{<N+1} = P_{(-1)^{N+1}}V_{<N} + \mathbf{C}\Omega_{0\leftarrow+\infty}(w_N) \tag{2.78}$$

and

$$P_{(-1)^N}V_{<N} = P_{(-1)^N}V_{<N+1} \cap (\mathbf{C}\Omega_{0\leftarrow+\infty}(v_N))^\perp; \quad P_{(-1)^{N+1}}V_{<N} = P_{(-1)^{N+1}}V_{<N+1} \cap (\mathbf{C}\Omega_{0\leftarrow-\infty}(w_N))^\perp. \tag{2.79}$$

We now do a "Gram-Schmidt" procedure to extract an orthonormal basis from these two nested sequences of subspaces. First, to avoid degeneracy, we must show that the inner product between the unit vectors $\Omega_{0\leftarrow-\infty}(v_N)$ and $\Omega_{0\leftarrow+\infty}(v_N)$ is non-zero. Using (2.59), we have

$$\langle \Omega_{0\leftarrow-\infty}(v_N), \Omega_{0\leftarrow+\infty}(v_N)\rangle_{\mathbf{H}} = \langle \Omega_{+\infty\leftarrow-\infty}(v_N), v_N\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}.$$

Applying (2.58) we thus see that

$$\langle \Omega_{0\leftarrow-\infty}(v_N), \Omega_{0\leftarrow+\infty}(v_N)\rangle_{\mathbf{H}} = \int_{\mathbf{T}} \frac{1}{a(\zeta^{-2})} = \frac{1}{a(0)}.$$

In particular, the vectors $\langle \Omega_{0\leftarrow-\infty}(v_N), \Omega_{0\leftarrow+\infty}(v_N)\rangle_{\mathbf{H}}$ are not orthogonal. A similar argument (using (2.76) if desired) gives that

$$\langle \Omega_{0\leftarrow-\infty}(w_N), \Omega_{0\leftarrow+\infty}(w_N)\rangle_{\mathbf{H}} = \frac{1}{a(0)} > 0.$$

In particular, from this and (2.78), (2.79) we see that $P_\pm V_{<N}$ is a codimension one subspace inside $P_\pm V_{<N+1}$. Thus the orthogonal complement of $P_\pm V_{<N}$ in $P_\pm V_{<N+1}$ is a complex line. If $\pm = (-1)^N$, then this orthogonal complement is not orthogonal to $\Omega_{0\leftarrow-\infty}(v_N)$, and so we can define a unique unit vector v'_N in this complement such that $\langle v'_N, \Omega_{0\leftarrow-\infty}(v_N)\rangle$ is real and positive, while if $\pm = (-1)^{N+1}$ then the orthogonal complement is not orthogonal to $\Omega_{0\leftarrow+\infty}(w_N)$, and so we can define w'_N to be the unique unit vector in this complement such that $\langle w'_N, \Omega_{0\leftarrow+\infty}(w_N)\rangle$ is real. Note that by construction we have $\sigma v'_N = (-1)^N v'_N$ and $\sigma w'_N = (-1)^{N+1} w'_N$. Also since $*$ must map $P_\pm V_N$ to $P_\mp V_N$ (by the $*$-invariance of V_N and (2.75)) we see that we must have $*v'_N = w'_N$. Thus the parity and conjugation operator act on v'_N and w'_N just like they do on v_N and w_N in $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$.

The unit vectors v'_N and w'_N are orthogonal to each other since they lie in different eigenspaces of σ. By construction, we see that

$$V_{<N+1} = V_{<N} \oplus \mathbf{C}v'_N \oplus \mathbf{C}w'_N.$$

Iterating this we see that

$$V_{<N_2} = V_{<N_1} \oplus \bigoplus_{N_1 \leq N < N_2} \mathbf{C}v'_N \oplus \mathbf{C}w'_N$$

for all $N_1 < N_2$. Sending N_2 to $+\infty$ and N_1 to $-\infty$ and using the hypotheses in (2.77), we thus see that

$$\mathbf{H} = \bigoplus_{N\in\mathbf{Z}} \mathbf{C}v'_N \oplus \mathbf{C}w'_N$$

or in other words that $\{v'_N, w'_N : N \in \mathbf{Z}\}$ is an orthonormal basis of $\mathbf{H}$. Thus we may apply a unitary transformation if necessary to replace $\mathbf{H}$ with $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$, and v'_N, w'_N with v_N, w_N; henceforth we shall do so, and erase the distinction between v'_N and v_N, between w'_N and w_N, and between $\mathbf{H}$ and $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$. Note that by the above discussion we have

$$V_{<N} = \bigoplus_{N'<N} \mathbf{C}v_{N'} \oplus \mathbf{C}w_{N'} \tag{2.80}$$

and

$$V_{<N}^\perp = \bigoplus_{N'\geq N} \mathbf{C}v_{N'} \oplus \mathbf{C}w_{N'}. \tag{2.81}$$

Now let us investigate the action of L on the basis vectors v_N, w_N. We begin with computing[15] Lw_N. From the hypothesis $LV_{<N+1} \subseteq V_{<N+2}$ we see that $Lw_N \in V_{<N+2}$. But we also know that w_N lies in $V_{<N+1}$, which is orthogonal to $\Omega_{0\leftarrow-\infty}(w_{N+2})$ by (2.77). Thus Lw_N is orthogonal to $L\Omega_{0\leftarrow-\infty}(w_{N+2})$, which equals $\Omega_{0\leftarrow-\infty}(w_{N+1})$ by (2.76). But Lw_N is also orthogonal to $\Omega_{0\leftarrow+\infty}(v_{N+1})$ by parity considerations, since by (2.75) we see that Lw_N lies in a different eigenspace of σ than v_{N+1}. Thus Lw_N actually lies in $V_{<N+2}\cap(\mathbf{C}\Omega_{0\leftarrow-\infty}(w_N))^\perp\cap(\mathbf{C}\Omega_{0\leftarrow+\infty}(v_N))^\perp = V_{<N+1}$. On the other hand, we know that w_N is orthogonal to $V_{<N}$, hence Lw_N is orthogonal to $LV_{<N-1}$. To understand this space, first observe from (2.77) that $LV_{<N-2}$ lies in $V_{<N-1}$, hence (by applying $*$ and (2.75)) $L^{-1}V_{<N-2}$ also lies in $V_{<N-1}$. Hence $LV_{<N-1}$ contains $V_{<N-2}$, and so Lw_N is orthogonal to $V_{<N-2}$.

To summarize, we have shown that Lw_N is contained in $V_{<N+1}$ and is orthogonal to $V_{<N-2}$. From (2.80), (2.81) that Lw_N must be a linear combination of v_N, w_N, v_{N-1}, and w_{N-1}. But since $L\sigma = -\sigma L$, Lw_N must have the opposite parity to w_N, and thus Lw_N must be a linear combination of v_N and w_{N-1}.

Since w_N has a positive inner product with $\Omega_{0\leftarrow+\infty}(w_N)$ by construction, Lw_N has a positive inner product with $L\Omega_{0\leftarrow+\infty}(w_N) = \Omega_{0\leftarrow+\infty}(w_{N-1})$. And w_{N-1} also has positive inner product with $\Omega_{0\leftarrow-\infty}(w_{N-1})$ by construction. Meanwhile, v_N is orthogonal to all of $V_{<N}$, and in particular to $\Omega_{0\leftarrow+\infty}(w_{N-1})$, which is contained in $V_{<N}$ by (2.77). Since Lw_N was a combination of w_{N-1} and v_N, we thus see that the inner product of Lw_N with w_{N-1} is strictly positive. Since Lw_N, w_{N-1}, and v_N are all unit vectors, we may therefore find a complex number $|F_{N-1}| < 1$ such that

$$Lw_N = -F^*_{N-1}v_N + \sqrt{1-|F_{N-1}|^2}w_{N-1},$$

or upon incrementing N,

$$Lw_{N+1} = -F^*_N v_{N+1} + \sqrt{1-|F_N|^2}w_N. \tag{2.82}$$

Applying $*$ and (2.75) we obtain

$$L^{-1}v_{N+1} = -F_N w_N + \sqrt{1-|F_N|^2}v_N;$$

applying L to both sides and using (2.82), we obtain (after some algebra)

$$Lv_N = \sqrt{1-|F_N|^2}v_{N+1} + F_N w_N.$$

In other words, $L = L[F]$ is given by (2.4). We now show that F is admissible, i.e. that $\prod_N \sqrt{1-|F_N|^2}$ is non-zero.

An inspection of the proof of Lemma 2.9 reveals that it does not use that F is admissible (basically because one only applies L a finite number of times). In particular, we see that

$$\langle L[F]^m v_n, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \prod_{n\le n'<n+m} \sqrt{1-|F_{n'}|^2} \tag{2.83}$$

for any $n \in \mathbf{Z}$ and $m > 0$. Thus it will suffice to prove that these inner products are bounded away from zero; in fact we will show that

$$\langle L[F]^m v_n, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} \ge \frac{1}{a(0)}. \tag{2.84}$$

[15] Readers familiar with Jacobi matrices may recognize the following argument as essentially the same as the one which shows that the operation of multiplication by z on $L^2(\mu)$ is given by a Jacobi matrix in the basis of orthogonal polynomials; the main new feature of the Dirac operator is that it introduces parity considerations.

To prove this, first recall that v_{n+m} lies in the orthogonal complement of $P_{(-1)^{n+m}} V_{<n+m}$ in $P_{(-1)^{n+m}} V_{<n+m+1}$. Meanwhile, the vector $\Omega_{0\leftarrow+\infty}(v_{n+m})$ is also orthogonal to $P_{(-1)^{n+m}} V_{<n+m}$, but has a non-zero inner product with v_{n+m} (since $v_{n+m} \notin V_{<n+m}$). Thus we have

(2.85)

$$\langle v, \Omega_{0\leftarrow+\infty}(v_{n+m})\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \langle v, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} \langle v_{n+m}, \Omega_{0\leftarrow+\infty}(v_{n+m})\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}$$

for all $v \in P_{(-1)^{n+m}} V_{<n+m+1}$. Applying this in particular to $v = \Omega_{0\leftarrow-\infty}(v_{n+m})$, we obtain

$$\frac{1}{a(0)} = \langle \Omega_{0\leftarrow-\infty}(v_{n+m}), v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} \langle v_{n+m}, \Omega_{0\leftarrow+\infty}(v_{n+m})\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}.$$

In particular, since the former inner product was positive real by hypothesis, so is the latter, and since all vectors are unit vectors we thus have

$$\frac{1}{a(0)} \leq \langle \Omega_{0\leftarrow\pm\infty}(v_{n+m}), v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} \leq 1. \tag{2.86}$$

Now consider $L[F]^m v_n$. From Lemma 2.9 we see that $L[F]^m v_n$ can be written as a linear combination of vectors $\{v_{n'}, w_{n'} : n' \leq n+m\}$ and hence lies in $V_{\leq n+m}$. By parity we see that it in fact lies in $P_{(-1)^{n+m}} V_{<n+m}$. In particular by (2.85) we have

$$\begin{aligned}\langle L[F]^m v_n, \Omega_{0\leftarrow+\infty}(v_{n+m})\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} =& \langle L[F]^m v, v_{n+m}\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} \\ & \langle v_{n+m}, \Omega_{0\leftarrow+\infty}(v_{n+m})\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}.\end{aligned}$$

The left-hand side is equal to $\langle v_n, \Omega_{0\leftarrow+\infty}(v_n)\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}$, which by (2.86) is at least $1/a(0)$, and the claim (2.84) follows. Thus F is admissible. In particular, we can define the scattering maps $\tilde{\Omega}_{0\leftarrow\pm\infty} : L^2(\mathbf{T}) \oplus L^2(\mathbf{T}) \to L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$ by

$$\tilde{\Omega}_{0\leftarrow\pm\infty} v := \lim_{m\to\pm\infty} L[F]^{-m} L[0]^m v.$$

We do not know yet whether these scattering maps $\tilde{\Omega}_{0\leftarrow\pm\infty}$ co-incide with the map $\Omega_{0\leftarrow\pm\infty}$ given to us by hypothesis, but if they did we would be able to deduce from (2.56), (2.57) (which holds for $\Omega_{0\leftarrow\pm\infty}$ and $M[a,b]$ by hypothesis, and holds for $\tilde{\Omega}_{0\leftarrow\pm\infty}$ and $\widehat{F}$ by previous discussion) that $\widehat{F} = M[a,b]$ as desired.

It remains to show that the partial isometries $\tilde{\Omega}_{0\leftarrow\pm\infty}$ and $\Omega_{0\leftarrow\pm\infty}$ coincide. By (2.76) it suffices to do so for v_0. We first show that $\Omega_{0\leftarrow\pm\infty}(v_0)$ lies in $\tilde{\Omega}_{0\leftarrow\pm\infty}(L^2(\mathbf{T}) \oplus L^2(\mathbf{T}))$, i.e. that

$$\lim_{m\to\pm\infty} L[F]^{-m} L[0]^m v = \Omega_{0\leftarrow\pm\infty}(v_0) \tag{2.87}$$

for some vector v (ideally we would have $v = v_0$, but we will not show this yet). Equivalently, we wish to show that the sequence $L[0]^{-m} L[F]^m \Omega_{0\leftarrow\pm\infty}(v_0)$ converges; by (2.76) we may rewrite this sequence as $L[0]^{-m} \Omega_{0\leftarrow\pm\infty}(v_m)$.

We shall just do this for the $\pm = +$ case; the $\pm = -$ case is similar, and consists basically of replacing all the spaces below by their orthogonal complements and changing the sign of m.

We know that $\Omega_{0\leftarrow+\infty}(v_m)$ is orthogonal to $V_{<m}$, hence $L[0]^{-m}\Omega_{0\leftarrow+\infty}(v_m)$ is orthogonal to $V_{<0}$. Thus we may write

$$L[0]^{-m}\Omega_{0\leftarrow+\infty}(v_m) = \sum_{n\geq 0} c_{n,m} v_n + d_{n,m} w_n$$

for some coefficients $c_{n,m}$, $d_{n,m}$; since v_m is a unit vector we have

$$\sum_{n\geq 0} |c_{n,m}|^2 + |d_{n,m}|^2 = 1.$$

Now observe that

$$c_{n,m} = \langle L[0]^{-m} L[F]^m \Omega_{0\leftarrow+\infty}(v_0), v_n\rangle = \langle \Omega_{0\leftarrow+\infty}(v_0), L[F]^{-m} L[0]^m v_n\rangle$$

and hence $c_{n,m}$ converges pointwise to $c_n := \langle \Omega_{0\leftarrow+\infty}(v_0), \tilde\Omega_{0\leftarrow+\infty}(v_n)\rangle$ as $m \to +\infty$. Similarly $d_{n,m}$ converges to $d_n := \langle \Omega_{0\leftarrow+\infty}(v_0), \tilde\Omega_{0\leftarrow+\infty}(w_n)\rangle$. If we can upgrade this pointwise convergence to l^2 convergence then we will have established the claim (2.87). To show this, pick any $\varepsilon > 0$, and choose N so large that

$$\sum_{n\geq N} |c_{n,0}|^2 + |d_{n,0}|^2 \leq \varepsilon^2.$$

Then pick $m \gg N$ so large that

$$\sum_{n\geq m-N} |F_n|^2 \leq \delta^2 \tag{2.88}$$

where $\delta = \delta(\varepsilon, N) > 0$ is a small number to be chosen later. We can then split $\Omega_{0\leftarrow+\infty}(v_0) = A + B$, where A is linear combination of the finite set of vectors $\{v_n, w_n : 0 \leq n < N\}$ and B has norm at most ε. By finite speed of propagation we then see that $L[F]^m A$ is $V_{<N+m}$. But $L[F]^m \Omega_{0\leftarrow+\infty}(v_0)$ is orthogonal to $V_{<N}$, thus $L[F]^m \Omega_{0\leftarrow+\infty}(v_0)$ is equal to a linear combination of $\{v_n, w_n : m \leq n < N + m\}$ plus an error of norm at most ε. We now dispose of the w_n terms. We write $L[F]^m A = L[0]^N L[F]^{m-N} A + (L[F]^N - L[0]^N) L[F]^{m-N} A$. From finite speed of propagation we know that $L[F]^{m-N} A \in V_{<N}$, hence $L[0]^N L[F]^{m-N} A$ is orthogonal to $\{w_n : m \leq n \leq N + m\}$. The second term $(L[F]^N - L[0]^N) L[F]^{m-N} A$ can have coefficients in the set $\{w_n : N \leq n \leq N + m\}$, but the size of these coefficients is at most $C(N)\delta$ thanks to (2.88). Thus by making δ sufficiently small, we can ensure that $L[F]^m \Omega_{0\leftarrow+\infty}(v_0)$ is equal to a linear combination of $\{v_n : m \leq n \leq N+m\}$ plus an error of norm at most 2ε. Thus $L[0]^m L[F]^{-m} \Omega_{0\leftarrow+\infty}(v_0)$ is a linear combination of $\{v_n : 0 \leq n \leq N\}$ plus an error of at most 2ε. This is uniform in m and thus can be used (together with the pointwise convergence of coefficients) to show the convergence of $L[0]^m L[F]^{-m} \Omega_{0\leftarrow+\infty}(v_0)$. This proves (2.87) for $\pm = +$. Notice also that the above argument shows that the $d_{n,m}$ are converging to 0 as $m \to +\infty$, so $d_n = 0$ and thus

$$\langle \Omega_{0\leftarrow\pm\infty}(v_0), \tilde\Omega_{0\leftarrow+\infty}(w_n)\rangle = 0. \tag{2.89}$$

We still have to prove that $\Omega_{0\leftarrow+\infty}(v_0) = \tilde\Omega_{0\leftarrow+\infty}(v_0)$. Since we now know that $\Omega_{0\leftarrow+\infty}(v_0)$ lies in the range of $\tilde\Omega_{0\leftarrow+\infty}$ (i.e. it is in the absolutely continuous portion of $L^2(\mathbf{T}) \oplus L^2(\mathbf{T})$) it will suffice to show that $\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 1$ and $\tilde\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 0$, where $\tilde\alpha_{\pm\infty\leftarrow 0}$ and $\tilde\beta_{\pm\infty\leftarrow 0}$ are defined as in (2.48) but with $\tilde\Omega_{0\leftarrow\pm\infty}$ instead of $\Omega_{0\leftarrow\pm\infty}$.

The claim $\tilde\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 0$ follows from (2.89), so now we turn to showing $\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 1$. First recall that $\Omega_{0\leftarrow+\infty}(v_m) = L[F]^m \Omega_{0\leftarrow+\infty}(v_0)$ is orthogonal to v_{m-r} for any $r > 0$ and $m \in \mathbf{Z}$ (since the latter vector lies in $V_{<m}$), hence $L[0]^{-m} L[F]^m \Omega_{0\leftarrow+\infty}(v_0)$ is orthogonal to v_{-r}. Taking limits as $m \to \pm\infty$ we see that $\langle \Omega_{0\leftarrow+\infty}(v_0), \tilde\Omega_{0\leftarrow\pm\infty}(v_{-r})\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = 0$ for all $r < 0$. In particular,

from (2.48) we thus see that $\tilde\alpha_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))$ lies in $H^2(\mathcal{D})$ for both choices of sign $\pm$. By (2.76) and (2.50) we have
(2.90)
$$\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_n)) = \zeta^n\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)); \quad \tilde\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_n)) = 0.$$
In particular we have
$$\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(\sum_n c_n v_n)) = \sum_n c_n\zeta^n\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))$$
$$\tilde\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(\sum_n c_n v_n)) = 0$$
for any compactly supported sequence c_n. Applying (2.49), the fact that $\Omega_{0\leftarrow+\infty}$ is an isometry, and that v_n is orthonormal, we thus have
$$(\sum_n |c_n|^2)^{1/2} = \|(\sum_n c_n\zeta^n)\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))\|_{L^2(\mathbf{T})}.$$
By Plancherel and a limiting argument we thus have
$$\|f\|_{L^2(\mathbf{T})} = \|f\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))\|_{L^2(\mathbf{T})}$$
for all $f \in L^2(\mathbf{T})$; this implies that $|\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))| = 1$. Since we already know that $\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) \in H^2(\mathcal{D})$, this implies that $\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))$ is an inner function on $\mathcal{D}$. A similar argument shows that $\tilde\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_0))$ is an inner function on $\mathcal{D}^*$.

Write $M[\tilde a, \tilde b] = \overbrace{F}$. From (2.60) and the claim $\tilde\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 0$ already shown, we have
$$\tilde\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))(\zeta) = \frac{1}{\tilde a^*(\zeta^{-2})}\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))(\zeta).$$
Meanwhile, for any integer n we have
$$\tilde\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_n)) = \zeta^n\tilde\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_0)); \quad \tilde\beta_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_n)) = 0.$$
by the analogue of (2.90) for $\Omega_{0\leftarrow-\infty}(w_n)$. Thus by (2.49) we have
$$\langle\Omega_{0\leftarrow+\infty}(v_0), \Omega_{0\leftarrow-\infty}(v_n)\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \int_{\mathbf{T}} \frac{1}{\tilde a^*(\zeta^{-2})}\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))(\zeta)\zeta^{-n}\tilde\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_0))^*(\zeta).$$
On the other hand, from (2.49) we have
$$\begin{aligned}\langle\Omega_{0\leftarrow+\infty}(v_0), \Omega_{0\leftarrow-\infty}(v_n)\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \int_{\mathbf{T}} &\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_n))^* \\ &+ \beta_{-\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))\beta_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_n))^*.\end{aligned}$$
From (2.48) we see that $\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_n)) = \zeta^n$ and $\beta_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_n)) = 0$. Similarly $\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 1$ and $\beta_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 0$, so by (2.60) we have
$$\alpha_{-\infty\leftarrow 0}(\Omega(v_0))(\zeta) = \frac{1}{a^*(\zeta^{-2})}.$$
Thus we have
$$\langle\Omega_{0\leftarrow+\infty}(v_0), \Omega_{0\leftarrow-\infty}(v_n)\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \int_{\mathbf{T}} \frac{1}{a^*(\zeta^{-2})}\zeta^{-n}.$$
Thus we have
$$\int_{\mathbf{T}} \frac{1}{\tilde a^*(\zeta^{-2})}\tilde\alpha_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))(\zeta)\zeta^{-n}\tilde\alpha_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_0))^*(\zeta) = \int_{\mathbf{T}} \frac{1}{a^*(\zeta^{-2})}\zeta^{-n}$$

for all n. This forces

$$\frac{1}{\tilde{a}^*(\zeta^{-2})}\tilde{\alpha}_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))(\zeta)\zeta^{-n}\tilde{\alpha}_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_0))^*(\zeta) = \frac{1}{a^*(\zeta^{-2})}$$

for almost every $\zeta \in \mathbf{T}$. But $\frac{1}{a^*(\zeta^{-2})}$ and $\frac{1}{\tilde{a}^*(\zeta^{-2})}$ are outer on $\mathcal{D}$, while $\tilde{\alpha}_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))$ and $\tilde{\alpha}_{-\infty\leftarrow 0}(\Omega_{0\leftarrow-\infty}(v_0))^*$ are inner on $\mathcal{D}$. By the uniqueness of inner and outer factorizations, we thus see that $\tilde{\alpha}_{\pm\infty\leftarrow 0}(\Omega_{0\leftarrow\pm\infty}(v_0))$ are constant functions with unit magnitude. However, since $\Omega_{0\leftarrow+\infty}(v_0)$ has positive inner product with $L[F]^{-m}L[0]^m v_0 = L[F]^{-m}v_m$ for every integer m (e.g. by (2.86)), we see from (2.48) that the constant coefficient of $\tilde{\alpha}_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0))$ is real and non-negative. Thus have $\tilde{\alpha}_{+\infty\leftarrow 0}(\Omega_{0\leftarrow+\infty}(v_0)) = 1$ as desired. This completes the proof of Theorem 2.15. □

Fix $M[a,b] \in \mathcal{L}^2(\mathbf{T})$. We can apply this Theorem to invert the NLFT for this choice of scattering data; in fact we will now give two such inversions.

We first have to choose $\mathbf{H}$. Notice that the scattering map $\Omega_{+\infty\leftarrow-\infty}$ defined in (2.58) depends purely on the scattering datum $M[a,b]$ and is unitary on $L^2(\mathbf{T})\oplus L^2(\mathbf{T})$. We will set $\mathbf{H}$ to be the graph of this map, i.e. $\mathbf{H}$ is the space of all pairs

$$\mathbf{H} := \{(u_{-\infty}, u_{+\infty}) \in (L^2(\mathbf{T})\oplus L^2(\mathbf{T}))\times(L^2(\mathbf{T})\oplus L^2(\mathbf{T})) : u_{+\infty} = \Omega_{+\infty\leftarrow-\infty}u_{-\infty}\}$$

equipped with the norm

$$\|(u_{-\infty}, v_{+\infty})\|_{\mathbf{H}} := \|u_{-\infty}\|_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})} = \|u_{+\infty}\|_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})}.$$

We define operators L, $*$, σ on this space by

$$\begin{aligned}
L(u_{-\infty}, u_{+\infty}) &= (L[0]u_{-\infty}, L[0]u_{+\infty}) = \zeta(u_{-\infty}, u_{+\infty})\\
*(u_{-\infty}, u_{+\infty}) &= (*u_{+\infty}, *u_{-\infty})\\
\sigma(u_{-\infty}, u_{+\infty}) &= (\sigma u_{-\infty}, \sigma u_{+\infty});
\end{aligned}$$

One can easily verify that these operations map $\mathbf{H}$ to itself and also obey the commutation relations (2.75). One can then define the wave operators $\Omega_{0\leftarrow\pm\infty} : L^2(\mathbf{T}) \oplus L^2(\mathbf{T}) \to \mathbf{H}$ by

(2.91)

$$\Omega_{0\leftarrow+\infty}(u_{+\infty}) := (\Omega_{-\infty\leftarrow+\infty}u_{+\infty}, u_{+\infty}), \quad \Omega_{0\leftarrow-\infty}(u_{-\infty}) := (u_{-\infty}, \Omega_{+\infty\leftarrow-\infty}u_{-\infty});$$

the adjoint operators are given by

$$\Omega_{\pm\infty\leftarrow 0}(u_{-\infty}, u_{+\infty}) = u_{\pm\infty}.$$

Note that the wave operators are surjective here (so $\mathbf{H} = \mathbf{H}_{ac}$), and thus unitary. This will mean that any potential constructed via Theorem 2.15 using this space will have purely absolutely continuous spectrum. (One can also easily verify the converse to this statement, any solution to the inverse NLFT of $M[a,b]$ with purely absolutely continuous spectrum can be constructed using this space $\mathbf{H}$. This space corresponds to the spaces $L^2_{s_\pm}$ used by Yuditskii and Volberg [**34**] in the Jacobi matrix case.)

We now have a choice as to how to define the spaces $V_{<N}$. As before we define

$$V_{<N}^{min} := \sum_{n<N} \mathbf{C}\Omega_{0\leftarrow+\infty}(w_n) + \mathbf{C}\Omega_{0\leftarrow-\infty}(v_n)$$

and

$$V_{<N}^{max} := \bigcap_{n\geq N} (\mathbf{C}\Omega_{0\leftarrow-\infty}(w_n))^\perp + (\mathbf{C}\Omega_{0\leftarrow+\infty}(v_n))^\perp.$$

Observe that $V_{<N}^{min} \subseteq V_{<N}^{max}$ for all N; this comes directly from the definitions, noting that a is outer on $\mathcal{D}$ and hence $\int_{\mathbf{T}} \frac{1}{a^*(\zeta^{-2})}\zeta^n = 0$ for all $n > 0$.

As noted earlier, we must have $V_{<N}$ lie between $V_{<N}^{min}$ and $V_{<N}^{max}$. Suppose we choose $V_{<N}$ to equal $V_{<N}^{min}$; we claim that this choice obeys all the properties in (2.77). The first and third claims are clear, while the fourth follows from (2.76). To prove the second, we observe from the inclusion $V_{<N}^{min} \subseteq V_{<N}^{max}$ that

$$V_{<N} \subseteq V_{<N+1} \cap (\mathbf{C}\Omega_{0\leftarrow -\infty}(w_N))^{\perp} \cap (\mathbf{C}\Omega_{0\leftarrow -\infty}(v_N))^{\perp}.$$

But from the first property of (2.77) we see that $V_{<N}$ has codimension at most two in $V_{<N+1}$. But since $\Omega_{0\leftarrow +\infty}(w_N)$ has non-zero inner product with $\Omega_{0\leftarrow -\infty}(w_N)$ (indeed, by (2.58) the inner product is $1/a(0)$) and is orthogonal to $\Omega_{0\leftarrow +\infty}(v_n)$, and vice versa for $\Omega_{0\leftarrow -\infty}(v_N)$, we see that the codimension is exactly two, and the above inclusion must be equality. To prove the fifth claim of (2.77), it suffices to prove the stronger statement that

$$\bigcap_N V_{<N}^{max} = \{0\}.$$

To see this, suppose we have a vector $(u_{-\infty}, u_{+\infty}) \in \mathbf{H}$ which lies in $\bigcap_N V_{<N}^{max}$, then by construction of $V_{<N}^{max}$ we thus see that $(u_{-\infty}, u_{+\infty})$ is orthogonal to $\Omega_{0\leftarrow -\infty}(w_N)$ and $\Omega_{0\leftarrow +\infty}(v_N)$ for every N. Thus $u_{-\infty}$ is orthogonal to every w_N, and $u_{+\infty}$ is orthogonal to every v_N. If one then writes

$$u_{\pm\infty} = \begin{pmatrix} \alpha_{\pm\infty} \\ \beta_{\pm\infty} \end{pmatrix}$$

then by (2.58) we see that $\beta_{-\infty} = \alpha_{+\infty} = 0$, and thus from (2.58) (or (2.56), (2.57))

$$0 = \frac{1}{a(\zeta^{-2})}\alpha_{-\infty}(\zeta)$$
$$\beta_{+\infty}(\zeta) = \frac{\zeta^{-1}b(\zeta^{-2})}{a(\zeta^{-2})}\alpha_{-\infty}(\zeta).$$

Since a is a.e. finite on $\mathbf{T}$, we thus see that $(v_{-\infty}, v_{+\infty})$ vanishes, as desired. This proves the fifth property of (2.77). Finally, the sixth property of (2.77) is a dual version of the (stronger version) of the fifth.

We can then invoke Theorem 2.15 to create an admissible potential $F := F^{right}[a,b]$ such that $\overbrace{F^{right}[a,b]} = M[a,b]$. We call this F the *rightmost* inverse NLFT of $M[a,b]$; as we shall see, it has the largest amount of energy on right half-lines amongst all the inverse nonlinear Fourier transforms of $M[a,b]$. As remarked eralier, $L[F^{right}[a,b]]$ has purely absolutely continuous spectrum on $\mathbf{T}$.

A similar argument (which we omit) shows that the choice $V_{<N} := V_{<N}^{max}$ also obeys all the properties in (2.77), and thus gives rise to another admissible potential $F := F^{left}[a,b]$ whose nonlinear Fourier transform is equal to $M[a,b]$. This will turn out to be the leftmost inverse NLFT, which contains the largest amount of energy on left half-lines amongst all inverse nonlinear Fourier transforms of $M[a,b]$. The Dirac operator associated to this potential also has purely absolutely continuous spectrum on $\mathbf{T}$. Note also from (2.17) that $F^{left}[a,b]$ and $F^{right}[a,b]$ must have exactly the same energy on the line $\mathbf{Z}$, namely $1/a(0)$. However as we shall see the energy of these two potentials can be distributed in different ways on the line.

In particular, we have now shown that every scattering datum $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ arises in at least one way (and possibly two) as the non-linear Fourier transform of an admissible potential. In the next section we shall clarify this statement further, when we prove the triple factorization theorem (Theorem 2.7). For now, we at least show that these two examples of an inverse NLFT can be used to characterise the uniqueness of the inverse NLFT.

PROPOSITION 2.16. *Let $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ be a scattering datum, and let $\mathbf{H}$, $V^{min}_{<N}$, $V^{max}_{<N}$, $F^{right}[a,b]$, $F^{left}[a,b]$ be as above. Then the following statements are equivalent.*

- *(i) The vectors $\{\Omega_{0\leftarrow+\infty}(w_n)\}_{n<0}$, $\{\Omega_{0\leftarrow-\infty}(v_n)\}_{n<0}$ $\{\Omega_{0\leftarrow+\infty}(v_n)\}_{n\geq 0}$, $\{\Omega_{0\leftarrow-\infty}(w_n)\}_{n\geq 0}$ span a dense subspace of $\mathbf{H}$.*
- *(ii) $V^{min}_{<0} = V^{max}_{<0}$.*
- *(iii) $V^{min}_{<N} = V^{max}_{<N}$ for all N.*
- *(iv) $F^{left}[a,b] = F^{right}[a,b]$.*
- *(v) $E(F^{left}[a,b]|_{[N,+\infty)}) = E(F^{right}[a,b]|_{[N,+\infty)})$ for some integer N.*
- *(vi) $E(F^{left}[a,b]|_{(-\infty,-N)}) = E(F^{right}[a,b]|_{(-\infty,N)})$ for some integer N.*
- *(vii) $M[a,b]$ has a unique inverse NLFT.*

PROOF. The equivalence of (i) and (ii) follows from the definitions of $V^{min}_{<0}$ and $V^{max}_{<0}$ (and recalling that $V^{min}_{<0}$ is contained in $V^{max}_{<0}$. The implication $(iii) \implies (ii)$ is trivial. To see that (ii) implies (iii), merely note that both $V^{min}_{<N}$ and $V^{max}_{<N}$ obey (2.77), which allow $V_{<N}$ to be constructed recursively from $V_{<0}$. The implication (iii) implies (iv) is trivial, as is the implication that (iv) implies (v). The equivalence of (v) and (vi) is just the statement that $F^{max}[a,b]$ and $F^{min}[a,b]$ have the same energy on $\mathbf{Z}$. To see that (vi) implies (iii), suppose for contradiction that $V^{min}_{<N}$ was strictly smaller than $V^{max}_{<N}$ for some N, and hence for all N (again by using (2.77)). By parity and conjugation invariance of these spaces, this also implies that $P_\pm V^{min}_{<N}$ is strictly smaller than $P_\pm V^{max}_{<N}$ for both choices of sign $\pm$. Let v^{min}_N be the basis of $\mathbf{H}$ constructed in Theorem 2.15, i.e. the unique unit vector in $P_{(-1)^N} V^{min}_{<N+1}$ orthogonal to the codimension one subspace $P_{(-1)^N} V^{min}_{<N}$ which had a positive inner product with $\Omega_{0\leftarrow-\infty}(v_N)$. Since $\Omega_{0\leftarrow-\infty}(v_N)$ was orthogonal to $P_{(-1)^N} V^{min}_{<N}$ by (2.77), we thus see that the vector

$$\frac{v^{min}_N}{\langle \Omega_{0\leftarrow-\infty}(v_N), v^{min}_N\rangle_{\mathbf{H}}}$$

is nothing more than the orthogonal projection of $\Omega_{0\leftarrow-\infty}(v_N)$ to $P_{(-1)^N} V^{min}_{<N+1}$. Similarly

$$\frac{v^{max}_N}{\langle \Omega_{0\leftarrow-\infty}(v_N), v^{max}_N\rangle_{\mathbf{H}}}$$

is the orthogonal projection of the same vector $\Omega_{0\leftarrow-\infty}(v_N)$ to the larger space $P_{(-1)^N} V^{max}_{<N+1}$. Since the projection to the larger space clearly has the larger norm, we thus have

$$\langle \Omega_{0\leftarrow-\infty}(v_N), v^{max}_N\rangle_{\mathbf{H}} < \langle \Omega_{0\leftarrow-\infty}(v_N), v^{min}_N\rangle_{\mathbf{H}}.$$

But by the limiting version of the boundary value formula in Lemma 2.9 (or (2.83)) we have

$$\langle \Omega_{0\leftarrow-\infty}(v_N), v^{max}_N\rangle_{\mathbf{H}} = \prod_{n<N} \sqrt{1 - |F^{max}_N[a,b]|^2}$$

and

$$\langle \Omega_{0\leftarrow-\infty}(v_N), v_N^{min}\rangle_{\mathbf{H}} = \prod_{n<N} \sqrt{1-|F_N^{min}[a,b]|^2}$$

and hence

$$E(F^{max}[a,b]|_{(-\infty,N)}) > E(F^{min}[a,b]|_{(-\infty,N)}).$$

Since N is arbitrary, this contradicts[16] (vi) as desired.

Finally, observe that (vii) trivially implies (iv). Now assume (iv) (and hence (i)-(vi), by the previous discussion), and suppose that F is an inverse NLFT to $M[a,b]$, and form the transfer matrices $M[a_-,b_-] = M_{-\infty\leftarrow 0}$ and $M[a_+,b_+] = M_{0\leftarrow+\infty}$; by Theorem 2.4, Corollary 2.5, Lemma 2.6 these matrices solve the Riemann-Hilbert problem (2.34).

Observe that the pair

$$(u_{-\infty}, u_{+\infty}) := (\begin{pmatrix} \frac{a_+^*(\zeta^{-2})}{a^*(\zeta^{-2})} \\ -\frac{\zeta^{-1}b_-(\zeta^{-2})}{a^*(\zeta^{-2})} \end{pmatrix}, \begin{pmatrix} \frac{a_-(\zeta^{-2})}{a(\zeta^{-2})} \\ \frac{\zeta^{-1}b_+(\zeta^{-2})}{a(\zeta^{-2})} \end{pmatrix})$$

is an element of $\mathbf{H}$. Indeed, by (2.63) this is nothing more than the adjoint wave operators of F applied to v_0:

$$(u_{-\infty}, u_{+\infty}) = (\Omega_{-\infty\leftarrow 0}[F](v_0), \Omega_{+\infty\leftarrow 0}[F]v_0),$$

and the claim then follows from (2.59). (Alternatively, one could use Lemma 2.12 to verify that all the factors in $(u_{-\infty}, u_{+\infty})$ are square integrable, and then use (2.58) and (2.34) to verify that $u_{+\infty} = \Omega_{+\infty\leftarrow-\infty} u_{-\infty}$).

Observe that $u_{-\infty}$ extends holomorphically to $\mathcal{D}$, while $u_{+\infty}$ extends holomorphically to $\mathcal{D}^*$. Thus $u_{-\infty}$ is orthogonal to v_n for all $n < 0$ and w_n for all $n > 0$, while $u_{+\infty}$ is similarly orthogonal to w_n for all $n < 0$ and v_n for all $n > 0$. This implies that $(u_{-\infty}, u_{+\infty})$ is orthogonal to $V_{<0}^{min}$ and $(V_{<1}^{max})^\perp$. But we are assuming that $V_{<1}^{max}$ is equal to $V_{<1}^{min}$, hence $(u_{-\infty}, u_{+\infty})$ lies in the orthogonal complement of $V_{<0}^{min}$ in $V_{min}^{<1}$. As we know from the proof of Theorem 2.15, this space is two-dimensional and is spanned by v_0^{min} and w_0^{min}. But from applying the parity operator to the definition of $(u_{-\infty}, u_{+\infty})$ we see that this vector lies in the range of P_+, and hence must be a constant multiple of v_0^{min}, say $(u_{-\infty}, u_{+\infty}) = cv_0^{min}$. Applying (2.63) to determine the adjoint wave operators of v_0^{min} this means that

$$a_+ = c^* a_+^{min}; \quad b_- = cb_-^{min}; \quad a_- = ca_-^{min}; \quad b_+ = cb_-^{min},$$

where $M[a_\pm^{min}, b_\pm^{min}]$ are the half-line transfer matrices associated with $F^{min}[a,b]$. But since a_+ and a_+^{min} (for instance) are both real and positive at the origin we know that c is real; since $M[a_-,b_-]M[a_+,b_+] = M[a_-^{min}, b_-^{min}]M[a_+^{min}, b_+^{min}] = M[a,b]$ we see that $c^2 = 1$. Thus $c = 1$, which implies that $a_\pm = a_\pm^{min}$ and $b_\pm = b_\pm^{min}$. By Theorem 2.4 and Corollary 2.5 we thus see that $F = F^{min}[a,b]$, and the uniqueness is proved. $\square$

This proposition gives, at least in principle, a means of determining whether a potential $M[a,b]$ has a unique inverse NLFT, but it does not seem very easy to work with. It seems of interest to determine a better criterion for uniqueness.

---

[16] In fact this argument shows more: that enlarging the space $V_{<N}$ leads to shifting more of the energy of the corresponding potential F to the left of N instead of to the right.

2.10. Proof of triple factorization

We can now begin the proof of Theorem 2.7. Define the Riesz projections $P_{[0,+\infty)} : L^2(\mathbf{T}) \to H^2(\mathcal{D})$ and $P_{(-\infty,0)} : L^2(\mathbf{T}) \to H^2_0(\mathcal{D}^*)$ to be the orthogonal projections of $L^2(\mathbf{T})$ to $H^2_0(\mathcal{D}^*)$. We begin with a key observation that the left component of $F^{right}[a,b]$ depends on only one component of a, b, namely $P_{(-\infty,0)}(b/a) = P_{(-\infty,0)}(r)$.

Proposition 2.17. *Let $M[a,b]$ be a scattering datum in $\mathcal{L}^2(\mathbf{T})$. Then the restriction of $F^{min}[a,b]$ to $(-\infty,0)$ depends only on $P_{(-\infty,0)}(r) = P_{(-\infty,0)}(b/a)$; in other words, if $M[\tilde{a},\tilde{b}]$ is another scattering datum in $\mathcal{L}^2(\mathbf{T})$ with $P_{(-\infty,0)}(b/a) = P_{(-\infty,0)}(\tilde{b}/\tilde{a})$, then the corresponding potential $F^{right}[\tilde{a},\tilde{b}]$ agrees with $F^{right}[a,b]$ on $(-\infty,0)$.*

As we shall see later, the converse of this proposition is also true: one can recover $P_{(-\infty,0)}(b/a)$ from the values of $F^{right}[a,b]$ on $(-\infty,0)$. The linear analogue of this is that one can recover $P_{(-\infty,0)}(\hat{F})$ from the values of F on $(-\infty,0)$ and vice versa.

Proof. The idea is to run the inversion procedure in Theorem 2.15 carefully and note that to recover the left half of $F^{min}[a,b]$ only requires knowledge of $P_{(-\infty,0)}(b/a)$.

We recall the Hilbert space $\mathbf{H}$ introduced in the previous section, and recall the space $V^{min}_{<0}$ defined in that section as

$$V^{min}_{<0} = \sum_{n<0} \mathbf{C}\Omega_{0\leftarrow+\infty}(w_n) + \mathbf{C}\Omega_{0\leftarrow-\infty}(v_n).$$

The set of vectors

$$\{\Omega_{0\leftarrow+\infty}(w_n), \Omega_{0\leftarrow-\infty}(v_n) : n < 0\} \tag{2.92}$$

thus span (a dense subspace of) $V^{min}_{<0}$. Let us now understand the Hilbert space structure of (2.92); in other words, let us compute all the inner products between the elements of (2.92). They are all unit vectors, and the vectors $\{\Omega_{0\leftarrow+\infty}(w_n) : n < 0\}$ and $\{\Omega_{0\leftarrow-\infty}(v_n) : n < 0\}$ are separately orthonormal (since the wave maps $\Omega_{0\leftarrow\pm\infty}$ are isometries). But the first set of vectors are not orthogonal to the second. Indeed, we have

$$\langle \Omega_{0\leftarrow+\infty}(w_n), \Omega_{0\leftarrow-\infty}(v_m)\rangle_{\mathbf{H}} = \langle w_n, \Omega_{+\infty\leftarrow-\infty} v_m\rangle_{L^2(\mathbf{T})\oplus L^2(\mathbf{T})},$$

and hence by (2.58) and (2.3)

$$\langle \Omega_{0\leftarrow+\infty}(w_n), \Omega_{0\leftarrow-\infty}(v_m)\rangle_{\mathbf{H}} = \int_{\mathbf{T}} \zeta^{-n-m+1} \frac{b^*(\zeta^{-2})}{a^*(\zeta^{-2})}.$$

Since $n, m < 0$, we see that we may replace b/a by $P_{(-\infty,0)}(b/a)$ without affecting the above integral, thus

$$\langle \Omega_{0\leftarrow+\infty}(w_n), \Omega_{0\leftarrow-\infty}(v_m)\rangle_{\mathbf{H}} = \int_{\mathbf{T}} \zeta^{-n-m+1} (P_{(-\infty,0)}(b/a))^*(\zeta^{-2}).$$

Thus, the Hilbert space structure of the vectors (2.92) which span $V^{min}_{<0}$ is determined entirely by $P_{(-\infty,0)}(b/a)$.

Define the map $\Psi : V^{min}_{<0} \to l^2((-\infty,0)) \oplus l^2((-\infty,0))$ by

$$\Psi(v) := ((\langle v, \Omega_{0\leftarrow+\infty}(w_n)\rangle_{\mathbf{H}})_{n<0}, (\langle v, \Omega_{0\leftarrow-\infty}(v_n)\rangle_{\mathbf{H}})_{n<0});$$

this is a continuous linear map since the vectors $\{\Omega_{0\leftarrow+\infty}(w_n) : n < 0\}$ and $\{\Omega_{0\leftarrow-\infty}(v_n) : n < 0\}$ are separately orthonormal. It is also injective since (2.92) spans $V^{min}_{<0}$. Thus we can define the Hilbert space $\Psi(V^{min}_{<0})$, endowed with the Hilbert space structure pushed forward from $V^{min}_{<0}$ via Ψ (so the Hilbert space structure on $\Psi(V^{min}_{<0})$ is *not* the one induced from the ambient space $l^2((-\infty,0)) \oplus l^2((-\infty,0)))$. This space is spanned by the vectors $\Psi(\Omega_{0\leftarrow+\infty}(w_n))$ and $\Psi(\Omega_{0\leftarrow-\infty}(v_n))$ for $n < 0$. But by the previous discussion, these vectors (thought of as elements of "$l^2((-\infty,0)) \oplus l^2((-\infty,0)))$ are completely determined by $P_{(-\infty,0)}(b/a)$, as are the inner product in $\Psi(V^{min}_{<0})$ between these vectors. This means thata the space $\Psi(V^{min}_{<0})$, and the Hilbert space structure on this space, is determined completely by $P_{(-\infty,0)}(b/a)$. Also the action of the parity operator σ, pushed forward to $\Psi(V^{min}_{<0})$ by Ψ, is also completely determined by $P_{(-\infty,0)}(b/a)$ because its action on the basis vectors $\Psi(\Omega_{0\leftarrow+\infty}(w_n))$ and $\Psi(\Omega_{0\leftarrow-\infty}(v_n))$ (which are determined by $P_{(-\infty,0)}(b/a)$) is known. In particular, the parity projection operators $P_\pm$, pushed forward by Ψ, can be defined on the parity-invariant space $\Psi(V^{min}_{<0})$ and are completely determined by $P_{(-\infty,0)}(b/a)$.

Now consider the subspaces $V^{min}_{<n}$ in $V^{min}_{<0}$ for $n \leq 0$. This space, by definition, is spanned by a specific subset of (2.92). Thus $\Psi(V^{min}_{<n})$ can be determined as the span inside $\Psi(V^{min}_{<0})$ of a set of vectors which is completely determined by $P_{(-\infty,0)}(b/a)$. One can then compute the orthogonal complements of $P_\pm\Psi(V^{min}_{<n})$ in $P_\pm\Psi(V^{min}_{<n+1})$, and thus determine the unit vectors $\Psi(v^{min}_n)$, $\Psi(w^{min}_n)$ for $n < 0$ up to a complex phase, again using only $P_{(-\infty,0)}(b/a)$ and no other knowledge of a or b. But we know that $\Psi(v^{min}_n)$ has a positive inner product with $\Psi(\Omega_{0\leftarrow-\infty}v_n)$ and so in fact we can determine $\Psi(v^{min}_n)$ exactly in terms of $P_{(-\infty,0)}(b/a)$. Similarly we can reconstruct $\Psi(w^{min}_n)$ from $P_{(-\infty,0)}(b/a)$.

Recall that the operator L maps $V^{min}_{<-1}$ to $V^{min}_{<0}$, and hence we can push this forward by Ψ to create a map from $\Psi(V^{min}_{<-1})$ to $\Psi(V^{min}_{<0})$. This map can be completely determined by $P_{(-\infty,0)}(b/a)$ because we know its action on basis vectors, namely it maps $\Psi(\Omega_{0\leftarrow-\infty}(v_n))$ to $\Psi(\Omega_{0\leftarrow-\infty}(v_{n+1}))$ and $\Psi(\Omega_{0\leftarrow+\infty}(w_n))$ to $\Psi(\Omega_{0\leftarrow-\infty}(w_{n-1}))$ by (2.45). In particular, we can determine $\Psi(Lv^{min}_n)$ and $\Psi(Lw^{min}_n)$ for $n < -1$ entirely in terms of $P_{(-\infty,0)}(b/a)$. By (2.4) we can thus reconstruct $F^{right}_n[a,b]$ for $n < 0$ entirely in terms of $P_{(-\infty,0)}(b/a)$, as desired. $\square$

A similar argument gives

PROPOSITION 2.18. *Let $M[a,b]$ be a scattering datum in $\mathcal{L}^2(\mathbf{T})$. Then the restriction of $F^{left}[a,b]$ to $[0,+\infty)$ depends only on $P_{[0,+\infty)}(s^*) = P_{[0,+\infty)}(b/a^*)$.*

To apply Proposition 2.17 and Proposition 2.18 we make the following observation:

LEMMA 2.19. *Let $M[a_-,b_-] \in \mathcal{H}^2_0(\mathcal{D}^*)$ and $M[a_+,b_+] \in \mathcal{H}^2(\mathcal{D})$, and define $M[a,b]$ by $M[a,b] = M[a_-,b_-]M[a_+,b_+]$ (thus $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ by Lemma 2.6). Then $P_{(-\infty,0)}(b/a) = P_{(-\infty,0)}(b_-/a_-)$ and $P_{[0,+\infty)}(b/a^*) = P_{[0,+\infty)}(b/a^*)$.*

PROOF. From the identity

$$M[a_+,b_+] = M[a_-,b_-]^{-1}M[a,b] = M[a_-^*,-b_-]M[a,b] = M[a_-^*a - b_-^*b, a_-b - b_-a]$$

we thus see that $b_+ = a_-b - b_-a$ and thus

$$\frac{b}{a} - \frac{b_-}{a_-} = \frac{b_+}{aa_-}.$$

The left-hand side lies in $L^2(\mathbf{T})$, and the right-hand side extends holomorphically to $[0,+\infty)$, thus the left-hand side actually lies in $H^2(\mathcal{D})$ and vanishes when $P_{(-\infty,0)}$ is applied, which proves the first claim. The second claim similarly follows from the identity

$$M[a_-,b_-] = M[a,b]M[a_+,b_+]^{-1} = M[a,b]M[a_+^*,-b_+] = M[aa_+^*-b^*b_+, ba_+^*-a^*b_+]$$

and hence $b_- = ba_+^* - a^*b_+$, and thus

$$\frac{b}{a^*} - \frac{b_+}{a_+^*} = \frac{b_-}{a^*a_+^*}$$

and one argues as before. □

Combining this lemma with the above two Propositions we immediately obtain a preliminary result on the Riemann-Hilbert problem.

Corollary 2.20. *Let $M[a,b] \in \mathcal{L}^2(\mathbf{T})$ be scattering data, and let $M[a_-,b_-] \in \mathcal{H}_0^2(\mathcal{D}^*)$ and $M[a_+,b_+] \in \mathcal{H}^2(\mathcal{D})$ be a solution to the Riemann-Hilbert problem* (2.34). *Then*

$$F^{right}[a,b]|_{(-\infty,0)} = F^{right}[a_-,b_-]|_{(-\infty,0)}$$

and

$$F^{left}[a,b]|_{[0,+\infty)} = F^{left}[a_+,b_+]|_{[0,+\infty)}$$

We are now in a position to prove the triple factorization theorem.

Proof of Theorem 2.7. Fix $M[a,b] \in \mathcal{L}^2(\mathbf{T})$. Define F_{--} to be the restriction of $F^{right}[a,b]$ to $(-\infty,0)$ and F_{++} to be the restriction of $F^{left}[a,b]$ to $[0,+\infty)$; thus F_{--} and F_{++} are admissible potentials on the left and right half lines respectively. In particular, if we define $M[a_{--},b_{--}]$ and $M[a_{++},b_{++}]$ to be the nonlinear Fourier transforms of F_{--} and F_{++} then we have $M[a_{--},b_{--}] \in \mathcal{H}_0^2(\mathcal{D}^*)$ and $M[a_{++},b_{++}] \in \mathcal{H}^2(\mathcal{D})$ by Theorem 2.4 and Corollary 2.5.

Now suppose we have some solution $M[a,b] = M[a_-,b_-]M[a_+,b_+]$ to the Riemann-Hilbert problem (2.34). By Corollary 2.20 we know that $F^{right}[a_-,b_-]$ is equal to F_{--} on $(-\infty,0)$, and thus we may write

$$F^{right}[a_-,b_-] = F_{--} + F_{-0}$$

for some admissible potential F_{-0} supported on $[0,+\infty)$. If we write $M[a_{-0},b_{-0}]$ for the nonlinear Fourier transform of F_{-0}, we thus see from (2.16) that

$$M[a_-,b_-] = M[a_{--},b_{--}]M[a_{-0},b_{-0}].$$

Now let us investigate the holomorphicity properties of $M[a_{-0},b_{-0}]$. From Theorem 2.4 we already know that $M[a_{-0},b_{-0}]$ lies in $\mathcal{H}^2(\mathcal{D})$, so a_{-0} is outer on $\mathcal{D}$ and b_{-0}/a_{-0} lies in $H^2(\mathcal{D})$. But from the identity

$$\begin{aligned} M[a_{--},b_{--}] &= M[a_-,b_-]M[a_{-0},b_{-0}]^{-1} \\ &= M[a_-,b_-]M[a_{-0}^*,-b_{-0}] \\ &= M[a_-a_{-0}^* - b_-^*b_{-0}, b_-a_{-0}^* - a_-b_{-0}] \end{aligned}$$

we have $b_{--} = b_-a_{-0}^* - a_-^*b_{-0}$ and hence

$$\frac{b_{-0}}{a_{-0}^*} = \frac{b_-}{a_-^*} - \frac{b_{--}}{a_-^*a_{-0}^*}.$$

Observe that the right-hand side extends holomorphically to $\mathcal{D}^*$, and hence so must the left-hand side. Since the left-hand side is also in $L^2(\mathbf{T})$, we see that $b_{-0}/a^*_{-0} \in H^2(\mathcal{D}^*)$ and hence $M[a_{-0}, b_{-0}]$ lies in $\mathcal{H}^2_0(\mathcal{D}^*)$ as well as $\mathcal{H}^2(\mathcal{D})$. Thus it in fact lies in the space $\mathcal{H}^0$ defined in the discussion after (2.34).

A similar argument allows us to write

$$M[a_+, b_+] = M[a_{0+}, b_{0+}]M[a_{++}, b_{++}]$$

where $M[a_{0+}, b_{0+}] \in \mathcal{H}^0$. In particular we have

$$M[a, b] = M[a_-, b_-]M[a_+, b_+] = M[a_{--}, b_{--}]M[a_{-0}, b_{-0}]M[a_{0+}, b_{0+}]M[a_{++}, b_{++}].$$

If we define

$$M[a_0, b_0] := M[a_{--}, b_{--}]^{-1}M[a, b]M[a_{++}, b_{++}]^{-1}$$

then we have

$$M[a_{-0}, b_{-0}]M[a_{0+}, b_{0+}] = M[a_0, b_0]$$

for all solutions to the Riemann-Hilbert problem (2.34). By Lemma 2.8 we see that $M[a_0, b_0]$ is also in $\mathcal{H}^0$ (note that we have at least one solution to the Riemann-Hilbert problem (2.34)).

We have established a solution to (2.35), but we have not yet verified that $M[a_{--}, b_{--}]$ lies in $\mathcal{H}^-$ and that $M[a_{++}, b_{++}]$ lies in $\mathcal{H}^+$. We shall just prove the former claim, as the latter is similar. Suppose that $M[a_{--}, b_{--}]$ is not in $\mathcal{H}^-$, so it does not have a unique inverse RHP. By Corollary 2.5, there must therefore be a solution to the RHP

$$M[a_{--}, b_{--}] = M[\tilde{a}_-, \tilde{b}_-]M[\tilde{a}_+, \tilde{b}_+] \tag{2.93}$$

with $M[\tilde{a}_-, \tilde{b}_-] \in \mathcal{H}^2_0(\mathcal{D}^*)$, $M[\tilde{a}_+, \tilde{b}_+] \in \mathcal{H}^2(\mathcal{D})$, and with $M[\tilde{a}_+, \tilde{b}_+]$ not equal to the identity $M[1, 0]$. But then $M[\tilde{a}_-, \tilde{b}_-]$ is one part of a solution to the Riemann-Hilbert problem (2.34), since

$$M[a, b] = M[\tilde{a}_-, \tilde{b}_-](M[\tilde{a}_+, \tilde{b}_+]M[a_0, b_0]M[a_{++}, b_{++}])$$

and the expression in parentheses is in $\mathcal{H}^2(\mathcal{D})$ by Lemma 2.8. Thus by the previous analysis we have a factorization

$$M[\tilde{a}_-, \tilde{b}_-] = M[a_{--}, b_{--}]M[\tilde{a}_{0-}, \tilde{b}_{0-}]$$

for some $M[\tilde{a}_{0-}, \tilde{b}_{0-}] \in \mathcal{H}^0$. Taking energies of both sides using Lemma 2.6 we obtain in particular that

$$E(M[\tilde{a}_-, \tilde{b}_-]) = E(M[a_{--}, b_{--}])E(M[\tilde{a}_{0-}, \tilde{b}_{0-}])$$

while from (2.93) we similarly have

$$E(M[a_{--}, b_{--}]) = E(M[\tilde{a}_-, \tilde{b}_-])E(M[\tilde{a}_+, \tilde{b}_+]).$$

But all energies are finite and greater than or equal to one, which forces $E(M[\tilde{a}_+, \tilde{b}_+])$ to equal 1, which forces (e.g. by inverting the NLFT and using (2.17)) $M[\tilde{a}_+, \tilde{b}_+]$ to equal $M[1, 0]$, a contradiction. Thus $M[a_{--}, b_{--}]$ does lie in $\mathcal{H}^-$, and similarly $M[a_{++}, b_{++}]$ lies in $\mathcal{H}^+$. This gives the factorization (2.35).

Now we show uniqueness. Suppose that there is an alternative factorization

$$M[a, b] = M[\tilde{a}_{--}, \tilde{b}_{--}]M[\tilde{a}_0, \tilde{b}_0]M[\tilde{a}_{++}, \tilde{b}_{++}]$$

with the three factors on the right in $\mathcal{H}^-$, $\mathcal{H}^0$, $\mathcal{H}^+$ respectively. Then $M[\tilde{a}_{--}, \tilde{b}_{--}]$ is again a left factor of a solution to the Riemann-Hilbert problem (2.34), and so again we have a factorization

$$M[\tilde{a}_{--}, \tilde{b}_{--}] = M[a_{--}, b_{--}]M[a_{-0}, b_{-0}]$$

for some $M[a_{-0}, b_{-0}] \in \mathcal{H}^0$. But we of course also have the factorization

$$M[\tilde{a}_{--}, \tilde{b}_{--}] = M[\tilde{a}_{--}, \tilde{b}_{--}]M[1, 0].$$

But $M[\tilde{a}_{--}, \tilde{b}_{--}]$ lies in $\mathcal{H}^-$ and hence should have unique inverse RHP, which is only possible of $M[\tilde{a}_{--}, \tilde{b}_{--}] = M[a_{--}, b_{--}]$. A similar argument shows that $M[\tilde{a}_{++}, \tilde{b}_{++}] = M[a_{++}, b_{++}]$, and hence $M[\tilde{a}_0, \tilde{b}_0] = M[a_0, b_0]$. This establishes uniqueness of the factorization (2.35).

The energy identity (2.36) follows from two applications of Lemma 2.6. The factorization (2.37) and (2.38) were already established by above discussion. Finally, the converse implication, that every solution to (2.38) induces a solution to (2.34) follows from several applications of Lemma 2.8. This completes the proof of Theorem 2.7. □

By Theorem 2.7, every scattering datum $M[a, b]$ can be split canonically into three components, a left-line component $M[a_{--}, b_{--}]$ which has a unique inverse NLFT, supported on $(-\infty, 0)$, a right-line component $M[a_{++}, b_{++}]$ which has a unique inverse NLFT, supported on $[0, +\infty)$, and a central component $M[a_0, b_0]$ which can be inverted either on the left line, the right line, or some combination of the two. In a future paper we will compute this factorization more explicitly in the case when b (and hence a) are rational functions; it then turns out that b_{--} is generated by the poles of b in $\mathcal{D}$, b_{++} is generated by the poles of b in $\mathcal{D}^*$, and b_0 is generated by the poles of b in $\mathbf{T}$.

Note also that out of all the solutions to the Riemann-Hilbert problem (2.34), the factorization $M[a, b] = M[a_{--}, b_{--}](M[a_0, b_0]M[a_{++}, b_{++}])$ has the largest energy on the right factor (and hence the least energy on the left factor), while the other extreme factorization $M[a, b] = (M[a_{--}, b_{--}]M[a_0, b_0])M[a_{++}, b_{++}]$ behaves of course in the converse direction. This explains our earlier remark that F^{right} is the solution to the inverse NLFT of $M[a, b]$ with the most mass in $[0, +\infty)$ (or indeed in $[N, +\infty)$ for any N), and F^{left} is the solution with the most mass in $(-\infty, 0)$ (or $(-\infty, N)$).

As discussed in the previous section, the two extreme solutions of the Riemann Hilbert problem both had Dirac operators $L[F]$ with purely absolutely continuous spectrum. Thus if $L[F]$ has some singular spectrum, the inverse NLFT for $\overbrace{F}$ cannot be unique. It seems reasonable to conjecture a converse, that if $M[a, b]$ does not have unique inverse NLFT, then there exists a solution F to the inverse NLFT problem such that $L[F]$ has some singular spectrum. In the case when b and a are rational functions, we have verified this conjecture (indeed in this case we can construct a continuous family of $L[F]$ each of which contains embedded eigenvalues at the poles of b); we shall detail this in a later paper.

If $M[a, b]$ lies in $\mathcal{H}^2_0(\mathcal{D})$, then the right-line component $M[a_{++}, b_{++}]$ becomes trivial, i.e. $M[a_{++}, b_{++}] = M[1, 0]$, since clearly $M[a, b] = M[a, b]M[1, 0]$ is the solution to (2.34) with the least amount of energy on the second factor. Thus we see (from Lemma 2.8) that $\mathcal{H}^2_0(\mathcal{D})$ factorizes uniquely as $\mathcal{H}^- \cdot \mathcal{H}^0$, and similarly $\mathcal{H}^2(\mathcal{D})$ factorizes uniquely as $\mathcal{H}^0 \cdot \mathcal{H}^+$.

It seems of interest to determine which admissible potentials F on the left halfline $(-\infty, 0)$ (for instance) have non-linear Fourier transforms in $\mathcal{H}^-$, and which ones have non-linear Fourier transforms in $\mathcal{H}^0$. We do not know the answer to this question, but a tentative conjecture would be that the former occurs if and only if $L[F]$ has no singular spectrum on the half-line for any choice of boundary condition at 0. Again, in the case of rational scattering data we know that this is a necessary and sufficient condition to lie in $\mathcal{H}^-$, and we will detail this in a later paper.

We close this section with some (rather weak) sufficient conditions to guarantee uniqueness of the inverse NLFT.

Proposition 2.21. *If $M[a, b] \in \mathcal{H}^2_0(\mathcal{D}^*)$ and $b \in L^2(\mathbf{T})$, then $M[a, b]$ has unique inverse NLFT. Similarly if $M[a, b] \in \mathcal{H}(\mathcal{D})$ and $b \in L^2(\mathbf{T})$.*

It seems reasonable to conjecture that this $L^2(\mathbf{T})$ condition can be relaxed to $L^1(\mathbf{T})$, in analogy with the well-known result that a non-constant $L^1(\mathbf{T})$ function cannot have holomorphic extensions to both $\mathcal{D}$ and $\mathcal{D}^*$ simultaneously. The rational examples given by (2.74) shows that uniqueness breaks down if $L^1(\mathbf{T})$ is replaced by $L^{1,\infty}(\mathbf{T})$.

Proof. We just show this when $M[a, b] \in \mathcal{H}^2_0(\mathcal{D}^*)$, as the other claim is similar. As discussed above, this means that $M[a_{++}, b_{++}] = M[1, 0]$, which implies by construction of $M[a_{++}, b_{++}]$ that $F^{left}[a, b]$ vanishes on $[0, +\infty)$.

By Lemma 2.16 it suffices to show that $V^{max}_{<1}$ is contained in $V^{min}_{<1}$. In fact it will suffice to show that the single vector v^{max}_0 lies in $V^{min}_{<1}$. To see this, observe from $*$-invariance that this will imply w^{max}_0 also lies in $V^{min}_{<1}$. Also Lw^{max}_0 will then lie in $LV^{min}_{<1} \subseteq V^{min}_{<2}$ by (2.77), but this vector is orthogonal to both $\Omega_{0\leftarrow-\infty}(w_2)$ and $\Omega_{0\leftarrow+\infty}(v_2)$ (the former by parity considerations and the latter by (2.77) applied to $V^{max}_{<N}$ and (2.45)), and so Lw^{max}_0 lies in $V^{min}_{<1}$. Applying (2.4) we thus see that w^{max}_{-1} lies in $V^{min}_{<1}$, and then by $*$-invariance so does v^{max}_{-1}. Continuing in this manner we see in fact that all the basis vectors of $V^{max}_{<1}$ are contained in $V^{min}_{<1}$, and hence $V^{max}_{<1} = V^{min}_{<1}$ as desired.

It remains to show that v^{max}_0 lies in $V^{min}_{<1}$. Since v^{max}_0 is one of the basis vectors associated with $F^{left}[a, b]$, which vanishes on $[0, +\infty)$, we see from (2.4) that $v^{max}_0 = L[F]^{-m} v^{max}_m$ for all $m \geq 0$. Letting $m \to +\infty$ we thus see that $v^{max}_0 = \Omega_{0\leftarrow+\infty}(v_0)$, which by (2.58) is equal to

$$v^{max}_0 = \Omega_{0\leftarrow+\infty}(v_0) = (\begin{pmatrix} \frac{1}{a^*(\zeta^{-2})} \\ \frac{-\zeta^{-1} b(\zeta^{-2})}{a^*(\zeta^{-2})} \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot).$$

Using $aa^* - bb^* = 1$, we can split this as

$$v^{max}_0 = (\begin{pmatrix} a(\zeta^{-2}) \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ \zeta^{-1}\beta(\zeta^{-2}) \end{pmatrix} \cdot) + (\begin{pmatrix} \frac{-b(\zeta^{-2}) b^*(\zeta^{-2})}{a^*(\zeta^{-2})} \\ \frac{-\zeta^{-1} b(\zeta^{-2})}{a^*(\zeta^{-2})} \end{pmatrix}, \begin{pmatrix} 0 \\ -\zeta^{-1}\beta(\zeta^{-2}); \end{pmatrix} \cdot)$$

since b (and hence a) lie in $L^2(\mathbf{T})$, we can verify from (2.58) that both summands lie in $\mathbf{H}$. Furthermore, the first summand lies in the span of the orthonormal set $\{\Omega_{0\leftarrow-\infty}(v_n) : n \leq 0\}$ since $a(\zeta^{-2})$ lies in $H^2(\mathcal{D}^*)$, while the second summand similarly lies in the span of $\{\Omega_{0\leftarrow+\infty}(w_n) : n \leq 0\}$ since $-\zeta^{-1}\beta(\zeta^{-2})$ lies in $H^2(\mathcal{D})$. This shows that v^{max}_0 lies in $V^{min}_{<1}$ as desired. □

Corollary 2.22. *If $M[a, b] \in \mathcal{L}^2(\mathbf{T})$ and $b \in L^\infty(\mathbf{T})$, then $M[a, b]$ has unique inverse NLFT.*

Again, it seems that $L^\infty(\mathbf{T})$ should be weakened, at least to $L^2(\mathbf{T})$ and perhaps even to the real-variable Hardy space $H^1(\mathbf{T})$.

PROOF. We solve the Riemann-Hilbert problem (2.34) (e.g. using Theorem 2.7) to obtain a factorization $M[a,b] = M[a_-, b_-]M[a_+, b_+]$. Since b (and hence a) are bounded, we see from Lemma 2.12 that $a_\pm$, $b_\pm$ lie in L^2. Thus by Lemma 2.21 $M[a_-, b_-]$ lies in $\mathcal{H}^-$ and $M[a_+, b_+]$ lies in $\mathcal{H}^+$. Thus in the triple factorization of $M[a,b]$, $M[a_0, b_0] = M[1,0]$ is trivial, and so one has unique inverse NLFT by Theorem 2.7. $\square$

In the case where a and b is bounded one can in fact solve the Riemann-Hilbert problem (2.34) directly by means of inverting Hankel operators, by the method of Gelfand, Levitan, and Marcenko (see e.g. [**11**] for more details). Indeed, if one rewrites (2.34) as

$$M[a_-, b_-] = M[a,b]M[a_+, b_+]^{-1} = M[a,b]M[a_+^*, -b_+] = M[aa_+^* - b^*b_+, ba_+^* - a^*b_+]$$

we obtain the identities

$$a_+ = \frac{a_-^*}{a^*} + \frac{b}{a^*}b_+^*; \quad b_+ = -\frac{b_-}{a^*} + \frac{b}{a^*}a_+^*.$$

Since b (and hence a) is bounded, $b_\pm$, $a_\pm$ must lie in $L^2(\mathbf{T})$ thanks to Lemma 2.12. The function $\frac{a_-^*}{a^*}$ lies in $H^2(\mathcal{D}^*)$, while $\frac{b_-}{a^*}$ lies in $H_0^2(\mathcal{D}^*)$. Thus if we apply the projection operators $P_{[0,+\infty)}$ to these equations we obtain

$$a_+ = C + P_{[0,+\infty)}(sb_+^*); \quad b_+ = P_{[0,+\infty)}(sa_+^*)$$

where C is the value of $\frac{a_-^*}{a^*}$ at infinity. Since a is bounded, the reflection coefficient $s = b/a^*$ is bounded in magnitude by $1 - \varepsilon$ for some $\varepsilon > 0$, which implies that the map $f \mapsto P_{[0,+\infty)}(sf^*)$ is a strict contraction on $L^2(\mathbf{T})$. Thus we may use the contraction mapping theorem (or Neumann series) to solve for a_+ and b_+ up to a constant; one can then recover this constant by recalling that a_+ is positive at zero, and that the constant C equals $\frac{a_-(0)}{a(0)} = \frac{1}{a_+(0)}$.

Observe that a and b are necessarily bounded when the potential F is absolutely summable; this can be seen directly from the infinite product representation (2.2), which is absolutey convergent in this case. (This is the non-linear analogue of the fact that $l^1(\mathbf{Z})$ sequences have bounded Fourier transforms). But as we can already see from Theorem 2.4, the best size estimate we can expect for the non-linear Fourier transform of $l^2(\mathbf{Z}; \mathcal{D})$ sequences is that a (and hence b) are only log-integrable.

2.11. Lax pair

In this section we give the Lax pair formulation of the Ablowitz-Ladik equation (2.7). We begin with some basic remarks on well-posedness of this equation. Observe that if F is an $l^2(\mathbf{Z})$ sequence, then so is the right-hand side of (2.7); in fact, the dependence of the right-hand side is locally Lipschitz from $l^2(\mathbf{Z})$ to itself, and from this and an easy application of the Picard existence theorem we see that this equation is locally well-posed in $l^2(\mathbf{Z})$ (i.e. for any choice of $l^2(\mathbf{Z})$ initial data $F_n(0)$, there exists a unique $l^2(\mathbf{Z})$ solution existing for time depending on the $l^2(\mathbf{Z})$ norm of the data, and furthermore the map from data to solution is continuous in $l^2(\mathbf{Z})$). Furthermore, from the identity

$$\partial_t |F_n|^2 = i(1 - |F_n|^2)(F_n^* F_{n-1} - F_n F_{n-1}^* - F_n F_{n+1}^* + F_{n+1} F_n^*) \tag{2.94}$$

we see that the $l^2(\mathbf{Z})$ norm of F is conserved, and hence one in fact has global well-posedness in $l^2(\mathbf{Z})$. Indeed, from (2.94) and summation by parts we obtain the estimate

$$\partial_t \sum_n (1+n^2)^{k/2}|F_n|^2 \leq C(k, \|F\|_{l^2(\mathbf{Z})} \sum_n (1+n^2)^{k/2}|F_n|^2$$

for any $k \geq 0$. Thus we see that if F is rapidly decreasing in n at time $t = 0$, then it is also rapidly decreasing for all later times t.

The operator $L(t) = L[F(t)]$ also evolves in time. In fact the evolution is given by a Lax pair:

Lemma 2.23. *Let F be a potential evolving under the flow* (2.7)*. Then we have*

$$\partial_t L = [P, L]$$

where $P(t)$ is the skew-adjoint operator defined by

$$P = i(-\frac{LJL + L^*JL^*}{2} + D), \tag{2.95}$$

J is the reflection operator defined by

$$Jv_n := v_n; \quad Jw_n := -w_n$$

and $D = D(t)$ is the diagonal operator defined by

$$D = \frac{[L,J]^2 + [L^*,J]^2}{8} J$$

or equivalently

$$Dv_n := \frac{F_{n-1}^* F_n + F_{n-1} F_n^*}{2} v_n; \quad Dw_n := -\frac{F_{n-1}^* F_n + F_{n-1} F_n^*}{2} w_n.$$

Proof. P is clearly skew-adjoint, since J and D are self-adjoint. We first compute $\partial_t L$ on basis vectors. To abbreviate the notation we write $e_n := \sqrt{1-|F_n|^2}$ and $F_{n\pm1} := F_{n-1} + F_{n+1}$. From (2.7) we see that

$$\partial_t e_n = \frac{i}{2} e_n (F_n F_{n\pm1}^* - F_n^* F_{n\pm1}).$$

Differentiating (2.4), we thus obtain that

$$\begin{aligned} \partial_t L v_n &:= \frac{i}{2} e_n (F_n F_{n\pm1}^* - F_n^* F_{n\pm1}) v_n + i e_n^2 F_{n\pm1} w_n \\ \partial_t L w_{n+1} &:= i e_n^2 F_{n\pm1}^* v_{n+1} + \frac{i}{2} e_n (F_n F_{n\pm1}^* - F_n^* F_{n\pm1}) w_n. \end{aligned} \tag{2.96}$$

Now we compute $[P, L]$. We may split $L = A + B$, where A and B are given by

$$\begin{aligned} Av_n &:= e_n v_{n+1} + F_n w_n \\ Aw_{n+1} &:= 0 \\ Bv_n &:= 0 \\ Bw_{n+1} &:= -F_n^* v_{n+1} + e_n w_n. \end{aligned}$$

Then we clearly have $LJ = A - B$. Thus

$$[LJL, L] = [LJ, L]L = [A-B, A+B]L = 2[A,B]L.$$

Taking adjoints, we obtain

$$[L^*, L^*JL^*] = 2L^*[A,B]^*;$$

since L is unitary, we have $[X, L] = L[L^*, X]L$ for any operator X, and hence

$$[L^* J L^*, L] = 2[A, B]^* L.$$

Thus we have

$$[P, L] = -i([A, B] + [A, B]^*)L + i[D, L].$$

A direct computation shows that

$$[A, B]v_{n+1} = -BAv_{n+1} = -F_{n+1}Bw_{n+1} = F_{n+1}F_n^* v_{n+1} - F_{n+1}e_n w_n$$

and

$$[A, B]w_n = ABw_n = -F_{n-1}^* Av_n = -F_{n-1}^* e_n v_{n+1} - F_{n-1}^* F_n w_n.$$

Taking adjoints, we obtain

$$[A, B]^* v_{n+1} = F_{n+1}^* F_n v_{n+1} - F_{n-1} e_n w_n$$

and

$$[A, B]^* w_{n+1} = -F_{n+1}^* e_n v_{n+1} - F_{n-1} F_n^* w_n$$

and hence

$$\begin{aligned} ([A, B] + [A, B]^*)v_{n+1} &= (F_{n+1}^* F_n + F_{n+1} F_n^*)v_{n+1} - F_{n\pm 1} e_n w_n \\ ([A, B] + [A, B]^*)w_n &= -F_{n\pm 1}^* e_n v_{n+1} - (F_{n-1}^* F_n + F_{n-1} F_n^*)w_n. \end{aligned}$$

Combining this with (2.4), we obtain

$$\begin{aligned} ([A, B] + [A, B]^*)Lv_n =& e_n(F_{n+1}^* F_n + F_{n+1} F_n^* - F_n F_{n\pm 1}^*)v_{n+1} \\ & - (F_{n\pm 1} e_n^2 + F_n(F_{n-1}^* F_n + F_{n-1} F_n^*))w_n \\ ([A, B] + [A, B]^*)Lw_{n+1} =& - (F_n^*(F_{n+1}^* F_n + F_{n+1} F_n^*) + F_{n\pm 1} e_n^2)v_{n+1} \\ & + (F_n^* F_{n\pm 1} - F_{n-1}^* F_n - F_{n-1} F_n^*)e_n w_n \end{aligned}$$

Meanwhile, a direct computation shows that

$$[D, L]v_n = \frac{F_{n+1}^* F_n + F_{n+1} F_n^* - F_{n-1}^* F_n - F_{n-1} F_n^*}{2} e_n v_{n+1} - (F_{n-1}^* F_n + F_{n-1} F_n^*)F_n w_n$$

$$[D, L]w_{n+1} = -(F_{n+1}^* F_n + F_{n+1} F_n^*)F_n^* v_{n+1} + e_n \frac{F_{n+1}^* F_n + F_{n+1} F_n^* - F_{n-1}^* F_n - F_{n-1} F_n^*}{2} w_n.$$

Comparing these equations with (2.96), the claim follows. □

Formally, Lemma 2.23 implies that the Ablowitz-Ladik equation (2.7) is completely integrable and can be inverted by means of the non-linear Fourier transform. We now make this more rigorous.

PROPOSITION 2.24. *Let $F(t)$ be an $l^2(\mathbf{Z}; \mathcal{D})$ solution to (2.7). Then the non-linear Fourier transform $\overbrace{F(t)} = M[a(t), b(t)]$ obeys the equation*

$$\overbrace{F(t)} = M[\exp(-i(z + z^{-1})t/2), 0]\overbrace{F(t)}M[\exp(i(z + z^{-1})t/2), 0] \tag{2.97}$$

for all $t \in \mathbf{R}$, or in other words we have the relation (2.8).

PROOF. Although this argument is by now very well known, we include it here for completeness. Because we are in the discrete setting there will be very little difficulty in making the Lax pair formalism rigorous (for instance, spatial derivatives in this setting are just finite difference operators, which are bounded on every reasonable space).

We may restrict the time parameter to a fixed compact interval $[-T, T]$, and allow all our bounds to depend on T. It suffices to verify this relation for solutions $F(t)$ which are rapidly decreasing in space, since these solutions are dense in the class of $l^2(\mathbf{Z}; \mathcal{D})$ solutions, and we have well-posedness of the discrete equation (2.7) in $l^2(\mathbf{Z}; \mathcal{D})$ and continuity of the non-linear Fourier transform (from Lemma 2.6. One can then easily verify from (2.7) and repeated differentiation in time that F is infinitely differentiable in time and all of its time derivatives are also rapidly decreasing in space.

At time zero, we define generalized eigenfunctions

$$\Phi(0, \zeta) = \sum_n a_n(0, \zeta)\zeta^n v_n + b_n(0, \zeta)\zeta^{1-n} w_n$$

(i.e. we use the ansatz (2.42)) for each $\zeta \in \mathbf{T}$ by solving the eigenfunction equation

$$L[F(0)]\Phi(0, \zeta) = \zeta\Phi(0, \zeta)$$

with initial data

$$\lim_{n\to+\infty} a_n(0, \zeta) = 1; \quad \lim_{n\to+\infty} b_n(0, \zeta) = 0.$$

This eigenfunction is well-defined since F is rapidly decreaasing. Indeed from (2.43) and a limiting argument we see that

$$M[a_n(0, \zeta), b_n(0, \zeta)] = M_{n\leftarrow+\infty}(0, \zeta^2)M[1, 0],$$

and so in particular by taking limits as $n \to -\infty$

$$M[a_{-\infty}(0, \zeta), b_n(0, \zeta)] = \widehat{F(0)}(\zeta^2).$$

Because F is rapidly decreasing, it is absolutely integrable and hence all the transfer matrices are bounded. Thus $\Phi(0, \zeta)$ has bounded coefficients. We abuse notation slightly and use l^∞ to denote the Banach space of generalized functions $\sum_{n\in\mathbf{Z}} \phi_n v_n + \psi_n w_n$ with bounded coefficients, equipped with the norm

$$\| \sum_{n\in\mathbf{Z}} \phi_n v_n + \psi_n w_n \|_{l^\infty} := \max(\sup_n |\phi_n|, \sup_n |\psi_n|).$$

Now fix $\zeta \in \mathbf{T}$. We evolve Φ in time by the equation

$$\Phi_t = P(t)\Phi \tag{2.98}$$

where $P(t)$ is the operator defined in Lemma 2.23. Observe that the difference operator $P(t)$ is bounded on l^∞ since F stays bounded. Thus the Picard existence theorem guarantees a local-in-time solution to this equation in l^∞ as long as the l^∞ norm of Φ stays bounded; repeated differentiation then shows this solution is smooth in time. We now use Lemma 2.23 to compute

$$\begin{aligned}\partial_t(L[F(t)]\Phi(t, \zeta)) &= L_t[F(t)]\Phi(t, \zeta) + L[F(t)]\Phi_t(t, \zeta)\\ &= [P(t), L[F(t)]]\Phi(t, \zeta) + L[F(t)]\Phi_t(t, \zeta)\\ &= P(t)L[F(t)]\Phi(t, \zeta).\end{aligned}$$

Combinign this with the previous equation we see in particular that

$$\partial_t(L[F(t)]\Phi(t, \zeta) - \zeta\Phi(t, \zeta)) = P(t)(L[F(t)]\Phi(t, \zeta) - \zeta\Phi(t, \zeta)).$$

Since $(L[F(t)]\Phi(t, \zeta) - \zeta\Phi(t, \zeta))$ lies in l^∞, equals 0 at time $t = 0$, and $P(t)$ is bounded on l^∞, we can appeal to the uniqueness component of the Picard existence theorem to then conclude that $L[F(t)]\Phi(t, \zeta) - \zeta\Phi(t, \zeta)$ vanishes for all times t near

0. Thus we see that $\Phi(t,\zeta)$ solves the eigenfunction equation (2.39) for each t near 0. In particular, since $F(t)$ is rapidly decreasing, we may write

$$\Phi(t,\zeta) = \sum_n a_n(t,\zeta)\zeta^n v_n + b_n(t,\zeta)\zeta^{1-n} w_n \tag{2.99}$$

where $a_n(t,\zeta)$ and $b_n(t,\zeta)$ are rapidly convergent to some limiting value $a_{\pm\infty}(t,\zeta)$ and $b_{\pm\infty}(t,\zeta)$ as $n \to \pm\infty$, and furthermore

$$M[a_{-\infty}(t,\zeta), b_{-\infty}(t,\zeta)] = \widehat{F(t)}(\zeta^2) M[a_{+\infty}(t,\zeta), b_{+\infty}(t,\zeta)]. \tag{2.100}$$

Since the time derivatives of F exist and are also rapidly decreasing, it is easy to establish that the derivatives of $a_n(t,\zeta)$ and $b_n(t,\zeta)$ converge to those of $a_{\pm\infty}(t,\zeta)$ and $b_{\pm\infty}(t,\zeta)$ as $n \to \pm\infty$.

Now let us look at the action on $L(t)$ and $P(t)$ on basis vectors v_n, w_n as $|n| \to +\infty$. Since $F(t)$ is rapidly decreasing, $L[F(t)]$ behaves like $L[0]$ modulo an error rapidly decreasing in n. Since $L[0]$ and $L[0]^*$ commute with J, we thus see that Dv_n and Dw_n are rapidly decreasing in n. Thus we have

$$Pv_n = -i\frac{v_{n-2} + v_{n+2}}{2} + \dots; \quad Pw_n = +i\frac{w_{n-2} + w_{n+2}}{2} + \dots$$

where the error $\dots$ is rapidly decreasing in n. Applying this to (2.98), (2.99) we obtain

$$\partial_t a_n = -i\frac{\zeta^2 + \zeta^{-2}}{2} a_n + \dots; \quad \partial_t b_n = i\frac{\zeta^2 + \zeta^{-2}}{2} b_n + \dots.$$

Passing to the limits at $\pm\infty$, the $\dots$ errors decay rapidly to zero and we obtain

$$\partial_t a_{\pm\infty} = -i\frac{\zeta^2 + \zeta^{-2}}{2} a_{\pm\infty}; \quad \partial_t b_{\pm\infty} = i\frac{\zeta^2 + \zeta^{-2}}{2} b_{\pm\infty}$$

which can of course be solved explicitly as

$$a_{\pm\infty}(t) = \exp(-i\frac{\zeta^2 + \zeta^{-2}}{2} t) a_{\pm\infty}(0); \quad b_{\pm\infty}(t) = \exp(+i\frac{\zeta^2 + \zeta^{-2}}{2} t) a_{\pm\infty}(0).$$

Reconciling this with (2.100) we obtain

$$M[\exp(-i\frac{\zeta^2 + \zeta^{-2}}{2} t), 0]\widehat{F(0)}(\zeta^2) = \widehat{F(t)} M[\exp(-i\frac{\zeta^2 + \zeta^{-2}}{2} t), 0]$$

which is (2.97), at least for short times t. Note that this also shows that $a_{\pm\infty}(t)$ and $b_{\pm\infty}(t)$ stay bounded (in fact, their magnitudes are constant); this combined with the rapid decrease of $F(t)$ shows that the eigenfunctions $\Phi(t,\zeta)$ stay bounded, and so the Picard existence theorem allows us to continue this argument indefinitely in time on the compact interval $[-T, T]$. $\square$

CHAPTER 3

The $SU(2)$ scattering transform

This chapter is a revised version of the PhD thesis of the third author.

3.1. Introduction

This chapter develops the theory of the $SU(2)$ non-linear Fourier series or transform, so named because the Fourier transform data is an $SU(2)$ valued function as opposed to the $SU(1,1)$ valued non-linear Fourier transform discussed in the first two chapters of this book. Recall that $SU(2)$ is the set of complex valued matrices of the form

$$\begin{pmatrix} a & b \\ -\overline{b} & \overline{a} \end{pmatrix}$$

with determinant 1, while a $SU(1,1)$ matrix is of the form

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix}$$

with determinant 1.

What begins with a seemingly harmless change of a sign in the definition of the Fourier transform results in different behaviour in several respects. This percolates to well know phenomena in other branches of mathematics including the theory of non-linear dispersive equations. For example, compare the defocusing and focusing nonlinear Schrödinger equations:

$$i\frac{\partial F}{\partial t}(x,t) = -\frac{\partial^2 F}{\partial x^2} + 2|F|^2 F(x,t) \ ,$$

$$i\frac{\partial F}{\partial t}(x,t) = -\frac{\partial^2 F}{\partial x^2} - 2|F|^2 F(x,t) \ .$$

The defocusing equation can formally be solved by the $SU(1,1)$ nonlinear Fourier transform, while the focusing equation can be solved by the $SU(2)$ nonlinear Fourier transform. A major difference in the theory of these two equations is the appearance of soliton solutions for the the focusing equations, which are not present in the theory of the defocusing equation. We will use the word "soliton" at the corresponding place in the theory of the nonlinear Fourier transform.

As in the previous chapters we restrict attention to nonlinear Fourier series, i.e. we will take the Fourier transform of data which are functions on the set of integers $\mathbf{Z}$. Our main interest is in data in $l^2(\mathbf{Z}, \mathbf{C})$, the square summable complex valued sequences on $\mathbf{Z}$.

Parallel to the previous chapters, we discuss the following questions:

(1) How does one define the non-linear Fourier transform data for sequences $(F_n) \in l_2(\mathbf{Z}, \mathbf{C})$?
(2) Characterize the range of the nonlinear Fourier transform on $l^2(\mathbf{Z}, \mathbf{C})$ and its topology.
(3) Existence, (non-) uniqueness, and construction of preimages under the NLFT. These aspects of the theory are not completely understood, so we will discuss in more detail the special case of rational Fourier transform data.
(4) Discuss continuity of the NLFT and the inverse NLFT whenever appropriate.

In Section 2 we introduce the discrete $SU(2)$ NLFT by defining it first on finite sequences and then l^1 sequences, and we will derive some basic properties of the NLFT. It will map sequences of complex numbers to $SU(2)$ valued functions on the unit circle $\mathbf{T}$. For convenience, we will write $SU(2)$ matrices by their first row and write the NLFT data as (a, b) where a, b are complex valued functions on $\mathbf{T}$ with $|a|^2 + |b|^2 = 1$.

In Section 3 we work on half line sequences, i.e. in the spaces $l^2(\mathbf{Z}_{\geq 0})$ and $l^2(\mathbf{Z}_{<0})$. We will define the NLFT and prove that it is injective on these spaces. We also characterize the image spaces which will be denoted by $\mathbf{H}$ and $\mathbf{H}_0^*$. To prove that the NLFT from $l^2(\mathbf{Z}_{\geq 0})$ to $\mathbf{H}$ is onto, we construct the inverse NLFT by measn of the *"layer stripping method"*. Finally, a metric on $\mathbf{H}$ is defined such that the NLFT is a homeomorphism between $l^2(Z_{\geq 0})$ and $\mathbf{H}$. However, $\mathbf{H}$ is not complete under that metric. This will lead to some discussion about the completion of $\mathbf{H}$. We will observe that the incompleteness is mainly due to the appearance of *soliton factors* $(a, 0)$ under the limiting process, where $\bar{a}$ is the boundary value of an inner function.

Having established some understanding of the NLFT on half line data, it is natural to approach the NLFT of full line $l^2(Z)$ data as a product of NLFT data of two half lines. The inverse process is a factorization problem known as *Riemann-Hilbert factorization* . In Section 4 we study the special case of *rational* NLFT data (a, b) and find a bijection between the set of decompositions $(a_-, b_-)(a_+, b_+) = (a, b)$ and some *extended scattering data* $\{b, Z, n_i, m_i, \gamma_i\}$, where Z is the zero set of a^*, $\{n_i\}$ represents the orders of the zeros, and $\{m_i, \gamma_i\}$ is additional information not contained in (a, b) that describes the fibers of the inverse NLFT. To derive this result, we apply some theorems from [**21**], more precisely we need a vbariant of their theorems becasue we need to allow the case in which a has zeros on $\mathbf{T}$, hence we will give a proof of the result. Moreover, we follow the discussions in [**14**] and construct the so called *matrix Blaschke-Potapov factors* which degenerate on given points with desired orders and residues.

In Section 5 we study the *soliton solutions*, i.e. the NLFT data $(a, 0)$ where $\bar{a}$ is the boundary value of an inner function. We have encountered this type of data in Section 3. They are not in the spaces $\mathbf{H}$ nor $\mathbf{H}_0^*$ but in their completions. We will characterize some of them as being the NLFT data of at least one but not necessarily a unique full line potential. We conjecture that all soliton solutions are in the range of the NLFT.

In fact, the exact formula for all inverse potentials of $(B^*, 0)$ is derived, where B is a finite Blaschke product. Also we prove that any half line rational data $(a_+, b_+) \in \mathbf{H}$ ($(a_-, b_-) \in \mathbf{H}_0^*$), after finitely many steps of layer stripping if necessary, can be paired with another half line rational data $(a_-, b_-) \in \mathbf{H}_0^*$ ($(a_+, b_+) \in \mathbf{H}$) so that $(a_-, b_-)(a_+, b_+)$ is a soliton solution.

Several open problems remain. In particular it would be desireable to develop a theory for the $SU(2)$ NLFT parallel to the $SU(1, 1)$ theory in Chapter 2 of this book.

3.2. $SU(2)$ NLFT on Finite Sequences and $l^1(\mathbf{Z}, \mathbf{C})$

We begin by defining the $SU(2)$ nonlinear Fourier transform on $l_0(\mathbf{Z}, \mathbf{C})$, the set of sequences of complex numbers with finite support.

Let $(F_n) \in l_0(\mathbf{Z}, \mathbf{C})$ be given. For a complex parameter z on the torus $\mathbf{T}$, we define the following recursion:

$$(3.1)\quad \begin{pmatrix} a_n & b_n \\ -\bar{b}_n & \bar{a}_n \end{pmatrix} = \frac{1}{(1+|F_n|^2)^{1/2}} \begin{pmatrix} a_{n-1} & b_{n-1} \\ -\bar{b}_{n-1} & \bar{a}_{n-1} \end{pmatrix} \begin{pmatrix} 1 & F_n z^n \\ -\bar{F}_n z^{-n} & 1 \end{pmatrix} ,$$

$$(3.2)\quad a_{-\infty} = 1,\ b_{-\infty} = 0 .$$

Here $a_{-\infty} = 1$ and $b_{-\infty} = 0$ are to be interpreted that a_n and b_n are constant equal 0 for sufficiently small n, more precisely for n to the left of the support of the sequence (F_n).

For each $z \in \mathbf{T}$ the matrices

$$\begin{pmatrix} a_n(z) & b_n(z) \\ -\bar{b}_n(z) & \bar{a}_n(z) \end{pmatrix}$$

are products of matrices in $SU(2)$ and therefore themselves in $SU(2)$. To abbreviate notation we will only write the first row $(a_n(z), b_n(z))$ to denote the $SU(2)$ matrices. Note that both a and b are finite Laurent polynomials and are in particular rational functions and have holomorphic extensions to $\mathbf{C} \setminus \{0\}$.

The $SU(2)$ NLFT of (F_n), denoted by $\widehat{(F_n)}$, is defined as

$$(a_\infty(z), b_\infty(z)) = \lim_{n\to\infty} (a_n(z), b_n(z))$$

for $z \in \mathbf{T}$. Since a_n and b_n eventually remain constant for sufficiently large n, the limiting process is trivial. The following are some basic properties of $\widehat{(F_n)} = (a(z), b(z))$.

Lemma 3.1. *Let (F_n) be a finite sequence and $\widehat{(F_n)} = (a, b)$. Then $b(z)$ is a finite Laurent series with lowest degree N_-, the minimum of the support of (F_n), and highest degree N_+, the maximum of the support of (F_n). On the other hand, $a(z)$ is a Laurent polynomial with lowest degree $N_- - N_+$, highest degree 0, and $a(\infty) = \prod (1 + |F_n|^2)^{-1/2} > 0$. Moreover, $|a(z)|^2 + |b(z)|^2 = 1$ for $z \in \mathbf{T}$.*

The proof is by induction on the length $1 + N_+ - N_-$ of the sequence (F_n) as in the first chapter of this book and is skipped here.

Lemma 3.2. *Suppose $\widehat{(F_n)} = (a, b)$. If (F_{n+1}) denotes the shifted sequence whose n-th entry is F_{n+1}, then $\widehat{(F_{n+1})} = (a, bz^{-1})$. If $|c| = 1$, then $\widehat{(cF_n)} = (a, cb)$.*

This is easily seen by conjugation with $\begin{pmatrix} z^{-1/2} & 0 \\ 0 & z^{1/2} \end{pmatrix}$ or $\begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}$ where $e^{i2\theta} = c$.

Let $f(z)$ be a complex function. We define f^* to be another complex function such that

$$f^*(z) = \overline{f(\frac{1}{\bar{z}})}.$$

Then it is easy to see that $aa^* + bb^* = 1$ if (a, b) is the image of a finite sequence via the NLFT. The next theorem characterizes the image of finite sequences as the collection of all such Laurent polynomials a, b. Moreover, we can prove that the NLFT is injective on finite sequences.

THEOREM 3.3. *The NLFT is bijective from* $l_0(\mathbf{Z}, \mathbf{C})$ *onto the space* $S = \{(a(z), b(z)) : a(z), b(z)$ *are Laurent polynominals such that* $aa^* + bb^* = 1$ *and* $0 < a(\infty) < \infty\}$.

PROOF. Clearly the NLFT maps l_0 sequences into the space S.

Suppose $(a, b) \in S$. By the assumption $0 < a(\infty) < \infty$ we know that a is a polynomial in z^{-1} with constant term > 0. Let N be the degree of this polynomial. Also denote the upper and lower degree of b by N_+ and N_-. Then we have $N = N_+ - N_-$ since $aa^* + bb^* = 1$.

Since the forward NLFT maps upper and lower degree of F to upper and lower degree of b by Lemma 3.1, we may prove bijectivity for fixed N_+ and N_-. By the shift symmetry between F and b stated in Lemma 3.2 we may assume $N_- = 0$.

We then prove bijectivity by induction on $N_+ = N$. For $N = 0$, $(a, b) \in S$ consists of constant functions a and b such that $|a|^2 + |b|^2 = 1$ and $a > 0$, which implies $b \in D$ and a is uniquely determined by b. Observe that the map $F_0 \to b$ defined by $b = F_0(1 + |F_0|^2)^{-1/2}$ is bijective from $\mathbf{C}$ to D. Thus $F \to (a, b)$ is bijective in the case $N = 0$.

Now assume we have proved bijectivity up to upper degree $N - 1$ and proceed to prove it for upper degree N. We first prove injectivity i.e. the sequence F can be recovered from (a, b). We first show that F_0 can be recovered by (a, b). Let F' denote the (unknown) truncated sequence which coincides with F except for $F_0' = 0$, and denote $\widehat{(F')} = (a', b')$. Then

$$(a', b') = (1 + |F_0|^2)^{-1/2}(1, -F_0)(a, b) \tag{3.3}$$

which implies that

$$b'(0) = (1 + |F_0|^2)^{-1/2}(b(0) + F_0 a^*(0)) = 0.$$

Hence

$$F_0 = -\frac{b(0)}{a^*(0)}.$$

Note that this quotient is well defined since $a^*(0) > 0$. Thus F_0 is determined by (a, b) and then (a', b') is derived by (3.3). By induction hypothesis, F' can be recovered by (a', b'). Therefore we have proved injectivity for upper degree N.

Now we prove surjectivity for upper degree $N > 0$. Let $(a, b) \in S$ and assume as before the lower degree and upper degree of the Laurent polynomial b are 0 and N respectively. Let $F_0 = -b(0)/a^*(0)$ and $(a', b') = (1+|F_0|^2)^{-1/2}(1, -F_0)(a, b)$. Then by taking the determinant of the last equation we see that the Laurent polynomials a' and b' satisfy

$$a'a'^* + b'b'^* = 1.$$

Moreover, b' is a polynomial with upper degree at most N and lower degree at least 1, and

$$a'(\infty) = (1 + |F_0|^2)^{1/2} a(\infty).$$

Thus $0 < a'(\infty) < \infty$ and $(a', b'/z) \in S$ where the upper degree of b'/z is less than N. By induction, there is a finite sequence (G_n) such that $\widehat{(G_n)} = (a', b')$. Let (F_n)

be the sequence such that $F_0 = -b(0)/a^*(0)$ and $F_n = G_{n-1}$ for $n > 0$. Then it is clear that $\widehat{(F_n)} = (a, b)$ and we have proved the surjectivity for upper degree N.

$\square$

Remark: The process we used here to produce the inverse potential is called the *layer stripping method* . In Section 3 this method will be used again to construct the inverse NLFT for half line l^2 sequences.

LEMMA 3.4. *Assume we are given any Laurent polynomial $b(z)$ with $|b(z)| \leq 1$ on Π but $|b|$ is not identically 1 on $\mathbf{T}$. For any set Z in D and prescribed multiplicities there is a unique Laurent polynomial a such that $(a, b) \in S$ and the zeros of a^* in D are exactly the ones in Z with the prescribed multiplicities.*

We will prove a more general lemma (see Lemma 3.18) in Section 4 about rational NLFT data and omit the proof of the present lemma.

Now we are ready to extend the NLFT to the first class of infinite sequences, $l^1(Z)$. The Fourier transform of such data will be an $SU(2)$ valued function on $\mathbf{T}$, written as before as (a, b). We will only partially have holomorphic extensions of these functions beyond $\mathbf{T}$.

Define a metric on $SU(2)$ by

$$dist(T, T') = ||T - T'||_{op} \ .$$

Obviously this makes $SU(2)$ a complete metric space because the space of 2 by 2 complex matrices is complete under this metric and $SU(2)$ is a closed subset of it.

Define $L^\infty(\mathbf{T}, SU(2))$ to be the metric space of all essentially bounded functions $A : \mathbf{T} \to SU(2)$ i.e. $\sup_z dist(id, A(z)) < \infty$, with the distance

$$dist(A, A') = \sup_z dist(A(z), A'(z)).$$

In fact, since $SU(2)$ has finite diameter, all measurable functions from $\mathbf{T}$ to $SU(2)$ are in $L^\infty(\mathbf{T}, SU(2))$. Also note that $C(\mathbf{T}, SU(2))$ is a closed subset in $L^\infty(\mathbf{T}, SU(2))$ and thus is complete under this metric.

As in the case of the linear and the SU(1,1) Fourier transform, we can easily extend the defining recursion formula to $l^1(\mathbf{Z})$. Here $l^1(\mathbf{Z})$ denotes the usual space of absolutely summable sequences on $\mathbf{Z}$ with have the usual norm $||F||_{l^1} = \sum |F_n|$.

THEOREM 3.5. *The NLFT on $l_0(Z)$ extends uniquely to a Lipschitz map from $l^1(Z)$ to $C(\mathbf{T}, SU(2))$.*

PROOF. We first derive a Lipschitz estimate on finite sequences. Given two finite sequences F and F', let

$$T_n(z) = \frac{1}{(1 + |F_n|^2)^{1/2}} \begin{pmatrix} 1 & F_n z^n \\ -\bar{F}_n z^{-n} & 1 \end{pmatrix} \text{ where } z \in \mathbf{T}.$$

Since T_n is unitary, $T_n T_n^* = Id$, $||T_n(z)||_{op} = 1$ for all $n \in \mathbf{Z}$, $z \in \mathbf{T}$. By Trotter's formula we have

$$|| \prod T_n(z) - \prod T_n'(z)||_{op} \leq \sum_n ||T_n(z) - T_n'(z)||_{op} \ .$$

Now we claim that $||T_n(z) - T'_n(z)||_{op} \leq 2|F_n - F'_n|$ for all $z \in \mathbf{T}$. But we have

$$||T_n(z) - T'_n(z)||_{op} \leq \left|\frac{1}{(1+|F_n|^2)^{1/2}} - \frac{1}{(1+|F'_n|^2)^{1/2}}\right| + \left|\frac{F_n}{(1+|F_n|^2)^{1/2}} - \frac{F'_n}{(1+|F'_n|^2)^{1/2}}\right|$$

and both terms are less than $|F_n - F_{n'}|$ since the functions $x \mapsto (1+x^2)^{-1/2}$ and $x \mapsto x(1+x^2)^{-1/2}$ are 1-Lipschitz on $\mathbf{R}$.

Hence $||\prod T_n(z) - \prod T'_n(z)||_{op} \leq 2\sum|F_n - F'_n| = 2||F - F'||_{l^1}$ for all $z \in \mathbf{T}$. As $C(\mathbf{T}, SU(2))$ is a complete metric space with the metric $dist(A, A') = \sup_z ||A(z) - A'(z)||_{op}$ and the image of finite sequences is contained in $C(\mathbf{T}, SU(2))$, this Lipschitz estimate says that we can uniquely extend the NLFT to a Lipschitz map from $l^1(Z)$ to $C(\mathbf{T}, SU(2))$.

$\square$

3.3. Extension to half line l^2 sequences

In this section we define the NLFT for half line sequences (in $l^2(Z_{\geq 0})$ or $l^2(Z_{<0})$), characterize the image space, and construct the inverse map. Moreover, we define a metric on the image space so that the NLFT is a homeomorphism, say, between $l^2(Z_{\geq 0})$ and the image space.

As in the case of the linear Fourier transform, for sequences in $l^2(Z)$ the defining recursion formula does not necessarily converge pointwise. Instead, one uses a Plancherel type identity to obtain convergence in a certain sense. Writing

$$\int_{\mathbf{T}} f(z) = \int_0^1 f(e^{2\pi i\theta})d\theta$$

for the average of the function f on $\mathbf{T}$, we have the following *Plancherel identity* for the NLFT.

LEMMA 3.6. *Let (F_n) be a finite sequence. $(a,b) = \widehat{(F_n)}$. We have*

$$-\left(\int_{\mathbf{T}} \log|a(z)|dz + \sum \log|z_k|\right) = \frac{1}{2}\sum_n \log(1+|F_n|^2)\ , \tag{3.4}$$

where (z_k) is the sequence of zeros of a^* in D and higher order zeros appear several times in the sequence according to their multiplicity.

PROOF. Since (a,b) is the NLFT of a finite sequence, a^* is a polynomial in z. If we let B be the Blaschke product formed by the zeros of a^* in D, then a^*/B is an outer function on D (see [**16**] for background on bounded analytic functions). Thus inside D, $\log|a^*/B(z)|$ is the harmonic extension of its boundary value. Also note that $B(0) \neq 0$ since $a^*(0) = (\prod(1+|F_n|^2))^{1/2} \neq 0$. Hence,

$$\int_{\mathbf{T}} \log|a(z)| = \int_{\mathbf{T}} \log\left|\frac{a^*}{B}(z)\right| = \log|a^*(0)| - \log|B(0)|$$
$$= -\frac{1}{2}\sum \log(1+|F_n|^2) - \sum \log|z_i|\ ,$$

where $\{z_i\}$ are the zeros of B. i.e. zeros of a^* in D.

□

Note that in equation (3.4) each term on the left hand side is positive. Hence each term on the left hand side is controlled by the right hand side, which is equivalent to the l^2 norm of (F_n) for (F_n) in a fixed ball about the origin of l^2. As the right hand side is defined for all $(F_n) \in l^2(Z)$, this will help us to extend the NLFT to all $l^2(Z)$ sequences.

Remark: For more general $a^* \in H^\infty(D)$, such as we will encounter for the NLFT of infinite sequences $(F_n) \in l^2$, a^* can be decomposed into $a^* = fBg$ where f is an outer function, B is a Blaschke product, and g is a singular inner function. The Plancherel identity will then be

$$\int_{\mathbf{T}} \log|a^*(z)| = \int_{\mathbf{T}} \log|f(z)| = \log|f(0)| = \log|a^*(0)| - \log|B(0)| - \log|g(0)|.$$

In this case $\log|g(0)|$ takes care of the "singular measure" part because it is just the total measure of σ_g where σ_g is the singular measure on the boundary that generates $\log|g|$.

Now we proceed to describe the range of NLFT on $l^2(Z_{\geq 0})$. Define

(1) The space $\mathbf{L}$ to be the space of all measurable $SU(2)$ matrix valued function (a,b) on $\mathbf{T}$ such that $\bar{a}$ is in the Hardy space $H^1(\mathbf{T})$ and thus has an analytic extension to D denoted by a^*, and $a^*(0)>0$.
(2) The space $\mathbf{H}$ to be the space of $(a,b)\in\mathbf{L}$ such that b is in the Hardy space $H^2(\mathbf{T})$ and thus has an analytic extension to D, which is also denoted by b. Moreover, a^* and b have no common inner factor.
(3) The space $\mathbf{H}_0^*$ to be the space of all $(a,b)\in\mathbf{L}$ such that $\bar{b}$ is in the Hardy space $H^2(D)$ and thus has an analytic extension to D denoted by b^*. Moreover, $b^*(0)=0$ and a^* and b^* have no common inner factor.

Note that the various functions in $H^1(\mathbf{T})$ and $H^2(\mathbf{T})$ in the above definitions are really bounded by 1 and thus in $H^\infty(\mathbf{T})$. The special choice of Hardy space exponents above is in reference to a convenient choice of metric used below.

Saying that a^* and b have no common inner factor means that there is no non-trivial inner function s such that a^*/s and b/s are bounded. Alternatively, consider for $f\in H^p(D)$, $p>0$, the factorization

$$f(z)=\exp\left(ic+\frac{1}{2\pi}\int_0^{2\pi}\frac{e^{it}+z}{e^{it}-z}\log|f(e^{it})|dt\right)$$
$$B_f(z)\exp\left(\frac{-1}{2\pi}\int_0^{2\pi}\frac{e^{it}+z}{e^{it}-z}d\sigma_f(t)\right)\ ,$$

where $B_f(z)$ is the Blaschke product consisting of zeros of $f(z)$, σ_f is a nonnegative singular measure, and c is a constant. Under these notations, a^* and b have no common inner factor if a^* and b have no common zeros and σ_{a*} and σ_b are mutually singular.

Note that if $(a,b)\in\mathbf{H}$ then the meromorphic function b/a^* on D will uniquely determine a^* and b. Namely, the inner parts of the functions of b and a^* can be determined since they have no common inner factor. The outer functions can be determined from the absolute values $|a|$ and $|b|$, which can be determined from $|b|/|a|$ and the identity $|a|^2+|b|^2=1$. Traditionally, b/a^* is called the *reflection coefficient*.

We consider the metric

$$\hat{d}\left((a,b),(a',b')\right)=\int_{\mathbf{T}}|a-a'|+\left(\int_{\mathbf{T}}|b-b'|^2\right)^{1/2}+|\log a^*(0)-\log(a')^*(0)|$$

on $\mathbf{L},\mathbf{H}$, and $\mathbf{H}_0^*$.

It is easy to see that $\mathbf{L}$ is complete under $\hat{d}$. Namely, the map $(a,b)\to(\overline{a},b,\log(a^*(0)))$ maps into a closed subset of $H^1(\mathbf{T})\times L^2(\mathbf{T})\times\mathbf{R}$ because its range is given by all triples (a,b,c) satisfying the closed conditions $|a|^2+|b|^2=1$ almsot everywhere and $\int_{\mathbf{T}}\overline{a}\geq 0$ and $\log\int_{\mathbf{T}}\overline{a}=c$. However, $\mathbf{H},\mathbf{H}_0^*$ are not complete under $\hat{d}$ because in general a^* and b may have a common inner factor when (a,b) is in the closure of $\mathbf{H}$. In the end of this section we will discuss the completion of $\mathbf{H}$. First, despite the incompleteness, we prove that the NLFT is a homeomorphism from $l^2(Z_{\geq 0})$ to $(\mathbf{H},\hat{d})$ and from $l^2(Z_{<0})$ to $(\mathbf{H}_0^*,\hat{d})$.

First we show that the NLFT of truncations of l^2 sequences converge in the metric of $\mathbf{L}$.

Lemma 3.7. *Let F be a sequence in $l^2(Z_{\geq 0})$ and let $F_{\leq n}$ denote the truncations to $[0,n]$. Then $(a_n,b_n)=\widehat{F_{\leq n}}$ is a Cauchy sequence in the metric space $(\mathbf{H},\hat{d})$.*

With this lemma we can define $\widehat{F} = (a,b)$ as a limit in $\mathbf{L}$, which is complete and contains $\mathbf{H}$. Later we need to do some work to prove that $(a,b) \in \mathbf{H}$ because $(\mathbf{H}, \hat{d})$ is not complete.

To prove Lemma 3.7, we first prove the following auxiliary lemma:

LEMMA 3.8. *If (a,b) is the NLFT of a a finite sequence (F_n), then*

$$\hat{d}((a,b),(1,0)) \leq C_1 \sum \log(1+|F_n|^2) + C_2 \left(\sum \log(1+|F_n|^2)\right)^{1/2} ,$$

where C_1, C_2 are fixed constants.

PROOF.

$$\begin{aligned}\hat{d}\,((a,b),(1,0)) &= \int |a-1| + \left(\int |b|^2\right)^{1/2} + |\log a^*(0)| \\ &\leq \int (1-|a|) + \int \left|\frac{a}{|a|} - 1\right| + \left(\int |b|^2\right)^{1/2} + |\log a^*(0)| \ .\end{aligned}$$

As (a,b) comes from a finite sequence, Lemma 3.1 says that $|\log a^*(0)| = 1/2 \sum \log(1+|F_n|^2)$, which is the desired estimate for this term. Using that a^* is a polynomial in z, we can write $a^*|_D = f_a B_a$ where f_a is an outer function, B_a is a finite Blaschke product and $f_a(0), B_a(0) > 0$. We claim that

$$\int \left|\frac{a}{|a|} - 1\right|^2 \leq C\left(\int |\log|a|| + \int |B_a - 1|^2\right)$$

for all (a,b) in the range of finite sequence. Before proving the claim we proceed to bound $\int |\log|a||$, $\int(1-|a|)$, $\int |b|^2$, and $\int |B_a - 1|^2$ by $\sum \log(1+|F_n|^2)$. The bound for the logarithmic integral is clear from the Plancherel identity. Observe that for $0 \leq x \leq 1$, $1 - x \leq |\log x|$. Thus,

$$\int (1-|a|) \leq \int |\log|a|| \text{ and } \int |b|^2 = \int (1-|a|^2) \leq 2\int |\log|a|| \ .$$

This proves the desired estimate for these integrals. Next,

$$\int |B_a - 1|^2 = \int (2 - 2\mathrm{Re}B_a) = 2(1 - \mathrm{Re}B_a(0)) \ .$$

As $B_a(0) > 0$, $\mathrm{Re}B_a(0) = B_a(0) = \prod |z_k|$ where $\{z_k\}$ are the zeros of B_a in D, i.e. zeros of a^* in D. Hence, $1 - \mathrm{Re}B_a(0) \leq |\log B_a(0)| = -\sum \log|z_k|$ and

$$\int_{\mathbf{T}} |B_a - 1|^2 \leq -2\sum \log|z_k|.$$

Finally, by Hölder's inequality and the Plancherel identity we conclude that $\hat{d}((a,b),(1,0)) \leq C_1 \sum \log(1+|F_n|^2) + C_2(\sum \log(1+|F_n|^2))^{1/2}$.

Now we prove the claim.The function f_a is outer and hence $f_a(0) > 0$ implies that on the circle

$$\frac{f_a}{|f_a|} = e^{ig}$$

where $g = p.v. \int_{\mathbf{T}} \log|a(\zeta)| \mathrm{Im}(\frac{\zeta+z}{\zeta-z})$, the Hilbert transform of $\log|a|$.

$$\int |f_a/|f_a| - 1|^2 = 2\int_0^2 t|\{|f_a/|f_a| - 1| > t\}|dt \leq 2\int_0^2 t|\{|g| > t\}|dt$$

By the weak type 1 bound for the Hilbert transform, $|\{|g| > t\}| \leq C||\log |a|||_{L^1}/t$. Thus

$$\int |f_a/|f_a| - 1|^2 \leq C||\log |a|||_{L^1}.$$

Since $|a/|a| - 1| = |\bar{a}/|a| - 1| = |B_a f_a/|f_a| - 1| \leq |f_a/|f_a| - 1| + |B_a - 1|$,

$$\begin{aligned}\int |a/|a| - 1|^2 &\leq 2\left(\int |f_a/|f_a| - 1|^2 + \int |B_a - 1|^2\right)\\ &\leq 2C\int |\log |a|| + 2\int |B_a - 1|^2.\end{aligned}$$

We have proved the claim.

□

Now we proceed to prove Lemma 3.7.

Proof. Observe that $\{a_n\}$ are polynomials in z^{-1}, and $\{b_n\}$ are polynomials in z. Thus, a_n^* and b_n have no singular inner factor. Furthermore, a_n^* and b_n have no common zeros in D because $a_n a_n^* + b_n b_n^* = 1$. Hence, a_n^* and b_n have no common inner factor and as a result $(a_n, b_n) \in \mathbf{H}$ for all n.

Let $G = (a, b)$, $G' = (a', b')$ be $SU(2)$ valued functions. Then $GG' = (aa' - b\bar{b}', ab' + b\bar{a}')$. If we assume that $\bar{b}b'$ is the boundary value of an analytic function on D which vanishes at 0, then

$$\hat{d}(GG', G) = \int_{\mathbf{T}} |aa' - b\bar{b}' - a| + \left(\int_{\mathbf{T}} |ab' + b\bar{a}' - b|^2\right)^{1/2} + |\log(a')^*(0)|.$$

Since

$$\int |aa' - b\bar{b}' - a| \leq \int |a||a' - 1| + \int |bb'| \leq \int |a' - 1| + \int |b'|$$

$$\text{and } ||ab' + b\bar{a}' - b||_{L^2(\mathbf{T})} \leq ||b'||_{L^2} + ||a' - 1||_{L^2},$$

with Hölder's inequality and the fact that $|a'| \leq 1$, it is easy to see

$$\hat{d}(GG', G) \lesssim \hat{d}(G', id) + (\hat{d}(G', id))^{1/2}\ .$$

Let $(a_{(n,m]}, b_{(n,m]}) = \overbrace{F_{(n,m]}}$ where $F_{(n,m]}$ is the restriction of F to the interval $(n, m]$. Then $\hat{d}((a_m, b_m), (a_n, b_n)) = \hat{d}((a_n, b_n)(a_{(n,m]}, b_{(n,m]}), (a_n, b_n))$ and $\bar{b}_n b_{(n,m]}$ is a polynomial with lowest degree ≥ 1. Thus, from the previous inequality and the auxiliary lemma, $\hat{d}((a_m, b_m), (a_n, b_n)) \lesssim \sum_{k=n+1}^m \log(1 + |F_k|^2) + (\sum_{k=n+1}^m \log(1 + |F_k|^2))^{1/4}$. Hence, $\{(a_n, b_n)\}$ is a Cauchy sequence in $\mathbf{H}$.

□

Lemma 3.9. *NLFT is injective on* $l^2(Z_{\geq 0})$.

Proof. Suppose $(a, b) \in \mathbf{L}$ and $(a, b) = \overbrace{F_n}$ for some $(F_n) \in l^2(Z_{\geq 0})$. We show that (F_n) can be determined by a and b and thus prove the injectivity.

As usual, we let $(a_n, b_n) = \overbrace{F_{\leq n}}$. From our discussion of finite sequences, $F_0 = b_n/a_n^*(0)$ for all n. Since b_n converges to b in $H^2(D)$, we have $\lim b_n(0) = b(0)$. Similarly, $\lim a_n^*(0) = a^*(0)$ because a_n^* converges to a^* in $H^1(D)$. Moreover, $a_n^*(0) = (\prod_{k=0}^n (1 + |F_k|^2))^{-1/2} \to (\prod_{k=0}^\infty (1 + |F_k|^2))^{-1/2} = a^*(0) > 0$. Hence

$F_0 = b_n/a_n^*(0) = \lim b_n/a_n^*(0) = b/a^*(0)$ i.e. F_0 is determined by a^* and b.

Let $(\tilde{F}_n)$ be the layer stripped sequence i.e. $\tilde{F}_n = F_n$ for $n > 0$ and $\tilde{F}_0 = 0$. Define $(\tilde{a}, \tilde{b}) = \overbrace{(\tilde{F}_n)}$. We have

$$(a_n, b_n) = \frac{1}{(1+|F_0|^2)^{1/2}} \begin{pmatrix} 1 & F_0 \\ -\bar{F}_0 & 1 \end{pmatrix} (\tilde{a}_n, \tilde{b}_n). \text{ Thus,}$$

$$\tilde{a}_n^* = \frac{1}{(1+|F_0|^2)^{1/2}} (a_n^* + \bar{F}_0 b_n) \ , \tag{3.5}$$

$$\tilde{b}_n = \frac{1}{(1+|F_0|^2)^{1/2}} (-F_0 a_n^* + b_n) \ . \tag{3.6}$$

As $n \to \infty$, $\tilde{a}_n^* \to \tilde{a}^*$, $a_n^* \to a^*$ in $H^1(D)$, and $\tilde{b}_n \to \tilde{b}$, $b_n \to b$ in $H^2(D)$. Because $|a_n^*|$ and $|b_n|$ are bounded by 1 and $\mathbf{T}$ has finite measure, we conclude that $a_n^* \to a^*$ and $b_n \to b$ in $H^p(D)$ for any $p \geq 1$. Thus, the left hand side of (3.5) converges to $\tilde{a}^*$ in $H^1(D)$ and its right hand side converges to $(1+|F_0|^2)^{-1/2}(a^* + \bar{F}_0 b)$ in $H^1(D)$. Therefore,

$$\tilde{a}^* = \frac{1}{(1+|F_0|^2)^{1/2}} (a^* + \bar{F}_0 b).$$

Similarly,

$$\tilde{b} = \frac{1}{(1+|F_0|^2)^{1/2}} (-F_0 a^* + b).$$

As a result, $(\tilde{a}^*, \tilde{b})$ can be determined by (a^*, b). By induction, we can determine F_n for all n. Hence, NLFT is injective on $l^2(Z_{\geq 0})$.

□

Remark : This proof also implies the following observation: Suppose $(F_n) \in l^2(Z_{\geq 0})$. Let $F_{\leq n}$ and $F_{>n}$ be the restrictions to $[0, n]$ and $[n+1, \infty)$ and $(a_n, b_n) = \overbrace{F_{\leq n}}$, $(a_{>n}, b_{>n}) = \overbrace{F_{>n}}$. Then

$$(a, b) = (a_n, b_n)(a_{>n}, b_{>n}) \ .$$

The base case $n = 0$ is shown in the proof of the lemma. Then we can use induction to prove this for all n.

The proof of the lemma shows in particular that the layer stripping method produces the *inverse NLFT* of data (a, b) in the range of the NLFT of sequences in $l^2(Z_{\geq 0})$. The next lemma shows that we can apply the layer stripping method on a more general class of (a^*, b) and obtain an l^2 sequence (F_n). However, $\overbrace{(F_n)}$ may not be (a, b). If $\overbrace{(F_n)} \neq (a, b)$, then (a, b) can not be in the range of $l^2(Z_{\geq 0})$.

LEMMA 3.10. *Given any $a^* \in H^\infty(D)$ and $b \in H^\infty(D)$ such that $|a^*|^2 + |b|^2 = 1$ on $\mathbf{T}$ and $a^*(0) > 0$. We can apply the layer stripping method on a^* and b and obtain a sequence $(F_n) \in l^2(Z_{\geq 0})$ with*

$$\prod_{k=0}^{\infty} (1 + |F_k|^2) \leq \frac{1}{(a^*(0))^2}.$$

Later we will see that the equality holds if and only if $\overbrace{(F_n)} = (a, b)$.

Proof. Applying the layer stripping method on a^* and b, we obtain $F_0 = b/a^*(0)$,

$$a^*_{\geq 1} = \frac{1}{(1+|F_0|^2)^{1/2}}(a^* + \bar{F}_0 b) \ , \ b_{\geq 1} = \frac{1}{z(1+|F_0|^2)^{1/2}}(b - F_0 a^*)$$

It is obvious that $a^*_{\geq 1}$ and $b_{\geq 1}$ also satisfy all the assumptions and

$$a^*_{\geq 1}(0) = \frac{1}{(1+|F_0|^2)^{1/2}}(a^*(0) + a^*(0)|F_0|^2) = (1+|F_0|^2)^{1/2} a^*(0)$$

In general, after $n+1$ steps we have

$$1 \geq a^*_{\geq n+1}(0) = (1+|F_n|^2)^{1/2} a^*_{\geq n}(0) = (\prod_{k=0}^{n}(1+|F_k|^2))^{1/2} a^*(0).$$

Thus,

$$\prod_{k=0}^{n}(1+|F_k|^2) \leq \frac{1}{(a^*(0))^2} \text{ for all n}$$

which says $(F_n) \in l^2(Z_{\geq 0})$ and $\prod_{n=0}^{\infty}(1+|F_n|^2) \leq (a^*(0))^{-2}$. □

Now we show that the range of $l^2(Z_{\geq 0})$ lies in $\mathbf{H}$.

Lemma 3.11. *If $(a,b) = \widehat{(F_n)}$ for some $(F_n) \in l^2(Z_{\geq 0})$, then $(a,b) \in \mathbf{H}$*

Proof. We already know that $(a,b) \in \mathbf{L}$. Since b is the limit in $L^2(\mathbf{T})$ of the elements $b_{\leq n} \in H^2(\mathbf{T})$, we have $b \in H^2(\mathbf{T})$. It remains to show that a^* and b^* have no common inner factor. Suppose to get a contradiction that they do have a common inner factor g. Then $a^*/g \in H^1(D)$ and $b/g \in H^2(D)$. It is easy to see that the layer stripping method applied $a^*/g, b/g$ produces the same potential as for a^*, b, namely the layer stripped data are the functions $a^*_{\geq n}$ and $b^*_{\geq n}$ divided by g. Lemma 3.9 shows that the potential obtained must be (F_n).

The previous lemma says that

$$\prod_{k=0}^{\infty}(1+|F_k|^2) \leq \frac{(g(0))^2}{(a^*(0))^2} < \frac{1}{(a^*(0))^2}.$$

However, since $(a,b) = \widehat{(F_n)}$,

$$a^*(0) = \lim_{n\to\infty} a^*_n(0) = \lim_{n\to\infty}(\prod_{k=0}^{n}(1+|F_k|^2))^{-1/2}$$

i.e.

$$\prod_{k=0}^{\infty}(1+|F_k|^2) = \frac{1}{(a^*(0))^2}$$

and we get a contradiction. Therefore a^* and b have no common inner factor. □

Finally, let us establish the range of the NLFT on $l^2(Z_{\geq 0})$.

Lemma 3.12. *NLFT is surjective from $l^2(Z_{\geq 0})$ to $\mathbf{H}$.*

PROOF. Given $(a,b)\in \mathbf{H}$, we apply the layer stripping method on it and obtain an l^2 sequence (F_n), $n\geq 0$. Let $(\tilde{a},\tilde{b})=\overbrace{(F_n)}$. We will prove that $(\tilde{a},\tilde{b})=(a,b)$.

According to the remark after Lemma 3.9,

$$(\tilde{a},\tilde{b})=\overbrace{(F_{\leq N})}\overbrace{(F_{>N})}=(\tilde{a}_N,\tilde{b}_N)(\tilde{a}_{>N},\tilde{b}_{>N})$$

for every nonnegative integer N. And $\tilde{b}_{>N}$ vanishes at 0 with order $>N$.

On the other hand, the layer stripping method says that

$$(a,b)=\overbrace{(F_{\leq N})}(a_{>N},b_{>N}) \tag{3.7}$$

where $(a_{>N},b_{>N})$ satisfies all the conditions described in $\mathbf{H}$ except that $a^*_{>N}$ and $b_{>N}$ might have common inner factor. Moreover, $b_{>N}$ vanishes at 0 with order $>N$. Hence

$$(\tilde{a},\tilde{b})^{-1}(a,b)=(\tilde{a}_{>N},\tilde{b}_{>N})^{-1}(a_{>N},b_{>N}) \text{ for all } N.$$

Note that the $(1,2)$ component of the right hand side is

$$\tilde{a}^*_{>N}b_{>N}-\tilde{b}_{>N}a^*_{>N}$$

which is an analytic function on D and vanishes at 0 with order $>N$. Because the expression is independent of N, we conclude that all Taylor coefficients at 0 vanish. Thus,

$$\tilde{a}^*_{>N}b_{>N}-\tilde{b}_{>N}a^*_{>N}=0 \text{ for all } N.$$

Especially,

$$b_{>N}=\frac{\tilde{b}_{>N}}{\tilde{a}^*_{>N}}a^*_{>N}\ . \tag{3.8}$$

Since $(\tilde{a}_{>N},\tilde{b}_{>N})$ comes from a tail of $(F_n)\in l^2(Z_{\geq 0})$, $||\tilde{b}_{>N}||_{L^2(\mathbf{T})}$ and $||\tilde{a}_{>N}-1||_{L^1(\mathbf{T})}$ are very small when N is large enough. Hence, given any $\epsilon>0$, we can find N_0 such that for all $N>N_0$,

$$\left|\frac{\tilde{b}_{>N}}{\tilde{a}^*_{>N}}\right|<\epsilon \text{ on } \mathbf{T} \text{ except on a set of measure } \leq\epsilon.$$

Moreover, $|b_{>N}|$ and $|a^*_{>N}|\leq 1$ a.e. on $\mathbf{T}$. Then from (3.8) we can easily see that

$$b_{>N}\to 0 \text{ in } H^2(D)$$

We claim that $a^*_{>N}\to g$ in $H^1(D)$ for some function g, the proof of this claim is postponed. Assuming the claim, g is an inner function. This follows from $|g|=1$ a.e. on $\mathbf{T}$, which in turn follows from $|a_{>N}|^2+|b_{>N}|^2=1$ on $\mathbf{T}$ for all N and $b_{>N}\to 0$ in $H^2(D)$.

By (3.7),

$$a^*=-\tilde{b}^*_N b_{>N}+\tilde{a}^*_N a^*_{>N}. \tag{3.9}$$

The term $\tilde{b}^*_N b_{>N}$ is analytic in D and on $\mathbf{T}$ its L^2 norm is smaller or equal to the L^2 norm of $b_{>N}$. Thus $\tilde{b}^*_N b_{>N}\to 0$ in $H^2(D)$ (also in $H^p(D)$ for all $p\geq 1$). Since $\tilde{a}^*_N\to\tilde{a}^*$, $a^*_{>N}\to g$ in $H^1(D)$ and $|\tilde{a}^*_N|,|a^*_{>N}|\leq 1$. Therefore, $\tilde{a}^*_N a^*_{>N}\to\tilde{a}^*g$ in $H^1(D)$ (also in $H^p(D)$ for all $p\geq 1$). Hence as $N\to\infty$, the right hand side of equation (3.9) goes to $\tilde{a}^*g$ in $H^1(D)$, and we conclude that $a^*=\tilde{a}^*g$. Similarly, $b=\tilde{b}g$.

If g is constant, then $g = 1$ (since $|g| = 1$ and $a^*(0), \tilde{a}^*(0) > 0$). In this case $(a,b) = (\tilde{a}, \tilde{b}) = \overbrace{(F_n)}$ i.e. (a,b) is in the range of $l^2(Z_{\geq 0})$. Otherwise, a^* and b have common inner factor which contradicts to the assumption $(a,b) \in \mathbf{H}$. Therefore we have proved the lemma.

Now we prove the claim. From equation (3.7),

$$\overbrace{(F_{(n,m]})}(a_{>m}, b_{>m}) = (a_{>n}, b_{>n})$$

for all $m > n$. Hence

$$a^*_{>n} = \tilde{a}^*_{(n,m]} a^*_{>m} - \tilde{b}^*_{(n,m]} b_{>m} \ ,$$
$$|a^*_{>n} - a^*_{>m}| \leq |a^*_{>m}||\tilde{a}^*_{(n,m]} - 1| + |\tilde{b}^*_{(n,m]} b_{>m}| \leq |\tilde{a}^*_{(n,m]} - 1| + |\tilde{b}^*_{(n,m]} b_{>m}| \ .$$

Since $(\tilde{a}_{(n,m]}, \tilde{b}_{(n,m]}) = \overbrace{(F_{(n,m]})}$, $||\tilde{a}^*_{(n,m]} - 1||_{L^1(\mathbf{T})} \to 0$ as $n, m \to \infty$. Moreover, $\tilde{b}^*_{(n,m]} b_{>m}$ is analytic in D and $||\tilde{b}^*_{(n,m]} b_{>m}||_{L^1(\mathbf{T})} \leq ||b_{>m}||_{L^1(\mathbf{T})}$. The later term goes to zero because $b_{>m} \to 0$ in $H^2(D)$. Hence $(a^*_{>n})$ is a Cauchy sequence in $H^1(D)$ and we have proved the claim

$\square$

This proof also shows the following fact:

LEMMA 3.13. *Given any $a^* \in H^\infty(D)$ and $b \in H^\infty(D)$ such that $|a^*|^2 + |b|^2 = 1$ on $\mathbf{T}$ and $a^*(0) > 0$. Let (F_n) be the $l^2(Z_{\geq 0})$ sequence produced by applying the layer stripping method on (a,b). Then $\overbrace{(F_n)}(g^*, 0) = (a,b)$ where g is the common inner factor of a^* and b.*

In particular, we have the identity

$$\prod_{k=0}^{\infty} (1 + |F_k|^2) = \frac{(g(0))^2}{(a^*(0))^2}.$$

Hence $\prod_{k=0}^{\infty}(1 + |F_k|^2) = 1/(a^*(0))^2$ if and only if $g(0) = 1$ which means g must be the constant 1 and $\overbrace{(F_n)} = (a,b)$.

The following two lemmas prove the continuity of the NLFT and the inverse NLFT.

LEMMA 3.14. *The NLFT is a continuous map from $l^2(Z_{\geq 0})$ to $\{\mathbf{H}, \hat{d}\}$.*

PROOF. Given $F \in l^2(Z_{\geq 0})$ and any $\epsilon > 0$, we will find $\delta > 0$ such that $\hat{d}\left(\overbrace{F}, \overbrace{F'}\right) \leq \epsilon$ for all F' with $||F - F'||_{l^2} \leq \delta$.

Again we let $(a,b) = \overbrace{F}$, $(a', b') = \overbrace{F'}$, and $(a_n, b_n) = \overbrace{F_{\leq n}}$. We write

$$\hat{d}((a,b),(a',b')) \leq \hat{d}((a,b),(a_n,b_n)) + \hat{d}((a_n,b_n),(a'_n,b'_n)) + \hat{d}((a'_n,b'_n),(a',b'))$$

From Lemma 3.7 and its proof we know that

$$\hat{d}((a_n,b_n),(a,b)) \leq C \left\{ \sum_{k>n} \log(1 + |F_k|^2) + \left(\sum_{k>n} \log(1 + |F_k|^2) \right)^{1/4} \right\}$$

for all $(a,b) \in \mathbf{H}$, where C is a fixed constant i.e. $\hat{d}((a_n,b_n),(a,b))$ are uniformly bounded depending only on the l^2 norm of the tail.

Therefore, we first choose N large such that $\hat{d}((a_N, b_N), (a, b)) \leq \epsilon/3$. Then we choose $\delta > 0$ such that for all F' with $||F - F'||_{l^2} \leq \delta$, $||F'_{>N}||$ is still small enough so that $\hat{d}((a'_N, b'_N), (a', b')) \leq \epsilon/3$. Moreover, when δ is small, the l^1 norm of $F_{\leq N} - F'_{\leq N}$ is small. Thus, from the results of l^1 sequences, $||a_n - a'_n||_{L^\infty}$ and $||b_n - b'_n||_{L^\infty}$ are small. Hence, we can choose δ such that the middle term $\hat{d}((a_n, b_n), (a'_n, b'_n))$ is less than $\epsilon/3$.

□

Remark : This proof does not give us uniform continuity, and one may pose the question whether the NLFT is uniformly continuous in the specified metrics.

Next we show that the inverse NLFT from $\{\mathbf{H}, \hat{d}\}$ to $\{l^2(Z_{\geq 0}), d\}$ is continuous where the quasi-metric on $l^2(Z_{\geq 0})$ is defined as

$$d((F_n), (F'_n)) = \sum_{n=0}^{\infty} \log(1 + |F_n - F'_n|^2).$$

It is easy to check that $d((F_n), (G_n)) = 0$ if and only if $(F_n) = (G_n)$ and the modified triangle inequality

$$d((F_n), (G_n)) \leq 2d((F_n), (H_n)) + 2d((H_n), (G_n))$$

holds. Moreover, $\{l^2(Z_{\geq 0}), d\}$ is complete.

Locally d is equivalent to the usual l^2 norm. Hence it is also true that the inverse NLFT from $\{\mathbf{H}, \hat{d}\}$ to $\{l^2(Z_{\geq 0}), ||\cdot||_{l^2}\}$ is continuous. However, with the quasi-metric d the proof will be easier.

Lemma 3.15. *The inverse NLFT from $\{\mathbf{H}, \hat{d}\}$ to $\{l^2(Z_{\geq 0}), d\}$ is continuous.*

Proof. The inverse NLFT is given by the layer stripping method. Given $(a, b) \in \mathbf{H}$, for any $(a', b') \in \mathbf{H}$ with $\hat{d}((a, b), (a', b'))$ sufficiently small we have

$$F_0 = \frac{b(0)}{a^*(0)} = \frac{\int_{\mathbf{T}} b}{\int_{\mathbf{T}} a^*} \text{ is close to } F'_0 = \frac{\int_{\mathbf{T}} b'}{\int_{\mathbf{T}} (a')^*}.$$

Moreover, the new data after the first step of layer stripping,

$$a^*_{\geq 1} = \frac{\bar{F}_0 b + a^*}{(1 + |F_0|^2)^{1/2}} \text{ , } b_{\geq 1} = \frac{b - F_0 a^*}{z(1 + |F_0|^2)^{1/2}} \text{ ,}$$

is very close to

$$(a'_{\geq 1})^* = \frac{\bar{F}'_0 b' + (a')^*}{(1 + |F'_0|^2)^{1/2}} \text{ , } b'_{\geq 1} = \frac{b' - F'_0 (a')^*}{z(1 + |F'_0|^2)^{1/2}}$$

in $\{\mathbf{H}, \hat{d}\}$ if $\hat{d}((a, b), (a', b'))$ is sufficiently small. Moreover, $|\int_{\mathbf{T}} (a'_{\geq 1})^*| = |(a'_{\geq 1})^*(0)| \geq |(a')^*(0)| > 0$. Hence,

$$F_1 = \frac{\int_{\mathbf{T}} b_{\geq 1}}{\int_{\mathbf{T}} a^*_{\geq 1}} \text{ is close to } F'_1.$$

And by induction, for any fixed N_0 and all $(a', b') \in \mathbf{H}$ with $\hat{d}((a, b), (a', b'))$ sufficiently small depending on (a, b) and N_0 we have $\sup_{n \leq N_0} |F_n - F'_n|$ is as small as desired.

Now given $\epsilon > 0$, let N_0 be the positive integer such that $\sum_{n>N_0} \log(1+|F_n|^2) \leq \epsilon/8$. We write

$$d((F_n),(F_n')) = \sum_{n=0}^{\infty} \log(1+|F_n - F_n'|^2) \tag{3.10}$$

$$\leq \sum_{n=0}^{N_0} \log(1+|F_n - F_n'|^2) + 2\sum_{n>N_0} \log(1+|F_n|^2) + 2\sum_{n>N_0} \log(1+|F_n'|^2)$$

Choose $\epsilon/16 > \delta > 0$ such that for all $\hat{d}((a,b),(a',b')) < \delta$, the first term in the above inequality is smaller than $\epsilon/8$, also

$$\left|\sum_{n=0}^{N_0} \log(1+|F_n|^2) - \sum_{n=0}^{N_0} \log(1+|F_n'|^2)\right| < \epsilon/16 \text{ , and}$$

$$\begin{aligned}\left|\sum_{n=0}^{\infty} \log(1+|F_n|^2) - \sum_{n=0}^{\infty} \log(1+|F_n'|^2)\right| &= 2|\log a^*(0) - \log(a')^*(0)| \\ &\leq 2\hat{d}((a,b),(a',b')) < \epsilon/8.\end{aligned}$$

Then the tail $\sum_{n>N_0} \log(1+|F_n'|^2)$ is also small since

$$\begin{aligned}&\sum_{n>N_0} \log(1+|F_n'|^2) \\ &= \sum_{n=0}^{\infty} \log(1+|F_n'|^2) - \sum_{n=0}^{N_0} \log(1+|F_n'|^2) \\ &\leq \left|\sum_{n=0}^{\infty} \log(1+|F_n|^2) - \sum_{n=0}^{\infty} \log(1+|F_n'|^2)\right| \\ &\quad + \sum_{n>N_0} \log(1+|F_n|^2) + \left|\sum_{n=0}^{N_0} \log(1+|F_n|^2) - \sum_{n=0}^{N_0} \log(1+|F_n'|^2)\right| \\ &< \epsilon/8 + \epsilon/8 + \epsilon/16 = \frac{5}{16}\epsilon.\end{aligned}$$

Thus, by (3.10)

$$d((F_n),(F_n')) < \frac{\epsilon}{8} + \frac{2\epsilon}{8} + \frac{5\epsilon}{8} = \epsilon$$

for all $(a',b') \in \mathbf{H}$ such that $\hat{d}((a,b),(a',b')) \leq \delta$.

□

Remark : We have shown that the NLFT is a homeomorphism between $\{\mathbf{H}, \hat{d}\}$ and $l^2(Z_{\geq 0})$ (with either the metric d or the usual l^2 norm). But note that the space $\{\mathbf{H}, \hat{d}\}$ is not complete while $l^2(Z_{\geq 0})$ is complete under both metrics. Hence the inverse NLFT can not be uniformly continuous.

Similarly we can prove that the NLFT is a homeomorphism between $l^2(Z_{<0})$ and $\{\mathbf{H}_0^*, \hat{d}\}$.

Next, we go a little further to study the completion of $\mathbf{H}$.

Let $\overline{\mathbf{H}}$ be the closure of $\mathbf{H}$ in the space $\{\mathbf{L}, \hat{d}\}$ Then,

$$\begin{aligned}\overline{\mathbf{H}} &= \{(a,b) : (a,b) \in \mathbf{L} \text{ and } b \text{ has an analytic exension on } D\ \} \\ &= \{(a\bar{g}, bg) : (a,b) \in \mathbf{H} \text{ and } g \in G\}\end{aligned}$$

where G is the collection of all inner functions which are positive at 0.

It is trivial that if (a,b) is in $\overline{\mathbf{H}}$ then b is the boundary value of an H^2 or even H^∞ function. On the other hand, it is easy to see that $(aB^*, bB) \in \overline{\mathbf{H}}$ for all $(a,b) \in \mathbf{H}$ and Blaschke products B with $B(0) > 0$, because we can approximate (aB^*, bB) by (aB^*, bB_n) where B_n is a Blaschke product with zero set disjoint from but close to the zero set of B. Moreover Frostman's Theorem (see [**16**]) says that the set of Blaschke products is dense in the set of inner functions under H^∞ norm. Hence $(a\bar{g}, bg)$ lies in $\overline{\mathbf{H}}$ for all $(a,b) \in \mathbf{H}$ and g an inner function such that $g(0) > 0$.

In a later section we discuss *"soliton data"*, which are data of the type $(\bar{g}, 0)$ where g is an inner function. For example, we will prove that for every Blaschke product B with $B(0) > 0$, $(B^*, 0)$ has a rapidly decaying (full line) inverse potential (G_n). It is easy to see that after the translation, (G_{n+m}) is still an inverse potential for $(B^*, 0)$. Thus we can approximate (aB^*, bB) in another way. Let (F_n) be the potential of (a,b) and (G_n) be the potential of $(B^*, 0)$. We define a sequence of l^2 potentials, $(H_n^{(k)})$, such that

$$H_n^{(k)} = F_n \text{ for } n < k \text{ , and } H_n^{(k)} = G_{n-2k} \text{ for } n \geq k.$$

Then as $k \to +\infty$,

$$\hat{d}\left(\overbrace{H_{<k}^{(k)}}, (a,b)\right) < c(||F_{\geq k}||_{l^2})^{1/4} \to 0 \text{ and}$$

$$\hat{d}\left(\overbrace{H_{\geq k}^{(k)}}, (B^*,0)\right) < c(||G_{<-k}||_{l^2})^{1/4} \to 0.$$

Therefore, it is easy to see that $\overbrace{H^{(k)}} = \overbrace{H_{<k}^{(k)}}\,\overbrace{H_{\geq k}^{(k)}} \in \mathbf{H}$ converges to $(a,b)(B^*,0) = (aB^*, bB) \notin \mathbf{H}$.

This example of a sequence of potentials $(H_n^{(k)})$ provides a typical picture. It is a sequence of l^2 potentials with a nontrivial tail which nearly produces a soliton data and is shifted farther and farther to the right. In the limit this tail produces the common inner factor. Note that $(H_n^{(k)})$ is not a Cauchy sequence in $l^2(Z_{\geq 0})$, but any finite truncation $(H_{<N}^{(k)})$ is a Cauchy sequence. With this picture in mind, we define the space $l^2(Z_{\geq 0}) \times G$ and give it the metric

$$d\left((F,g),(F',g')\right) = \inf_{N,M \in \mathbf{Z}} d\left((F,g),(F',g'),N,M\right)$$

where for any two integers N, M the distance $d\left((F,g),(F',g'),N,M\right)$ is defined as

$$||F_{<N} - F'_{<M}||_{l^2} + \hat{d}\left(\overbrace{F_{\geq N}}(\bar{g},0), \overbrace{F'_{\geq M}}(\bar{g}',0)\right) + \int_{\mathbf{T}} 1 - |a_{\geq N}| + \int_{\mathbf{T}} 1 - |a'_{\geq M}|\ .$$

The last two terms encourage to cut at N, M such that the tails are like soliton data. The second term combines the NLFT of these tails with the inner function part.

Obviously, d has symmetry, i.e. $d((F,g),(F',g')) = d((F',g'),(F,g))$. The triangle inequality holds as well, since for all $(F^i, g^i) \in \overline{\mathbf{H}}$ and positive integers N_i , $i = 1,2,3$, we have

$$\begin{aligned}
&d((F^1,g^1),(F^3,g^3),N_1,N_3)\\
&\leq d((F^1,g^1),(F^2,g^2),N_1,N_2) + d((F^2,g^2),(F^3,g^3),N_2,N_3)
\end{aligned}$$

since the l^2 norm and the metric $\hat{d}$ satisfy the triangle inequalities. By first taking $\inf_{N_1}$, $\inf_{N_3}$, then $\inf_{N_2}$ to the last inequality, we conclude that

$$d((F^1,g^1),(F^3,g^3)) \leq d((F^1,g^1),(F^2,g^2)) + d((F^2,g^2),(F^3,g^3)).$$

Noreover d satisfies definiteness, i.e. $d((F,g),(F',g')) = 0$ implies $F = F'$ and $g = g'$. To prove this, suppose first that the inf in the definition of $d((F,g),(F',g'))$ is obtained at some N_0, M_0. Then, $F_{<N_0} = F'_{<M_0}$ and $\overbrace{F_{\geq N_0}}(\bar{g},0) = \overbrace{F'_{\geq M_0}}(\bar{g}',0)$. The last equality means that g and g' as the common inner factor are the same and $F_{\geq N_0} = F'_{\geq M_0}$ since the NLFT on the half line is injective. Therefore, $F = F'$ and $g = g'$. Suppose the inf in the definition of $d((F,g),(F',g'))$ is not attained by any finite pair N, M. Let N_n, M_n be the sequence of pairs such that it approaches the inf. Then we can argue that both positive sequences N_n and M_n must go to infinity. Thus the first term in the definition of d says that $F = F'$. Moreover,

$$\begin{aligned}
||g-g'||_{H^1} &\leq \hat{d}((\bar{g},0),(\bar{g}',0))\\
&\leq \hat{d}\left((\bar{g},0),\overbrace{F_{\geq N_n}}(\bar{g},0)\right) + \hat{d}\left(\overbrace{F_{\geq N_n}}(\bar{g},0),\overbrace{F'_{\geq M_n}}(\bar{g}',0)\right)\\
&\quad + \hat{d}\left(\overbrace{F'_{\geq M_n}}(\bar{g}',0),(\bar{g}',0)\right)
\end{aligned}$$

and as $n \to \infty$ each term goes to zero. Hence $g = g'$. As a conclusion, d is a metric.

It would be interesting to understand whether $l^2(Z_{\geq 0}) \times G$ is complete under d.

Now we define an operator, also called the NLFT, from $l^2(Z_{\geq 0}) \times G$ to $\overline{\mathbf{H}}$ such that

$$NLFT: \ (F,g) \to \overbrace{F}(\bar{g},0).$$

And the inverse map from $\overline{\mathbf{H}}$ to $l^2(Z_{\geq 0}) \times G$ is defined as:

$$NLFT^{-1}: \ (a,b) \to (F,g)$$

where F is obtained by applying the layer stripping method on (a,b), and g is the common inner factor of a^* and b. We will prove that the NLFT is a homeomorphism between $l^2(Z_{\geq 0}) \times G$ and $\overline{\mathbf{H}}$.

Lemma 3.16. *The NLFT is continuous from $l^2(Z_{\geq 0}) \times G$ to $\overline{\mathbf{H}}$ with the metrics defined above.*

Proof. Given $(F,g) \in l^2(Z_{\geq 0}) \times G$ and $\epsilon > 0$. Since the NLFT is continuous on $l^2(Z_{\geq 0})$ and $\{(F_{<N})\}$ is a Cauchy sequence in $l^2(Z_{\geq 0})$, there is a $\delta_0 > 0$ such that if $H \in l^2(Z_{\geq 0})$ and $||H - F_{<N}||_{l^2} < \delta_0$ for some $N \in \mathbf{N}$, then $\hat{d}\left(\overbrace{H}, \overbrace{F_{<N}}\right) < \epsilon^2/100$.

Let $\delta < 1/2\min(\delta_0, \epsilon^2/100)$. Then for all $(F', g') \in \overline{\mathbf{H}}$ with $d((F,g),(F',g')) < \delta$, there is some N_0 and M_0 such that

$$||F_{<N_0} - F'_{<M_0}||_{l^2} + \hat{d}\left(\overbrace{F_{\geq N_0}}(\bar{g}, 0), \overbrace{F'_{\geq M_0}}(\bar{g}', 0)\right) < \min(\delta_0, \epsilon^2/100).$$

Thus $\hat{d}\left(\overbrace{F_{<N_0}}, \overbrace{F_{<M_0}}\right)$ and $\hat{d}\left(\overbrace{F_{\geq N_0}}(\bar{g}, 0), \overbrace{F'_{\geq M_0}}(\bar{g}', 0)\right) < \epsilon^2/100$. We will prove that

$$\hat{d}\left(\overbrace{F}(\bar{g}, 0), \overbrace{F'}(\bar{g}', 0)\right) = \hat{d}\left(\overbrace{F_{<N_0}}\overbrace{F_{\geq N_0}}(\bar{g}, 0), \overbrace{F_{<M_0}}\overbrace{F_{\geq M_0}}(\bar{g}', 0)\right) < \epsilon.$$

Let $\overbrace{F_{<N_0}} = (a, b)$, $\overbrace{F'_{<M_0}} = (a', b')$, $\overbrace{F_{\geq N_0}}(\bar{g}, 0) = (c, d)$, and $\overbrace{F'_{\geq M_0}}(\bar{g}', 0) = (c', d')$. Then on $\mathbf{T}$ $\overbrace{F}(\bar{g}, 0) = (ac - b\bar{d})$, $\overbrace{F'}(\bar{g}', 0) = (a'c' - b'\bar{d}', a'd' + b'\bar{c}')$. Observe that $||a - a'||_{L^1(\mathbf{T})}$, $||c - c'||_{L^1(\mathbf{T})}$, $||b - b'||_{L^2(\mathbf{T})}$, $||d - d'||_{L^2(\mathbf{T})}$, $|\log a^*(0) - \log(a')^*(0)|$, and $|\log c^*(0) - \log(c')^*(0)|$ all are smaller than $\epsilon^2/100$. Moreover, the sup norm of $|a|, |b|, |c|, |d| \ldots$ are less than 1. Thus with Hölder's inequality we can show that

$$||(ac - b\bar{d}) - (a'c' - b'\bar{d}')||_{L^1(\mathbf{T})} + ||(ad + b\bar{c}) - (a'd' + b'\bar{c}')||_{L^2(\mathbf{T})} < \epsilon/2.$$

Moreover,

$$(ac - b\bar{d})^*(0) = a^*(0)c^*(0)\ ,\ (a'c' - b'\bar{d}')^*(0) = (a')^*(0)(c')^*(0).$$

Hence

$$\begin{aligned}
&|\log(ac - b\bar{d})^*(0) - \log(a'c' - b'\bar{d}')^*(0)| \\
&\leq |\log a^*(0) - \log(a')^*(0)| + |\log c^*(0) - \log(c')^*(0)| \\
&\leq \epsilon^2/50 < \epsilon/2.
\end{aligned}$$

In short, we have proved that $\hat{d}\left(\overbrace{F}(\bar{g}, 0), \overbrace{F'}(\bar{g}', 0)\right) < \epsilon$ for all $(F', g') \in \overline{\mathbf{H}}$ such that $d((F,g),(F',g')) < \delta$. Hence the NLFT is continuous. $\square$

LEMMA 3.17. *The inverse NLFT from $\overline{\mathbf{H}}$ to $l^2(Z_{\geq 0}) \times G$ is continuous.*

PROOF. Given $(a, b) = (\tilde{a}\bar{g}, \tilde{b}g) \in \overline{\mathbf{H}}$ and $\epsilon > 0$ where $(\tilde{a}, \tilde{b}) \in \mathbf{H}$ and g is the common inner function. We will find $\delta > 0$ such that for all $(a', b') \in \overline{\mathbf{H}}$ with $\hat{d}((a,b),(a',b')) < \delta$ we have $d((F,g),(F',g')) < \epsilon$ where (F, g) and (F', g') are the images of (a, b) and (a', b') via the inverse NLFT.

Since $\overbrace{F} = (\tilde{a}, \tilde{b})$, first choose N_0 large such that $\int_{\mathbf{T}} 1 - |\tilde{a}_{\geq N_0}| < \epsilon/10$. By the proof of Lemma 3.15, there is $\delta_0 > 0$ such that for all $(a', b') \in \overline{\mathbf{H}}$ with $\hat{d}((a,b),(a',b')) < \delta_0$ we have $||F_{<N_0} - F'_{<N_0}||_{l^2} < \epsilon/10$ and because the NLFT is continuous on the half line we can further require that $\hat{d}\left(\overbrace{F_{<N_0}}, \overbrace{F_{<N_0}}\right) < \epsilon^2/100$. Then we claim that $\delta = \min(\delta_0, \epsilon^2/100)$ will do.

Suppose $\hat{d}((a,b),(a',b')) < \delta$. Then $||F_{<N_0} - F'_{<N_0}||_{l^2} < \epsilon/10$ and by assumption $\int 1 - |\tilde{a}_{\geq N_0}| < \epsilon/10$. Now, we will prove that $\hat{d}\left(\overbrace{F_{\geq N_0}}(\bar{g}, 0), \overbrace{F'_{\geq N_0}}(\bar{g}', 0)\right)$ automatically and $\int 1 - |\tilde{a}'_{\geq N_0}|$ are also small.

Since

$$(a,b) = \overbrace{F_{<N_0}}\overbrace{F_{\geq N_0}}(\bar{g},0) \text{ and } (a',b') = \overbrace{F'_{<N_0}}\overbrace{F'_{\geq N_0}}(\bar{g}',0)$$

we have

$$\overbrace{F_{\geq N_0}}(\bar{g},0) = (\overbrace{F_{<N_0}})^{-1}(a,b) \in \overline{\mathbf{H}}.$$

On $\mathbf{T}$, the right hand side is

$$(\bar{\tilde{a}}_{<N_0}a + \tilde{b}_{<N_0}\bar{b}, \bar{\tilde{a}}_{<N_0}b - \tilde{b}_{<N_0}\bar{a}).$$

We also have the similar expression for $\overbrace{F'_{\geq N_0}}(\bar{g}',0)$. Since $\hat{d}((a,b),(a',b')) < \delta \leq \epsilon^2/100$ and $\hat{d}\left(\overbrace{F_{<N_0}},\overbrace{F'_{<N_0}}\right) < \epsilon^2/100$, it is easy to see that

$$\int_{\mathbf{T}} |\tilde{a}_{\geq N_0}\bar{g} - \tilde{a}'_{\geq N_0}\bar{g}'| \leq \epsilon^2/10 < \epsilon/10 \text{ and}$$

$$\hat{d}\left(\overbrace{F_{\geq N_0}}(\bar{g},0), \overbrace{F'_{\geq N_0}}(\bar{g}',0)\right) < \epsilon/2.$$

Then we also have

$$\int_{\mathbf{T}} ||\tilde{a}_{\geq N_0}| - |\tilde{a}'_{\geq N_0}|| = \int_{\mathbf{T}} ||\tilde{a}_{\geq N_0}\bar{g}| - |\tilde{a}'_{\geq N_0}\bar{g}'|| \leq \int_{\mathbf{T}} |\tilde{a}_{\geq N_0}\bar{g} - \tilde{a}'_{\geq N_0}\bar{g}'| < \epsilon/10.$$

And together with the assumption $\int 1-|\tilde{a}_{\geq N_0}| < \epsilon/10$, we derive that $\int 1-|\tilde{a}'_{\geq N_0}| < \epsilon/5$. Hence

$$||F_{<N_0} - F'_{<N_0}||_{l^2} + \hat{d}\left(\overbrace{F_{\geq N_0}}(\bar{g},0), \overbrace{F'_{\geq N_0}}(\bar{g}',0)\right) + \int 1 - |\tilde{a}_{\geq N_0}| + \int 1 - |\tilde{a}'_{\geq N_0}| < \epsilon$$

Hence by definition $d((F,g),(F',g')) < \epsilon$.

$\square$

3.4. Rational Functions as Fourier Transform Data

In this section, we consider $(a, b) \in \mathbf{L}$ where a and b are rational functions and we investigate the Riemann Hilbert problem of finding $(a_+, b_+) \in \mathbf{H}$, and $(a_-, b_-) \in \mathbf{H}_0^*$ such that $(a_-, b_-)(a_+, b_+) = (a, b)$. We prove the existence of such Riemann Hilbert factorizations. If a^* has zeros inside D, the factorization is not unique, and we are able to construct all factorizations and parameterize them by a collection of subspaces $\gamma_i^{(j)}$ of $\mathbf{C}^2$.

We begin with some preliminary information about rational data (a, b).

THEOREM 3.18. *For every rational function b such that $|b| \leq 1$ but b is not identically* 1 *on $\mathbf{T}$, and for any finite subset $Z = \{z_1, z_2, ..., z_n\} \subset D \setminus \{0\}$ together with corresponding positive integers $\{m_1, m_2, ..., m_n\}$, there is a unique rational function a such that $(a, b) \in \mathbf{L}$ and Z is the zero set of a^* in D with corresponding multiplicities $m_1, m_2, ..., m_n$.*

Note that for rational functions a, b, $|a|^2 + |b|^2 = 1$ on $\mathbf{T}$ if and only if $aa^* + bb^*$ is constant 1 because $aa^* + bb^* = |a|^2 + |b|^2$ on $\mathbf{T}$.

We emphasize the comparison of this lemma with Lemma 3.4. In Lemma 3.4, to insure that a is a Laurent polynomial, we need the restriction that zeros of a^* form a subset of the zeros of $1 - bb^*$ while here zeros of a^* can be any finite set.

PROOF. We construct a rational function a with the desired properties. Define $P = 1 - bb^*$. Then P is a rational function with the symmetry $P = P^*$. Hence $z \in D$ is a pole or zero of P with order m if and only if $z^* = \bar{z}^{-1} \in D^*$ is a pole or zero of P with the same order. Since $|b| \leq 1$ on $\mathbf{T}$, we have $0 \leq P \leq 1$ on $\mathbf{T}$. Therefore P has no poles on $\mathbf{T}$ and each zero of P on the circle, being a local minimum, is of even order.

We describe the rational function a up to a scalar multiple by establishing the correct zeros and poles. First let a_0 be the rational function which has no poles and zeros in D^*, which has the same zeros as P on $\mathbf{T}$ but with half the multiplicities, and which has the same zeros and poles as P on D with the same multiplicities. Note that $a_0 a_0^*$ has the same zeros and poles as P, but a_0 does not have the desired zeros in D. To rectify this, we define

$$a = a \prod_{i=1}^{n} B_{z_i}^{m_i}$$

where B_{z_i} is the Blaschke factor which has a zero at z_i and a pole at z_i^*. Since each Blaschke factor B satisfies $BB^* = 1$ we have that aa^* still has the same zeros and poles as P and a has the desired set of zeros in D with multiplicities.

According to our construction of a, the rational function $f = P/aa^*$ has no poles and no zeros and therefore is a constant. Moreover, f is positive on $\mathbf{T}$. Hence we can normalize a by a positive factor so that $f = 1$. We can also normalize a by a phase factor to obtain $a^*(0) > 0$. Hence a is a rational function with $a^*(0) > 0$ and $aa^* + bb^* = 1$. Also, a^* is analytic in D. This means $(a, b) \in \mathbf{L}$ and a satisfies all the desired properties.

Now we prove that such rational a is unique. Observe that rational functions which are analytic in D are of the form $f_{outer}B$ when restricted to D where f_{outer} is an outer function and B is a Blaschke product. The Blaschke product part of a^* is determined when we are given its zeros in D. The outer function part is determined (up to a phase factor) by $|a^*|$ on $\mathbf{T}$ which is given as $\sqrt{1-|b|^2}$. And the phase factor is determined by $a^*(0) > 0$. Hence $a^*|_D$ is uniquely determined. Since a^* is rational, it is determined by its restriction to D. Therefore a^*, and thus a, is unique.

□

LEMMA 3.19. *Assume $(a,b) \in \mathbf{L}$ is rational. Then for any factorization*

$$(a_-, b_-)(a_+, b_+) = (a,b),\ (a_-, b_-) \in \mathbf{H}_0^*,\ (a_+, b_+) \in \mathbf{H},$$

we have that (a_-, b_-) and (a_+, b_+) are also rational.

PROOF. We first show that b_+ is a rational function.

Given that a and b are rational functions, and $|a|$, $|b| \leq 1$ on $\mathbf{T}$, we know that a and b have no poles on $\mathbf{T}$ and are bounded in a neighborhood around $\mathbf{T}$.

The identity $(a_-, b_-)(a,b) = (a_+, b_+)^{-1}$ implies $b_- = -ab_+ + ba_+$ and hence $ab_+ = ba_+ - b_-$. On the right hand side of the last equation, a_+ and b_- have analytic extensions on D^*, and b is a rational function. Hence ab_+ has a meromorphic extension to D^* with finitely many poles in D^*. Moreover, $|a_+|, |b_-| \leq 1$ on D^*, and b is bounded in a neighborhood of $\mathbf{T}$. Therefore, $\int_{\mathbf{T}} |ab_+|^2(r.)$ is bounded for $1 \leq r \leq 1+\epsilon$ for some $\epsilon \geq 0$. On the other hand, b_+ has an analytic extension on D with absolute value smaller or equal to 1 and a is a rational function which is bounded in a neighborhood of $\mathbf{T}$. Hence ab_+ is meromorphic in D with finitely many poles in D and $\int_{\mathbf{T}} |ab_+|^2(r.)$ is bounded for $1-\epsilon \leq r \leq 1$ for some $\epsilon \geq 0$.

Now we can remove the poles of ab_+ by the following recursive procedure. If z_∞ is a pole of ab_+, then subtract a constant from ab_+ so that the new function has a zero at z_0. Then we multiply the new function by $(z - z_\infty)/(z - z_0)$. This will reduce the order of the pole at z_∞ and leave the order of other poles unchanged and doesn't produce new poles. Iterating this procedure, we obtain a function g, which is holomorphic in D and D^* and $\int_{\mathbf{T}} |g|^2(r.)$ remains bounded for $1-\epsilon \leq r \leq 1+\epsilon$. That means $g \in H^2(D) \cap H^2(D^*)$ and thus g is a constant. This proves that ab_+ is a rational function and so is b_+.

Similarly, with the equation $a_-^* = a^*a_+ + b^*b_+$, we can prove that a_+ is rational. And by $(a_-, b_-) = (a,b)(a_+, b_+)^{-1}$, (a_-, b_-) is also rational. □

Given a rational data $(a,b) \in \mathbf{L}$, we now transform the factorization problem $(a,b) = (a_-, b_-)(a_+, b_+)$ with $(a_-, b_-) \in \mathbf{H}_0^*$ and $(a_+, b_+) \in \mathbf{H}$, into a more classical *Riemann-Hilbert problem*.

Given $aa^* + bb^* = 1$, it is easily checked that the following conditions

$$\left\{ \begin{array}{l} (a_-, b_-)(a_+, b_+) = (a,b) \\ a_+a_+^* + b_+b_+^* = 1,\ a_-a_-^* + b_-b_-^* = 1 \end{array} \right. \tag{3.11}$$

are equivalent to either of the following equivalent equations:

$$\begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix} \begin{pmatrix} a_+ & b_+ \\ -b_+^* & a_+^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -b^* & a^* \end{pmatrix} \tag{4.2a}$$

$$\text{or } \begin{pmatrix} a_+^* & -b_+ \\ b_+^* & a_+^* \end{pmatrix} \begin{pmatrix} a_+ & -b_- \\ -b_+^* & a_- \end{pmatrix} = \begin{pmatrix} 1 & -b \\ 0 & a \end{pmatrix} \tag{4.2b}$$

If we multiply equation (4.2b) by equation (4.2a), we obtain

$$\begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix}\begin{pmatrix} a_+ & b_- \\ -b_+^* & a_- \end{pmatrix} = \begin{pmatrix} 1 & -b \\ -b^* & 1 \end{pmatrix} \tag{3.12}$$

Hence, given two rational functions a and b such that $(a, b) \in \mathbf{L}$, (3.11) $\Rightarrow$ (3.12). Moreover, if we require $(a_+, b_+) \in \mathbf{H}$ and $(a_-, b_-) \in \mathbf{H}_0^*$, then on the left hand side of (3.12), all the entries of the first matrix (hereafter denoted by A_+) are analytic in D and a^* appears implicitly as its determinant. Similarly, all the entries of the second matrix(hereafter denoted by A_-) are analytic in D^* and a appears implicitly as its determinant. Thus (3.12) as a classical Riemann-Hilbert factorization of the matrix $\begin{pmatrix} 1 & -b \\ -b^* & 1 \end{pmatrix}$ (hereafter denoted by A) into a product

$$A = A_+ A_+^*$$

where

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^* \text{ is defined as } \begin{pmatrix} a^* & c^* \\ b^* & d^* \end{pmatrix}.$$

and all entries of A_+ have holomorphic extensions to D.

Conversely, assume we have such a Riemann-Hilbert factorization of A and assume it satisfies the following properties:

(1) The two elements in the first row of A_+ have no common inner factor. The two elements in the second row of A_+ have no common inner factor.
(2) $\det A_+$ is a rational function such that the zeros in D equal the zeros of a^* in D.
(3) $A_+(0)$ is of the form $\begin{pmatrix} + & * \\ 0 & + \end{pmatrix}$ where $+$ denotes a positive element.

Then, $(a_-, b_-)(a_+, b_+)$ is a decomposition of (a, b) into factors in $\mathbb{H}_0^*$ and $\mathbb{H}$ respectively, where

$$\begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix} = A_+.$$

To see this, note that the product of the first row of A_+ and the first column of A_+^* gives us $a_+a_+^* + b_+b_+^* = 1$. Hence according to the above Properties 1 and 3, $(a_+, b_+) \in \mathbf{H}$. Similarly, $(a_-, b_-) \in \mathbf{H}_0^*$. Moreover, the determinants of the left hand side and the right hand side of (3.12) give us $(\det A_+)(\det A_+)^* = 1 - bb^*$, which together with the above Property 2 implies that $\det A_+ = a_+^* a_-^* - b_+ b_-^* = a^*$ (by Lemma 3.19).

We call a Riemann Hilbert factorization of A with Properties 1,2,3 listed above an admissible Riemann Hilbert factorization.

To obtain the Riemann Hilbert factorization for matrices A whose determinant does not vanish on $\mathbf{T}$, we reproduce a result from [**21**]. There the authors define *a factorization in* L_p $(1 < p < \infty)$ of a measurable matrix function G on $\mathbf{T}$ to be a representation $G = G_+ \Lambda G_-$ where Λ is a diagonal matrix of the form $diag[z^{\kappa_1}, \cdots, z^{\kappa_n}]$ for some integers $\kappa_1 \geq \kappa_2 \geq \cdots \kappa_n$ and $G_+ \in H_p^+$, $G_- \in H_q^-$, $G_+^{-1} \in H_q^+$, $G_-^{-1} \in H_p^-$. In particular, $\det G_+$ and $\det G_-$ are non-vanishing on functions on $D \cup \mathbf{T}$ and $D^* \cup \mathbf{T}$ respectively.

The exponents κ_i are called *p-partial indices* of G. The sum $\kappa_1 + \cdots + \kappa_n$ is called the *p-total index* of X. It is shown that though the factorization is not unique, the partial indices are uniquely determined by G and p. Moreover, If $G = G_+^1 \Lambda G_-^1 = G_+^2 \Lambda G_-^2$ are two factorizations of G in L_p, then $G_+^2 = G_+^1 H$

and $G_-^2 = \Lambda^{-1}H^{-1}\Lambda G_-^1$ for some invertible matrix $H = (h_{ij})$ such that h_{ij} is a polynomial of degree $\leq \kappa_i - \kappa_j$ if $\kappa_i - \kappa_j \geq 0$ and $h_{ij} = 0$ if $\kappa_i - \kappa_j < 0$. In the case of a rational 2×2 matrix X as in our Riemann Hilbert problem, the p-partial indices n_1, n_2 are independent of $1 < p < \infty$, and one can actually show that the partial indices are all equal to 0.

We now state the precise result we need from [**21**]. Since we allow degeneracy on $\mathbf{T}$, which is slightly more general than stated in [**21**], and for the sake of self containment, we present a proof as well.

LEMMA 3.20. *Assume $|b| \leq 1$ on $\mathbf{T}$, but $|b|$ is not constant 1 on $\mathbf{T}$. Then there is a rational matrix function A_+ which is analytic and non-degenerate in D and $A_+A_+^* = A$. Such A_+ is unique up to multiplication by a unitary matrix from the right.*

PROOF. Note that on every point z of $\mathbf{T}$ the matrix $A(z)$ is positive semidefinite, and that $A = A^*$. Moreover, $A(z)$ is positive definite on all but finitely many points of $\mathbf{T}$.

Set $A_0 = A$. We will define a sequence of rational matrix functions A_i by the recursion $A_{i+1} = B_iA_iB_i^*$ with rational matrix functions B_i, such that A_i will have decreasing order of poles outside $\{0, \infty\}$ with increasing i. By virtue of the recursion we retain the property that A_i is positive definite on $\mathbf{T}$ and $A_i = A_i^*$, and we will also keep the matrix of $A_i(1)$ constant through most of the steps by imposing that $B_i(1)$ is the identity matrix.

Assume A_i is already defined and let $z_i \in D$ be a point such that some entry of A_i has a pole at z_i. By the symmetry relation $A = A^*$ this is equivalent to some entry of A_i having a pole at $z_i^* \in D^*$. Let f be the Möbius transformation of the Riemann sphere which sends z_i to ∞ and leaves the points 0 and 1 invariant. Let $B_i(z) = f(z)\mathrm{id}$, then clearly $A_{i+1} := B_i^{-1}A_i(B_i^*)^{-1}$ has a reduced order of pole at z_0 and z_0^* while the order of poles elsewhere outside $\{0, \infty\}$ is unchanged. We may repeat this process until all (finitely) many poles of any entry of A outside $\{0, \infty\}$ are removed.

Next we remove the zeros of $\det(A_i)$ in D and D^* by a similar process. First note that $\det(A_i)$ is not constant equal 0 by the assumption that $|b|$ is not constant 1 on $\mathbf{T}$, hence there are only finitely many zeros to be removed. Assume A_i is already defined and let $z_i \in D$ be a zero of $\det(A_i)$, which is equivalent to $z_i^* \in D^*$ being a zero of $\det(A_i)$. Then there is a unit vector v such that $v^TA_i(z_i) = 0$. Let w be a unit vector perpendicular to v. In the basis v, w the matrix z then necessarily takes the form

$$\begin{pmatrix} (z - z_0)r(z) & (z - z_0)s(z) \\ t(z) & u(z) \end{pmatrix}$$

with r, s, t, u analytic near z_0. Let f be again the Möbius transformation of the Riemann sphere which sends z_0 to ∞ and leaves the points 0 and 1 invariant. Then the determinant of the matrix

$$\begin{pmatrix} f(z) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} (z - z_i)r(z) & (z - z_i)s(z) \\ t(z) & u(z) \end{pmatrix}$$

has a zero of one order less at z_0 while the order of zeros at all other points outside $\{0, \infty\}$ is unchanged. If B_i denotes the left factor in this product as matrix in the original basis, we note that the determinant

$$A_{i+1} = B_iA_iB_i^*$$

has reduced order of zeros at z_0 and z_0^* and unchanged order of zeros at all other points outside $\{0, \infty\}$. We may iterate this process until all zeros of $\det(A)$ outside $\{0, \infty\} \cup \mathbf{T}$ are removed.

Next we remove the zeros of $\det(A_i)$ on $\mathbf{T}$ by a similar process. Let z_i be such a zero and let v as before be a unit vector such that $v^T A_i(z_i) = 0$. Then $v^T A_i(z_i)v = 0$ and since A_i is positive semidefinite on $\mathbf{T}$, this has to be a zero of even order and thus at least of order 2. In the basis given by v and a perpendicular unit vector w the matrix $A_i(z)$ in the vicinity of z_i takes the form

$$\begin{pmatrix} (z-z_0)^2 r(z) & (z-z_0)s(z) \\ (z-z_0)t(z) & u(z) \end{pmatrix} .$$

Note that both off diagonal terms vanish at z_i since $A_i(z_i) = A_i^*(z_i)$. Then with the notation as above,

$$\begin{pmatrix} f(z) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} (z-z_i)^2 r(z) & (z-z_i)s(z) \\ (z-z_i)t(z) & u(z) \end{pmatrix} \begin{pmatrix} f^*(z) & 0 \\ 0 & 1 \end{pmatrix}$$

is still analytic near z_i and has lower order of vanishing at z_i, while leaving the order of zeros outside $\{z_i, 0, \infty\}$ unchanged. Proceeding as before we remove in this way all zeros on $\mathbf{T}$.

Let A_{n-1} denote the final matrix in this process, hence A_{n-1} is analytic and non-singular outside $\{0, \infty\}$. Letting $A_n = B_{n-1}A_{n-1}B_{n-1}^*$ for some appropriate constant unitary matrix B_{n-1} we may assume that $A_n(1)$ is diagonal. While $\det(A_n)$ may have a pole at 0 or ∞, a winding number argument excludes this possibility since $\det(A_n)$ is positive on $\mathbf{T}$. Being a rational function, $\det(A_n)$ is then actually constant.

Each diagonal entry of A_n can be expressed by $e_i^T A_n e_i$ for unit vectors e_i. Again this rational function may in principle have a pole at 0 or ∞, but this possibility is excluded by positivity of this rational function on $\mathbf{T}$ using positive definiteness of A_n. Hence the diagonal entries of A_n are constant as well. The product of the diagonal entries is equal to the determinant, since this is the case at $z = 1$. Hence the off diagonal terms of A_n, which are conjugate to each other on $\mathbf{T}$, have to be constant equal to 0 on $\mathbf{T}$. By analyticity they have to be constant 0 everywhere. Finally setting $A_{n+1} = B_n A_n B_n^*$ for some appropriate constant diagonal matrix B_n with positive diagonal entries assures A_{n+1} is constant equal to the identity matrix. Note that these last two steps producing A_n and A_{n+1} are the only ones which change the matrix $A_i(1)$, but they do so by a controlled amount determined by the eigenvalues of $A(1)$ which are less than $\sup_{z \in \mathbf{T}} |b(z)|$ away from 1. Denoting by A_+ the product of the matrices B_i from $i = 1$ to $i = n$ we obtain the factorization $A = A_+ A_+^*$.

Now we show the uniqueness of A_+. If there is another rational matrix B which is analytic and non-degenerate in D and $BB^* = A = A_+ A_+^*$, then $A_+^{-1} B = A_+^*(B^*)^{-1}$ which will be denoted by C. Thus $C = A_+^{-1} B$ is analytic in D and similarly $C = A_+^*(B^*)^{-1}$ is analytic in D^*. Moreover, it is continuous on $\mathbf{T}$. This is clear if $\det(A)$ has no zeros on $\mathbf{T}$ since then A_+ and B are regular on $\mathbf{T}$. If $\det(A)$ has a zero at $z_0 \in \mathbf{T}$, we observe that there is unit vector v such that $v^T A = 0$, and hence also $v_T A^+ A_+^* v = ||v_T A_+||^2 = 0$ and hence $v_T A_+ = 0$ and similarly for B. Considering the matrix

$$B_f = \begin{pmatrix} f(z) & 0 \\ 0 & 1 \end{pmatrix}$$

as in the above process of removal of zeros we note that $B_f A_+$ and $B_f B$ are analytic near z_0 but with lower order vanishing of determinant. If the determinant still vanishes at z_0, we iterate this process until the matrix becomes regular, hence we find a sequence $B_1, \dots B_n$ so that $B_1 \dots B_n A_+$ and $B_1 \dots B_n B$ are analytic and regular near z_0. But then $C = A_+^{-1} B = A_+^{-1} B_n^{-1} \dots B_1^{-1} B_1 \dots B_n B$ is regular near z_0. Hence C is a constant matrix. In addition, $C^* = C^{-1}$ which means C is unitary. Therefore A_+ is unique up to a unitary matrix. □

Note that A_+ is non-degenerate in D. In particular $A_+(0)$ is invertible. It is a well-known fact that for an invertible 2 by 2 complex matrix G, there is a unique unitary matrix T and a unique matrix G' of the form $\begin{pmatrix} + & * \\ 0 & + \end{pmatrix}$ such that $GT = G'$. Therefore if we require A_+ to be of this form as stated in Property 3, then it is uniquely determined.

We call the factor A_+ described in the previous lemma and with the normalization of $A_+(0)$ being upper triangular with positive diagonal entries the *regular factor* of A, because it is regular on D, and we denote it by $\tilde{A_+}$. Note that $\tilde{A_+}$ gives an admissible Riemann-Hilbert factorization without singularities in D, in particular the entries in the first row of $\tilde{A_+}$ do not have a common inner factor since they do not simultaneously vanish in D, and similarly for the entires of the second row.

We now consider the Riemann-Hilbert factorization with zeros, i.e. we find A_+ which degenerates at given prescribed points $Z = \{z_1, z_2, \dots, z_n\} \in D \setminus \{0\}$.

We shall develop the matter through special cases of increasing complexity, we will follow the discussion in [**14**] using Blaschke Potapov factors as described in the following lemma.

LEMMA 3.21. *Given $z_1 \in D \setminus \{0\}$ and a one dimensional subspace γ_1 of $\mathbf{C}^2$, there is a unique meromorphic matrix function B on the Riemann sphere such that*

(1) *The entries of B have no pole outside z_1^*.*
(2) *For $z \in \mathbf{T}$ the matrix $B(z)$ is unitary.*
(3) *The determinant* $\det(B)$ *a vanishes only at z_1 and it vanishes of first order there.*
(4) *$B(0)$ is upper triangular and has positive diagonal entries*
(5) *For every vector $v \in \mathbf{C}^2$ we have $B(z_1)v \in \gamma_1$.*

PROOF. We first prove existence by explicitly describing such a matrix B. Let P be the orthogonal projection onto the space γ_1 and define the auxiliary matrix (Blaschke Potapov factor) $B'(z) = (P + (I - P)f(z))$, where $f(z)$ is a Blaschke factor vanishing at z_1. Since f has modulus 1 on $\mathbf{T}$, this matrix $B'(z)$ is unitary for $z \in T$. The determinant of B' is equal to $f(z)$ and thus satisfies the desired properties. The entries of B' can only have a pole where f has, and f only has a pole at z_1^*. The range of $B'(z_1)$ is the range of P, which is γ_1. Since $B'(0)$ is non-singular, there exists a unitary matrix T such that $B'(0)T$ is upper triangular and positive on the diagonal. Then $B(z) = B'(z)T$ has all the desired properties.

To prove uniqueness of the matrix B, assume we have some other matrix function B' satisfying the properties listed in the lemma. Note that $(B')^* B'$ is the identity matrix on $\mathbf{T}$. Since $(B')^* B'$ is meromorphic on the sphere, it is constant equal to the identity matrix. Hence $(B')^{-1} = (B')^*$ wherever defined, and similarly $B^{-1} = B^*$

Let v be a unit vector spannning γ_1 and let w be a unit vector perpendicular to v. Changing into the basis (v, w) the functions B, B' take the form

$$\begin{pmatrix} r(z) & s(z) \\ (z-z_1)t(z) & (z-z_1)u(z) \end{pmatrix}, \begin{pmatrix} r'(z) & s'(z) \\ (z-z_1)t'(z) & (z-z_1)u'(z) \end{pmatrix}$$

with functions $r, s, t, u, r', s', t', u'$ analytic near z_1, since the range of the matrix at z_1 is spanned by the vector $(1,0)^T$ in this basis. Since $\det(B)$ vanishes of first order, it follows that all entries of

$$B^{-1}B' = \frac{1}{\det(B)} \begin{pmatrix} (z-z_1)u(z) & -s(z) \\ -(z-z_1)t(z) & r(z) \end{pmatrix}, \begin{pmatrix} r'(z) & s'(z) \\ (z-z_1)t'(z) & (z-z_1)u'(z) \end{pmatrix}$$

are analytic near z_1. We claim that all its entries are analytic near $(z_1)^*$ as well. For this it suffices to check that the entries of $(B^{-1}B')^*$ are analytic near z_1. However,

$$(B^{-1}B')^* = (B')^*(B^{-1})^* = (B')^{-1}B$$

is analytic near z_1 by the symmetric argument as above. It follows that $B^{-1}B'$ is analytic everywhere and thus constant. This constant matrix is both unitary (by evaluation on $\mathbf{T}$) and upper triangular with positive diagonal entries (by evaluation at 0). Hence it is the identity matrix, which proves $B = B'$. □

We denote the matrix B in this lemma by B_{z_1,γ_1}. In Lemma 3.21, we chose to parameterize for fixed z_1 the matrix B_{z_1,γ_1} by its range γ_1 at z_1. Alternatively, we could have used the kernel at z_1 to parameterize these matrices. This is the content of the following lemma.

LEMMA 3.22. *For fixed $z_1 \in D\backslash\{0\}$, the map $\gamma_1 \to \ker(B_{z_1,\gamma_1})$ is a self-bijection on the set of one dimensional subspaces of C^2.*

PROOF. We construct the inverse map. For δ_1 a one dimensional subspace, Let Q be the orthogonal projection onto that subspace. Define $B'(z) = (1-Q)+Qf(z)$ with f as above a Blaschke factor vanishing at z_1. Let T be the unique unitary map such that $TB'(0)$ is upper triangular with positive entries on the diagonal. Then the function $B(z) = TB'(z)$ satisfies the assumptions of the Lemma above with γ_i the range of the rank one matrix $B(z_1)$. To establish the claimed bijection it remains to show that for each one dimensional space δ_i there is a unique matrix function B satisfying the properties of the previous lemma for some γ_i. For this we consider another such matrix B' with the same kernel. Analogously to the previous lemma, by passing to a basis spanned by a unit vector in δ_1 and a perpendicular unit vector, we prove that $B'B^{-1}$ is regular at z_1 and then also regular at z_1^*. hence it is constant and hence it is constant equal to the identity matrix. □

Returning to the discussion of factorizations $A = A_+(A_+)^*$, we first consider the case that $\det(A_+)$ has only one zero z_1 in $D \setminus \{0\}$ and this is a simple zero. We choose a one dimensional subspace $\gamma_1 \subset \mathbf{C}^2$ such that $\tilde{A_+}(z_1)\gamma_1$ is neither spanned by $(1,0)^T$ nor by $(0,1)^T$, note that this excludes exactly two subspaces since $\tilde{A_+}(z_1)$ is regular. Then we write $A_+ := \tilde{A_+}B_{z_1,\gamma_1}$. We have:

(1) A_+ is analytic in D.
(2) $A_+A_+^* = \tilde{A_+}\tilde{A_+}^* = A$.
(3) A_+ degenerates at z_1, $\det(A_+)$ has a simple zero there, and this is the only point of degeneracy in D.
(4) $A_+(0)$ is upper triangular with positive enties on the diagonal.

(5) The entries of the first row of A_+ do not have a common inner factor. The entries of the second row of A_+ do not have a common inner factor.

The last property follows since the entries of the first row of A_+ may only vanish at z_1, but they do not both vanish at z_1 by choice of γ_1. Similarly for the second row. We have thus produced an admissible Riemann Hilbert factorization of A with one prescribed simple zero of $det(A_+)$ in D.

Now consider $Z = \{z_1\}$ but z_1 is a multiple zero with $\mathrm{ord}(\det A_+, z_1) = n$. There are in principle two cases to be discussed:

Case 1.: $A_+(z_1) \neq 0$.
Case 2.: $A_+(z_1) = 0$.

However, in Case 2 the two entries of the first row of A_+ have a common zero at z_1, which is inadmissible. We are thus reduced to discussing Case 1. Then necessarily $A_+(z)$ has rank one. Similarly to above we write

$$A_+ = \tilde{A_+} B_{z_1,\gamma_1} B_{z_1,\gamma_2} \dots B_{z_1,\gamma_n}$$

for some sequence of $\gamma_1, \dots \gamma_n$ of one dimensional subspaces of $\mathbf{C}^2$. Non-vanishing of $A_+(z_1)$ requires that $B_{z_1,\gamma_i}\gamma_{i+1} \neq \{0\}$ for each $1 \leq i < n$. Note that this condition for all $1 \leq i < n$ is sufficient for non-vanishing of $A_+(z_1)$. As before, we also require $\widetilde{A_+(z_1)}\gamma_1$ to not be spanned by $(1,0)$ or $(0,1)$ so as to obtain an admissible Riemann Hilbert factorization.

Based on these special cases we are ready to describe the family of Riemann Hilbert factorizations for any prescribed set of zeros of $\det(A_+)$ in D.

Theorem 3.23. *Assume we are given a rational function b with $|b| \leq 1$ on $\mathbf{T}$ and b is not identically 1 on $\mathbf{T}$. Define*

$$A = \begin{pmatrix} 1 & -b \\ -b^* & 1 \end{pmatrix}.$$

Assume we are given a finite subset $Z = \{z_1, z_2, \dots, z_k\}$ of $D \setminus \{0\}$ and positive integers $n_1, n_2, \dots, n_k$. Then there is a bijection between

(1) *the set of data consisting of one dimensional subspaces $\gamma_j^{(i)}$ for $1 \leq i \leq k$ and $1 \leq j \leq n_k$ with $\tilde{A_+}\gamma_1^{(i)}$ not spanned by $(1,0)^T$ or $(0,1)^T$ and $B_{z_i,\gamma_j^{(i)}}\gamma_{j+1}^{(i)} \neq \{0\}$ for all $1 \leq i \leq k$ and $1 \leq j < n_k$ and*
(2) *the set of admissible Riemann Hilbert factorizations*

$$A = A_+(A_+)^*,$$

i.e., A_+ a rational matrix function analytic in D such that the zeros of $\det(A_+)$ inside D are exactly the points $z_1, \dots z_k$ with multiplicities $n_1, \dots n_k$, $A_+(0)$ is of the form $\begin{pmatrix} + & * \\ 0 & + \end{pmatrix}$, and the two entries of the first row of A_+ have no common inner factor and the two entries of the second row have no common inner factor.

Proof. We first describe a different parametrization of the set of admissible Riemann Hilbert factorizations that depends on the chosen enumeration of zeros $\{z_1, \dots, z_k\}$. We write

$$A_+ = \tilde{A_+} B_{z_1,\tilde{\gamma}^{(1)}1} \dots B_{z_1,\tilde{\gamma}^{(1)}_{n_1}} B_{z_2,\tilde{\gamma}^{(2)}1} \dots B_{z_1,\tilde{\gamma}^{(2)}_{n_2}} \dots B_{z_k,\tilde{\gamma}^{(k)}1} \dots B_{z_1,\tilde{\gamma}^{(k)}_{n_k}} \tag{3.13}$$

for some choices of $\tilde\gamma_j^{(i)}$ which satisfy $B_{z_i,\tilde\gamma_j^{(i)}}\tilde\gamma_{j+1}^{(i)} \neq \{0\}$ for all $1 \le i \le k$ and $1 \le j < n_k$, and $\tilde\gamma_1^{(i)}$ avoids the two subspaces which make $A_+(z_i)$ be spanned by $(1,0)^T$ or $(0,1)^T$, a condition which depends on all matrices to the left of $B_{z_i,\tilde\gamma_1^{(i)}}$ in the above product. Note that all these matrices are non-singular at z_i, hence this last condition excludes exactly two subspaces. Then clearly A_+ gives an admissible Riemann Hilbert factorization.

To see that this is a parametrization of all admissible Riemann Hilbert factorizations we will recover the data $\tilde\gamma_j^{(i)}$ from A_+. To do so we argue recursively by the total order $\sum_{i=1}^k n_k$. If this total order is zero, we are done by Lemma 3.20. Assume we are given a Riemann Hilbert factorization $A = A_+(A_+)^*$ with the desired zeros and multiplicities. Then choose $B = B_{z_k,\tilde\gamma_{n_k}^{(k)}}$ by Lemma 3.22 so as to have the same one dimensional kernel as A_+ and consider A_+B^{-1}. Note that A_+B^{-1} is regular at z_k by an analoguous discussion as in the Lemma 3.22. Then A_+B^{-1} is the left factor of a Riemann Hilbert factorization of lower total order than A_+, and by induction we find a unique factorization as in (3.13) of A_+B^{-1}. Multiplying from the right by B gives the desired factorization of A_+.

The particular parameterization by $\tilde\gamma_j^{(i)}$ in the above proof is unsatisfactory, since it depends on the enumeration of the zeros in Z. To find the parameterization that is independent of enumeration as claimed in the theorem, we choose for each i the numbers $\tilde\gamma_j^{(i)}$ which arise from the above argument by an enumeration that has z_i as first element. These numbers are independent of the exact enumeration chosen and depend only on local properties near z_i, since the Blaschke Potapov factors with superscript (1) are constructed first. To see that these new parameters are equivalent to the old ones for some fixed enumeration, we shall successively replace the old parameters by the new ones, running through the Blaschke Potapov factors from right to left.

Note that $\gamma_j^{(i)}$ describes the range of the Blaschke Botapov factor $B_j^{(i)}$ in the product

$$A_+ = \tilde{A_+}\dots B_j^{(i)}B_{j+1}^{(i)}\dots B_{n_i}^{(i)}B_{\text{right}}T$$

where B is the product of Blaschke factors with upper index (i') with $i' \neq i$, while $\tilde\gamma_j^{(i)}$ describes the range of the factor $\tilde B_j^{(i)}$ in the product

$$A_+ = \tilde{A_+}\tilde B_{\text{left}}\tilde B_1^{(i)}\dots\tilde B_j^{(i)}\tilde B_{j+1}^{(i)}\dots\tilde B_{n_i}^{(i)}\tilde B_{\text{right}}\tilde T$$

where $\tilde B_{\text{left}}$ and $\tilde B_{\text{right}}$ are the products of the Blaschke factors to the left and to the right of the ones indexed by (i) in the fixed order. To show that these ranges are in bijective correspondence with each other, it suffices to show by Lemma 3.22 that the kernels of these matrices are in bijective correspondence. For this it suffices to show that the matrix

$$\tilde B_{j+1}^{(i)}\dots\tilde B_{n_i}^{(i)}\tilde B_{\text{right}}\tilde T T^* B^{-1}(B_{n_i}^{(i)})^{-1}\dots(B_{j+1}^{(i)})^{-1}$$

has a regular continuation at the point z_i. This will follow by induction on the number $n_i - j$. If that number is zero, then the product only has regular factors at z_i and the assertion is clear. To prove the induction step, we assume the product as written is regular, and prove that the corresponding product with one extra factor on either side is regular as well. By construction $B_j^{(i)}$ is the right most factor of the matrix

$$\tilde{A_+} B_1^{(i)} \dots B_j^{(i)}$$
$$= A_+ T^{-1} B^{-1} (B_{n_i}^{(i)})^{-1} \dots (B_{j+1}^{(i)})^{-1}$$

and hence has the same kernel as the regular extension of this matrix at z_i. On the other hand, this same matrix can be written as

$$\tilde{A_+} \tilde{B}_{\text{left}} \tilde{B}_1^{(i)} \dots \tilde{B}_j^{(i)} \tilde{B}_{j+1}^{(i)} \dots \tilde{B}_{n_i}^{(i)} \tilde{B}_{\text{right}} \tilde{T} T^{-1} B^{-1} (B_{n_i}^{(i)})^{-1} \dots (B_{j+1}^{(i)})^{-1}$$

and the kernel of this matrix has to be the same as the kernel of the product of the right most factors

$$\tilde{B}_j^{(i)} \tilde{B}_{j+1}^{(i)} \dots \tilde{B}_{n_i}^{(i)} \tilde{B}_{\text{right}} \tilde{T} T^{-1} B^{-1} (B_{n_i}^{(i)})^{-1} \dots (B_{j+1}^{(i)})^{-1}$$

which by accounting orders of vanishing of the determinant of each factor is singular at z_i. Hence we may multiply $(B_j^{(i)})^{-1}$ to the right of this matrix and retain a regular extension to z_i. This is what had to be shown, and we have established that we can successively replace the $\gamma_j^{(i)}$ by $\tilde{\gamma}_j^i$ retaining a good parameterization.

□

3.5. Soliton Data

In this section we focus on data of the type $(a, 0) \in \mathbf{L}$. This type of data in the analoguous continuous model is related to soliton solutions for the focusing nonlinear Schrödinger equation [**14**]. Following this connection we will also call such data soliton data. Observe that $(a, 0) \in \mathbf{L}$ implies that a^* is an inner function.

In the previous section we have described all preimages of soliton data under the nonlinear Fourier transform in case a is a rational function, which for soliton data means that a is a finite Blaschke product. As part of our discussion of soliton data we will first extend these results to the case of a being an infinite Blaschke product. With Frostman's theorem in mind this is a step towards general inner functions, however we are not able to give a satisfactory description of the inner function case.

Then we will derive the exact formula for all inverse potentials of rational soliton data and see that such potentials are rapidly decaying.

Finally, we will show that every rational data $(a_+, b_+) \in \mathbf{H}$, after applying finitely many steps of layer stripping if necessary, can be paired with another rational data $(a_-, b_-) \in \mathbf{H}_0^*$ such that $(a_-, b_-)(a_+, b_+)$ is of the form $(a, 0)$. Similarly all rational data $(a_-, b_-) \in \mathbf{H}_0^*$, after applying finitely many steps of layer stripping if necessary, can be paired with another rational data $(a_+, b_+) \in \mathbf{H}$ such that $(a_-, b_-)(a_+, b_+) = (a, 0)$. Thus we can derive a formula for all inverse potentials of rational data $(a, b) \in \mathbf{L}$ and in particular establish that any inverse potential of rational data is rapidly decaying.

THEOREM 3.24. *Given a^* an infinite Blaschke product with zeros $\{z_i\} \subset D \setminus \{0\}$ and corresponding multiplicities $\{n_i\}$. There is a bijection between the set of sequences of subspaces $\{\gamma_j^(i)\}$ of $\mathbf{C}^2$ for $1 \leq j \leq n_i$ and $1 \leq i < \infty$ satisfying the conditions that $\gamma_1^{(i)}$ not spanned by $(1,0)^T$ or $(0,1)^T$ and $B_{z_i, \gamma_j^{(i)}} \gamma_{j+1}^{(i)} \neq \{0\}$ for all $1 \leq i \leq \infty$ and $1 \leq j < n_k$ and the set of factorizations $(a_-, b_-)(a_+, b_+) = (a, 0)$ in such a way that the spaces $\gamma_j^{(i)}$ describe the singularity of $A_+ = \begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix}$ at the point z_i as in Theorem 3.23.*

PROOF. First we prove that given $\{\gamma_j^{(i)}\}$ as in the theorem, there exists a corresponding factorization. We approximate a^* by the finite Blaschke product

$$a_n^* = \prod_{i=1}^{n} \left(-\frac{\bar{z}_i}{|z_i|} \frac{z - z_i}{1 - \bar{z}_i z} \right)^{n_i}.$$

Since a_n^* is rational, we may apply Theorem 3.23 to obtain an admissible Riemann Hilbert factorization with parameters $\{\gamma_j^(i)\}$, $1 \leq j \leq n_i$, $1 \leq i \leq n\}$:

$$A_+^n = \begin{pmatrix} (a_+^n)^* & -b_+^n \\ -(b_-^n)^* & (a_-^n)^* \end{pmatrix}$$

Note that in the notation of that theorem, $\tilde{A_+}$ is the identity matrix.

Since $(a_+^n)^*, b_+^n, (a_-^n)^*, (b_-^n)^*$ are analytic and bounded by 1 on D, the theorem of Arzela-Ascoli provides a subsequence such that $(a_+^{n_k})^*, b_+^{n_k}, (a_-^{n_k})^*, (b_-^{n_k})^*$ converge uniformly on compact subsets of D. Thus $(a_+^{n_k})^*, b_+^{n_k}, (a_-^{n_k})^*, (b_-^{n_k})^*$ converge to some $H^\infty(D)$ functions a_+^*, b_+, a_-^*, b_-^*. We claim that $(a_-, b_-) \in \mathbf{H}_0^*$, $(a_+, b_+) \in \mathbf{H}$ and $(a_-, b_-)(a_+, b_+) = (a, 0)$.

Let $A_+ = \begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix}$. Since $\det A_+^{n_k} = a_{n_k}^*$ and it converges to $\det A_+$ uniformly on compact sets inside D, we have $\det A_+ = a_+^* a_-^* - b_+ b_-^* = a^*$, which is half of the equation $(a_-, b_-)(a_+, b_+) = (a, 0)$.

Next observe that $||A_+^{n_k}(z)||_{op} \leq 1$ for all $z \in D$ since $||B_i^j(z)||_{op} = 1$ for all $z \in D$. Thus, $||A_+(z)||_{op} \leq 1$ on D. Since the entires of A_+ are bounded analytic functions in D, they have almost everywhere on $\mathbf{T}$ nontangential limits, which we also denote by A_+. The estimate $||A_+(z)||_{op} \leq 1$ then remains true almost everywhere on $\mathbf{T}$. However $|\det A_+| = |a^*| = 1$ a.e. on the circle, and thus $\det(A_+ A_+^*) = 1$ almost everywhere on $\mathbf{T}$. Since $A_+ A_+^*$ is also positive semidefinite on $\mathbf{T}$, and bounded in operator norm by 1, we see that $A_+ A_+^*$ is the identity matrix almost everywhere on $\mathbf{T}$.

Next since all $A_+^{n_k}(0)$ are of the form $\begin{pmatrix} + & * \\ 0 & + \end{pmatrix}$, so if their limit $A_+(0)$.

To establish that that $\{\gamma_1^(i)\}$ is the range of A_+ at $\{z_i\}$, pick a nonzero vector v perpendicular to the range and note that $v^T A_+^{n_k}(z_i) = 0$ for all k. Hence in the limit $v^T A_+(z_i) = 0$. We may now multiply A_+ by $(B_{z_1, \gamma_1^{(i)}, z_i})^{-1}$ from the left as in the discussions of Lemma **??**, and proceed inductively with identifying ranges as $\gamma_j^{(i)}$. After multiplication by all n_i factors of the form $(B_{z_i, \gamma_j^{(i)}})^{-1}$ we obtain a matrix function regular at z_i as we can verify by accounting the order of vanishing of the determinant of each factor. This establishes in particular that $A_+(z_i)$ is rank one. This establishes also that the entries of the first row of $A_+(z_i)$ and the entries of the second row of $A_+(z_i)$ do not simultaneously vanish at z_i, by the choice of $\gamma_1^{(i)}$.

If a_+^* and b_+ have common singular inner factor, say g, then $\det A_+$ also has the factor g. However, $\det A_+$ is the infinite Blaschke product $\prod_{i=1}^\infty B_i^{n_i}(z)$ and has no singular inner part. Thus, a_+^* and b_+ (a_-^* and b_-^*) can not have common singular inner factor. Hence, we have showed that $(a_+, b_+) \in \mathbf{H}$, $(a_-, b_-) \in \mathbf{H}_0^*$, and $(a_-, b_-)(a_+, b_+) = (a, 0)$ is the desired factorization corresponding to the spaces $\{\gamma_j^{(i)}\}$.

To prove the uniqueness of such decomposition, suppose that there is another factorization $A'_+(A'_+)^* = I$, $\det A'_+ = a^*$, and $\{\gamma_j^{(i)}\}$ represents the images of A'_+ at $\{z_i\}$. Then $A_+^{-1} A'_+$ is analytic on D expect for possible poles $\{z_i\}$. However, at $\{z_i\}$ the parameters $\gamma_j^{(i)}$ match and by an argument as before this means that the product $A_+^{-1} A'_+$ has an analytic extension to z_i. Since the (i,j) element of $A_+^{-1} A'_+$ is a $H^\infty(D)$ function, say f_{ij}, divided by the infinite Blaschke product $\det A_+ = a^*$ and has no poles in D. Hence f_{ij} must have the factor a^*. Therefore each element of $A_+^{-1} A'_+$ belongs to $H^\infty(D)$. Similarly, $(A'_+)^{-1} A_+$ belongs to $H^\infty(D)$ and thus $((A'_+)^{-1} A_+)^* = A_+^* (A'^*_+)^{-1} \in \mathbf{H}_0^*$. However, on $\mathbf{T}$, $A_+ A_+^* = A'_+ A'^*_+ = I$ i.e. $A_+^{-1} A'_+ = A_+^* (A'^*_+)^{-1}$. Thus $A_+^{-1} A'_+$ can be extended as a bounded analytic function on the whole Riemann sphere. Therefore, $A_+^{-1} A'_+$ is a constant matrix. Moreover, it is unitary on $\mathbf{\Pi}$ and of the form $\begin{pmatrix} + & * \\ 0 & + \end{pmatrix}$ at 0. Hence $A_+^{-1} A'_+ = I$.

So far we have constructed the map from $\{\gamma_j^{(i)}\}$ to the factorization $(a_-, b_-)(a_+, b_+) = (a, 0)$ and showed that it is injective. It is easy to see that this map is onto. Given any factorization $(a_-, b_-)(a_+, b_+) = (a, 0)$, let $A_+ =$

$\begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix}$ and collect the constants $\{\gamma_j^{(i)}\}$ that describe the local behaviour of the matrix A_+ at the degenerating points z_i. The $\{\gamma_j^{(i)}\}$ are well defined and satisfy the properties of the theorem since $A_+(z_i)$ has to have rank one and range different from the span of $(0,1)^T$ and the span of $(1,0)^T$, and there are only finitely many parameters $\gamma_j^{(i)}$ for fixed i since $\det(A_+)$ can only vanish up to finite order.

$\square$

For $(a,0) \in \mathbf{L}$ where a^* is a general inner function, it is tempting to use Frostman's theorem to approach a^* by $\{a_n^*\}$ where $\{a_n^*\}$ is a sequence of Blaschke products and converges to a^* uniformly on D. Each $(a_n,0)$ is factorable and the decompositions are described by constants $\{\gamma_j^{(i)}\}$. However, we do not have a good description of the variety of the factorizations in the limit, as we do not understand what data replaces the $\gamma_j^{(i)}$ in the limit.

We pass to the next item of our discussion of soliton data and construct a formula for the inverse potential of $(a,0)$ where a^* is a finite Blaschke product.

First assume that a^* has only simple zeros $\{z_i\}$, $i = 1,\ldots,n$ i.e.

$$a^* = \prod_{i=1}^{n} -\frac{\bar{z}_i}{|z_i|}\frac{z-z_i}{1-\bar{z}_i z}\ . \tag{3.14}$$

Then given any $\gamma_i \in C^*$ for $i = 1,\ldots,n$ there is a unique rational matrix A_+ analytic in D such that $A_+A_+^* = Id$, $\det A_+ = a^*$, with $\mathrm{Im}A_+(z_i) = \langle \begin{pmatrix} 1 \\ -\gamma_i \end{pmatrix} \rangle$, and $A_+(0)$ is normalized. Denote $A_+ = \begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix}$. Now we compute its inverse potential.

Lemma 3.25. *Given $(a,0) \in \mathbf{L}$ where a^* is of the form (3.14) with $z_i \neq z_j$ for all $i \neq j$ and constants $\gamma_i \in C^*$ for $i = 1,\ldots,n$. There is a unique $A_+ = \begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix}$ described as above. Applying the layer stripping method for (a_+,b_+) and (a_-,b_-), we obtain a geometrically decreasing sequence (F_k),*

$$F_k = -\prod_{i=1}^{i=n}|z_i|\frac{\det M_1(k)}{\det M(k)}\ ,\ \text{for } k \in Z$$

where $M(k)$ is a n by n matrix with

$$M(k)_{ij} = \frac{1+\bar{\gamma}_i\gamma_j\bar{z}_i{}^k z_j^k}{\bar{z}_i{}^{-1}-z_j}\ ,\ \text{and } M_1(k) = \begin{pmatrix} & & & \bar{\gamma}_1\bar{z}_1{}^k \\ & M(k) & & \vdots \\ & & & \bar{\gamma}_n\bar{z}_n{}^k \\ z_1^{-1} & \cdots & z_n^{-1} & 0 \end{pmatrix}.$$

If $\gamma_i \neq 0,\infty$ for all i, then (F_k) is the unique l_2 sequence such that $\overbrace{(F_{\geq 0})} = (a_+,b_+)$, $\overbrace{(F_{<0})} = (a_-,b_-)$ and $\overbrace{(F)} = (a,0)$. Suppose $\gamma_i = 0$ or ∞ for some i. Reorder the zeros of a^* so that $\gamma_j \neq 0,\infty$ for $1 \leq j \leq k$, $\gamma_j = 0$ for $k+1 \leq j \leq l$, and $\gamma_j = \infty$ for $l+1 \leq j \leq n$. Then $\overbrace{(F_{\geq 0})}((B^1)^*,0) = (a_+,b_+)$, $((B^2)^*,0)\overbrace{(F_{<0})} = (a_-,b_-)$ where B^1 is the Blaschke product with simple zeros $\{z_{l+1}, z_{l+2},\ldots,z_n\}$ such that

$B^1(0) > 0$ and B^2 is the Blaschke product with simple zeros $\{z_{k+1}, z_{k+2}, \ldots, z_l\}$ such that $B^2(0) > 0$.

PROOF. Denote $A_+(0) = \begin{pmatrix} a_+^*(0) & -b_+(0) \\ -b_-^*(0) & a_-^*(0) \end{pmatrix} = \begin{pmatrix} a_1 & b \\ 0 & a_2 \end{pmatrix}$, $a_1,\ a_2 > 0$. Then the layer stripping method give us $F_0 = b_+(0)/a_+^*(0) = -b/a_1$.

From our discussion in the previous section, A_+ is just the normalization of the product of Blaschke-Potapov factors,

$$B_i = I + \left(-\frac{\bar{z}_i}{|z_i|}\frac{z - z_i}{1 - \bar{z}_i z} - 1\right) P_i,$$

and thus $A_+^{-1}(1/\bar{z}_i) = A_+^*(z_i)$. Hence A_+^{-1} is meromorphic on C^* with simple poles z_i. Write

$$A_+^{-1}(z) = A_0 + \sum_{i=1}^{i=n} \frac{A_i}{z - z_i}$$

for some constant matrices $A_0, A_1, \cdots, A_n$. Then, $A_0 = A_+^{-1}(\infty) = A_+^*(0) = \begin{pmatrix} a_1 & 0 \\ \bar{b} & a_2 \end{pmatrix}$.

Claim that $A_i = \begin{pmatrix} p_i \\ q_i \end{pmatrix} \begin{pmatrix} \gamma_i & 1 \end{pmatrix}$ for some $p_i,\ q_i \in C$.

Since $A_+^{-1}(z) \to A_i/(z - z_i)$, and $A_+(z) \to C_i = A_+(z_i)$ as $z \to z_i$, we have $A_i C_i = C_i A_i = 0$. Moreover $\mathrm{Im} C_i = \langle \begin{pmatrix} 1 \\ -\gamma_i \end{pmatrix} \rangle$. Thus $A_i = \begin{pmatrix} p_i \\ q_i \end{pmatrix} \begin{pmatrix} \gamma_i & 1 \end{pmatrix}$ for some $p_i,\ q_i \in C$.

We can further solve for p_i and q_i. Since

$$\mathrm{Im} A_+(z_i) = \langle \begin{pmatrix} 1 \\ -\gamma_i \end{pmatrix} \rangle,\ Ker A_+^*(z_i) = Ker A_+^{-1}(1/\bar{z}_i) = \langle \begin{pmatrix} \bar{\gamma_i} \\ 1 \end{pmatrix} \rangle.$$

Hence,

$$A_+^{-1}(\frac{1}{\bar{z}_i}) \begin{pmatrix} \bar{\gamma_i} \\ 1 \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ \bar{b} & a_2 \end{pmatrix} \begin{pmatrix} \bar{\gamma_i} \\ 1 \end{pmatrix} + \sum_{j=1}^{j=n} \frac{1}{\bar{z_i}^{-1} - z_j} A_j \begin{pmatrix} \bar{\gamma_i} \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The first component gives us a system of linear equations for p_i which reads as

$$a_1 \bar{\gamma_i} + \sum_{j=1}^{j=n} \frac{\gamma_j \bar{\gamma_i} + 1}{\bar{z_i}^{-1} - z_j} p_j = 0 \text{ , for all } i.$$

Thus

$$M(0) \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix} = -a_1 \begin{pmatrix} \bar{\gamma_1} \\ \vdots \\ \bar{\gamma_n} \end{pmatrix}.$$

Observe that

$$A_+^{-1}(0) = \begin{pmatrix} 1/a_1 & -b/(a_1 a_2) \\ 0 & 1/a_2 \end{pmatrix} = \begin{pmatrix} a_1 & 0 \\ \bar{b} & a_2 \end{pmatrix} - \sum_{i=1}^{i=n} \frac{1}{z_i} \begin{pmatrix} p_i \gamma_i & p_i \\ q_i \gamma_i & q_i \end{pmatrix}$$

And the $(1,2)$ component tells us $-b/(a_1a_2) = -\sum p_i/z_i$ or

$$F_0 = -b/a_1 = a_2 \begin{pmatrix} -\frac{1}{z_1} & \cdots & -\frac{1}{z_n} \end{pmatrix} \begin{pmatrix} p_1 \\ \vdots \\ p_n \end{pmatrix}$$

$$= a_1a_2 \begin{pmatrix} \frac{1}{z_1} & \cdots & \frac{1}{z_n} \end{pmatrix} M(0)^{-1} \begin{pmatrix} \bar{\gamma_1} \\ \vdots \\ \bar{\gamma_n} \end{pmatrix}.$$

Since $a_1a_2 = \det A_+(0) = a^*(0) = \prod |z_i|$, together with Cramer's rule,

$$F_0 = -\left(\prod |z_i|\right) \frac{\det M_1(0)}{\det M(0)}.$$

To obtain F_1, we switch the cutting point from $n = 0$ to $n = 1$. That means we apply the layer stripping method for new A_+^1 which is

$$\begin{aligned} &\begin{pmatrix} a^*_{\geq 1} & -b_{\geq 1}/z \\ -zb^*_{\leq 0} & a^*_{\leq 0} \end{pmatrix} \\ &= (1+|F_0|^2)^{-1/2} \begin{pmatrix} a^*_{\geq 0} + \bar{F}_0 b_{\geq 0} & z^{-1}(F_0 a^*_{\geq 0} - b_{\geq 0}) \\ -z(\bar{F}_0 a^*_{<0} + b^*_{<0}) & a^*_{<0} - F_0 b^*_{<0} \end{pmatrix} \\ &= (1+|F_0|^2)^{-1/2} \begin{pmatrix} z^{-1/2} & 0 \\ 0 & z^{1/2} \end{pmatrix} A_+ \begin{pmatrix} 1 & F_0 \\ -\bar{F}_0 & 1 \end{pmatrix} \begin{pmatrix} z^{1/2} & 0 \\ 0 & z^{-1/2} \end{pmatrix} \end{aligned}$$

It is clear that A_+^1 degenerates at z_i with $\mathrm{Im}A_+^1(z_i) = \langle \begin{pmatrix} 1 \\ -\gamma_i z_i \end{pmatrix} \rangle$. The layer stripping method works similarly on new A_+^1 with new coefficients $\{\gamma_i z_i\}$. Thus

$$F_1 = -\prod |z_i| \frac{\det M_1(1)}{\det M(1)}.$$

And by induction, we prove the formula for all $k \in Z$.

We prove the claim that (F_k) is geometrically decaying. Since after applying k steps of layer stripping method, new $\gamma_i^k = \gamma_i z_i^k = \mathbf{O}(|z_i|^k)$ as $k \to +\infty$, $\mathrm{Im}A_+(z_i) \to \langle \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rangle$ with the same order. i.e. $|\gamma_i| = \mathbf{O}(|z_i|^k)$. Thus, the Blaschke-Potapov factor is of the form

$$B_i = \begin{pmatrix} 1 & 0 \\ 0 & -\frac{\bar{z}_i}{|z_i|} \frac{z - z_i}{1 - \bar{z}_i z} \end{pmatrix} + \mathbf{O}(r^k)$$

where $r = \max\{|z_1|, |z_2|, \ldots, |z_n|\}$. Therefore,

$$A_+^k = \begin{pmatrix} 1 & 0 \\ 0 & \prod_{i=1}^n -\frac{\bar{z}_i}{|z_i|} \frac{z - z_i}{1 - \bar{z}_i z} \end{pmatrix} + \mathbf{O}(r^k) \text{ and}$$

$$|F_k| = \left| \frac{(A_+^k(0))_{12}}{(A_+^k(0))_{11}} \right| = \mathbf{O}(r^k).$$

As $k \to -\infty$, $\mathrm{Im}A_+(z_i) \to \langle \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rangle + \mathbf{O}(|z_i|^{-k})$. And

$$A_+^k = \begin{pmatrix} \prod_{i=1}^n -\frac{\bar{z}_i}{|z_i|} \frac{z - z_i}{1 - \bar{z}_i z} & 0 \\ 0 & 1 \end{pmatrix} + \mathbf{O}(r^{-k}).$$

Hence $|F_k| = \mathbf{O}(r^{-k})$ as $k \to -\infty$.

If $\gamma_i \neq 0, \infty$ for all i, then $(a_+, b_+) \in \mathbf{H}$ and $(a_-, b_-) \in \mathbf{H}_0^*$. From the proof of Lemma 3.3, we know that $\overbrace{(F_{\geq 0})} = (a_+, b_+)$, $\overbrace{(F_{<0})} = (a_-, b_-)$ and thus $\overbrace{(F)} = (a, 0)$. Otherwise, $\gamma_i = 0$ or ∞ for some i. Since $\gamma_i = 0$ means $a_-^*(z_i) = b_-^*(z_i) = 0$ and $\gamma_i = \infty$ implies $a_+^*(z_i) = b_+(z_i) = 0$. Thus a_+^* and b_+ have common inner factor B^1 while a_-^* and b_-^* have common inner factor B^2. Thus Lemma 3.3 says that $\overbrace{(F_{\geq 0})}((B^1)^*, 0) = (a_+, b_+)$ and $((B^2)^*, 0)\overbrace{(F_{<0})} = (a_-, b_-)$.

□

For the one simple zero case, i.e. $a^* = B$ where B is the Blaschke factor vanishing at z_0, it seems that the inverse potential is described by a parameter $\gamma \in C \setminus \{0\}$. By the previous lemma,

$$F_n(\gamma) = \left(\frac{1}{|z_0|} - |z_0|\right) \frac{\bar{\gamma}\bar{z}_0^n}{1 + |\gamma|^2 |z_0|^{2n}}.$$

Hence,

$$F_n(c\gamma) = \bar{c}F_n(\gamma) \text{ for all } |c| = 1, \text{ and } F_n(z_0\gamma) = F_{n+1}(\gamma)$$

which means that the changes of the phase of γ only cause the changes of the phase of the inverse potential and multiplying γ by z_0 causes the shift of the potential. Lemma (3.2) says that if $\overbrace{F_n} = (a, b)$ then $\overbrace{cF_n} = (a, cb)$ for all $|c| = 1$ and $\overbrace{F_{n+1}} = (a, bz^{-1})$. Thus it may be natural to quotient by the equivalence relation:

$$(F_n) \sim (G_n) \text{ if there is some } |c| = 1 \text{ and } m \in \mathbf{N} \text{ such that } (F_n) = (cG_{n+m}).$$

After quotienting this equivalence relation, we can reduce the parameter γ to the interval $(|z_0|^2, |z_0|)$. However inside this interval, it is not clear how to relate potentials from different parameters.

We have constructed the inverse potential for rational $(a, 0)$ where a^* only has simple zeros. Based on this we can construct the inverse potential for general rational $(a, 0)$. Given

$$a^* = \prod_{i=1}^{k_0} B_i^{n_i}\ ,\ B_i = \frac{-\bar{z}_i}{|z_i|}\frac{z - z_i}{1 - \bar{z}_i z}\ ,\ z_i \neq z_j \text{ for all } i \neq j,$$

there is a bijection between the half line decomposition $(a_-, b_-)(a_+, b_+) = (a, 0)$ and the data $\{\gamma_i^j\}$ satisfying (**??**), (**??**). With such $\{\gamma_i^j\}$, we have

$$A_+(z) = \begin{pmatrix} a_+^* & -b_+ \\ -b_-^* & a_-^* \end{pmatrix} = \prod_{i=1}^{k_0}\prod_{j=1}^{n_i}(I - P_{ij} + B_i(z)P_{ij})$$

$$\text{where } P_{ij} = \frac{1}{1 + |\delta_{ij}|^2}\begin{pmatrix} |\delta_{ij}|^2 & \bar{\delta}_{ij} \\ \delta_{ij} & 1 \end{pmatrix}$$

and $\{\delta_{ij}\}$ are determined by $\{\gamma_i^j\}$. We approach A_+ by $\tilde{A}_+(z) = \prod_{i=1}^{k}\prod_{j=1}^{n_i}(I - P_{ij} + B_{ij}(z)P_{ij})$ where

$$B_{ij}(z) = \frac{-\bar{\omega}_{ij}}{|\omega_{ij}|}\frac{z - \omega_{ij}}{1 - \bar{\omega}_{ij} z}\ ,\ \text{with } \omega_{i1} = z_i \text{ for all i}, \omega_{ij} \neq \omega_{st} \text{ for all } (i, j) \neq (s, t)$$

and let $\omega_{ij} \to z_i$ for all $2 \le j \le n_i$. Observe that $\tilde{A}_+$ has simple zeros $\{\omega_{ij}\}$ and

$$\begin{aligned}\mathrm{Im}\tilde{A}_+(\omega_{ij}) = \langle \prod_{s=1}^{i-1}\prod_{t=1}^{n_s}(I - P_{st} + B_{st}(\omega_{ij})P_{st})\times \\ (I - P_{i1} + B_{i1}(\omega_{ij})P_{i1})(I - P_{i2} + B_{i2}(\omega_{ij})P_{i2})\cdots \\ (I - P_{ij-1} + B_{ij-1}(\omega_{ij})P_{ij-1})\begin{pmatrix} 1 \\ -\delta_{ij}\end{pmatrix}\rangle\end{aligned}$$

Define Γ_{ij} such that $\mathrm{Im}\tilde{A}_+(\omega_{ij}) = \langle\begin{pmatrix} 1 \\ -\Gamma_{ij}\end{pmatrix}\rangle$. Then it is clear that Γ_{ij} is a rational function of variables $\{\omega_{st}\}$ for $1 \le s \le i-1$, $1 \le t \le n_s$ and $\{\omega_{il}\}$ for $1 \le l \le j$.

When ω_{ij} is close to z_i,

$$\mathrm{Im}\tilde{A}_+(\omega_{ij}) \sim \langle \prod_{s=1}^{i-1}\prod_{t=1}^{s_i}(I - P_{st} + B_s(z_i)P_{st})\begin{pmatrix} 1 \\ -\delta_{i1}\end{pmatrix}\rangle = \mathrm{Im}A_+(z_i).$$

Since by assumption $\gamma_i^1 \ne 0, \infty$, we have $\Gamma_{ij} \ne 0, \infty$ for all (i,j). Thus we can apply the previous lemma and obtain the inverse potential $(\tilde{F}_n)$ for $\tilde{A}_+$.

For a better notation, we replace the double index (i,j) by single index $k = k(i,j) = (\sum_{s=1}^{i-1} n_s) + j$, where $1 \le k \le \sum_{i=1}^{k_0} n_i$, and let $N = \sum_{i=1}^{k_0} n_i$. Then

$$\tilde{F}_n = -\prod_k |\omega_{k(i,j)}|\frac{\det \tilde{M}_1(n)}{\det \tilde{M}(n)}$$, where $\tilde{M}(n)$ is a $N \times N$ matrix such that

$$(\tilde{M}(n))_{kl} = \frac{1 + \bar{\Gamma}_k\Gamma_l\bar{\omega}_k^n\omega_l^n}{\bar{\omega}_k^{-1} - \omega_l} \text{ and } \tilde{M}_1(n) = \begin{pmatrix} & & & \bar{\Gamma}_1\bar{\omega}_1^n \\ & \tilde{M}(n) & & \vdots \\ & & & \bar{\Gamma}_N\bar{\omega}_N^n \\ \omega_1^{-1} & \cdots & \omega_N^{-1} & 0\end{pmatrix}.$$

Observe that as $\omega_{k(i,j)} \to \omega_{k(i,l)}$ for some $l < j$, $(I - P_{il} + B_{il}(\omega_{ij})P_{il}) \to I - P_{il}$. Therefore $\mathrm{Im}\tilde{A}_+(\omega_{ij}) \to \mathrm{Im}\tilde{A}_+(\omega_{il})$ i.e. $\Gamma_{k(i,j)} \to \Gamma_{k(i,l)}$ and $k(i,j)$ column (row) of $\tilde{M}(n)$ converges to $k(i,l)$ column (row) of $\tilde{M}(n)$. And it is also true for $\tilde{M}_1(n)$. Hence to compute $\det \tilde{M}(n)$ we can do the following process:

Step 1.: Subtract $k(i,1)$ column from $k(i,2), k(i,3), \ldots, k(i,n_i)$ columns and pull out the factor $\omega_{k(i,j)} - \omega_{k(i,1)}$ from $k(i,j)$ column for all i, $j > 1$.

Note that after step 1, $k(i,j)$ column still converges to $k(i,2)$ column as $\omega_{k(i,j)} \to \omega_{k(i,2)}$ for $j > 2$. So we go further to the next step:

Step 2.: Subtract $k(i,2)$ column from $k(i,3), k(i,4), \ldots, k(i,n_i)$ columns and pull out the factor $\omega_{k(i,j)} - \omega_{k(i,2)}$ from $k(i,j)$ column for all i, $j > 2$.

And we can repeat this process inductively so that in step n, we subtract $k(i,n)$ column from $k(i,n+1), k(i,n+2), \ldots, k(i,n_i)$ columns and pull out the factor $\omega_{k(i,j)} - \omega_{k(i,n)}$ from $k(i,j)$ column for all i, $j > n$.

After $\max_i n_i - 1$ steps we will stop. Then we repeat the whole process for rows but pull out the factor $\bar{\omega}_{k(i,j)} - \bar{\omega}_{k(i,n)}$ (instead of $\omega_{k(i,j)} - \omega_{k(i,n)}$) from $k(i,j)$ row in step n for all n. Similarly, to compute $\det \tilde{M}_1(n)$ we apply the same process on first N columns and first N rows of $\tilde{M}_1(n)$.

Then when taking the limit $\omega_{k(i,j)} \to \omega_{k(i,1)} = z_i$, we will have

$$\tilde{F}_n \to F_n = -\prod_{i=1}^{k_0} |z_i|^{n_i} \frac{M(n)}{M_1(n)}$$

where $M(n)$ is obtained from $\tilde{M}(n)$ by applying operators $\frac{1}{(j-1)!}\frac{\partial^{j-1}}{\partial\omega_{k(i.j)}^{j-1}}$ to the $k(i,j)$ column of $\tilde{M}(n)$, operators $\frac{1}{(j-1)!}\frac{\partial^{j-1}}{\partial\bar{\omega}_{k(i.j)}^{j-1}}$ to the $k(i,j)$ row of $\tilde{M}(n)$ and taking value at $\omega_{k(i,j)} = z_i$ for all i, j. $M_1(n)$ is obtained from $\tilde{M}_1(n)$ in the same way.

We now prove that (F_n) is bounded by a geometric sequence.

First claim that shifting the cutting point from 0 to n can be viewed as changing the coefficients from $\{\gamma_i^j\}$ to $\{\gamma_i^j(n)\}$ where $\gamma_i^1(n) = \gamma_i^1 z_i^n$ and $\gamma_i^j(n) z_i^{-n} \to c_{i,j,+}$ as $n \to +\infty$, $(\gamma_i^j(n))^{-1} z_i^n \to c_{i,j,-}$ as $n \to -\infty$ where $c_{i,j,+}$, $c_{i,j,-} \in C$ are some constants. Then the same argument in the proof of Lemma 3.25 says that

$$A_+^n = \begin{pmatrix} 1 & 0 \\ 0 & \prod B_i^{n_i} \end{pmatrix} + \mathbf{O}(r^n) \text{ as } n \to +\infty$$

$$A_+^n = \begin{pmatrix} \prod B_i^{n_i} & 0 \\ 0 & 1 \end{pmatrix} + \mathbf{O}(r^{-n}) \text{ as } n \to -\infty$$

where $r = \max\{|z_1| \dots |z_{k_0}|\}$. And thus $|F_n| = |(A_+^n(0))_{12}/(A_+^n(0))_{11}| = \mathbf{O}(r^{|n|})$ for $|n| \to \infty$.

Define $U_z = \begin{pmatrix} \frac{1}{\sqrt{z}} & 0 \\ 0 & \sqrt{z} \end{pmatrix}$, and $T_n = (1 + |F_n|^2)^{-1/2} \begin{pmatrix} 1 & F_n \\ -\bar{F}_n & 1 \end{pmatrix}$. From the proof of Lemma 3.25, we know that changing the cutting point from 0 to n gives us new A_+^n which is

$$\begin{pmatrix} a_{\geq n}^* & -\frac{b_{\geq n}}{z^n} \\ -z^n b_{<n}^* & a_{<n}^* \end{pmatrix} = U_z^n A_+ T_0 U_z^{-1} T_1 U_z^{-1} \cdots T_{n-1} U_z^{-1}.$$

It is easy to see that A_+^n has image $\langle \begin{pmatrix} 1 \\ -z_i^n \gamma_i^1 \end{pmatrix} \rangle$ at z_i.

Let $\tilde{B}_i^j$, $j = 1 \dots n_i$, be the Blaschke-Potapov factors which degenerate at z_i and $(\tilde{B}_i^{n_i})^{-1}(\tilde{B}_i^{n_i-1})^{-1} \cdots (\tilde{B}_i^1)^{-1} A_+^n$ can be defined analytically at z_i. Then $\mathrm{Im}\tilde{B}_i^j(z_i) = \langle \begin{pmatrix} 1 \\ -\gamma_i^j(n) \end{pmatrix} \rangle$. Suppose we have shown that $(\gamma_i^j(n) z_i^{-n})^{\pm 1} \to c_{i,j,\pm}$ as $n \to \pm\infty$ for $j \leq k$. For $j = k+1$,

$$\begin{aligned} \mathrm{Im}\tilde{B}_i^{k+1}(z_i) &= \mathrm{Im}(\tilde{B}_i^k)^{-1}(\tilde{B}_i^{k-1})^{-1} \cdots (\tilde{B}_i^1)^{-1} \tilde{A}_+(z_i) \\ &= \mathrm{Im}(\tilde{B}_i^k)^{-1} \cdots (\tilde{B}_i^1)^{-1} U_z^n A_+(z_i) \\ &= \mathrm{Im} U_z^n [U_z^{-n} (\tilde{B}_i^k)^{-1} U_z^n] \cdots [U_z^{-n} (\tilde{B}_i^1)^{-1} U_z^n] A_+(z_i). \end{aligned}$$

And the induction assumption says that

$$\mathrm{Im}[U_z^{-n} (\tilde{B}_i^k)^{-1} U_z^n] \cdots [U_z^{-n} (\tilde{B}_i^1)^{-1} U_z^n] A_+(z_i)$$

will converge to some $V_\pm$ as n goes to $\pm\infty$. Now claim that $V_+ \neq \langle \begin{pmatrix} 0 \\ 1 \end{pmatrix} \rangle$ and $V_- \neq \langle \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rangle$. Then $\mathrm{Im}\tilde{B}_i^{k+1}(z_i) \to \mathrm{Im} U_{z_i}^n V_\pm$ as $n \to \pm\infty$ and thus $(\gamma_i^{k+1}(n) z_i^{-n})^{\pm 1} \to c_{i,k+1,\pm}$ as $n \to \pm\infty$ for some $c_{i,k+1,\pm}$. Hence we have proved it by the induction.

Assume that $V_+ = \langle \begin{pmatrix} 0 \\ 1 \end{pmatrix} \rangle$. Then as n large, $W_n(z_i) = [U_z^{-n}(\tilde{B}_i^k)^{-1}U_z^n] \cdots [U_z^{-n}(\tilde{B}_i^1)^{-1}U_z^n]A_+(z_i)$ is of the form

$$\begin{pmatrix} 0 & 0 \\ a_1 & a_2 \end{pmatrix} + \epsilon(n)$$

where a_1, a_2 are constants that are not both zeros and $\epsilon(n) \to 0$ as $n \to +\infty$. Let b_i be the Blaschke factor vanishing at z_i. Then

$$U_z^{-n}\tilde{B}_i^j U_z^n = U_z^{-n}\frac{1}{1+|\gamma_i^j(n)|^2}\begin{pmatrix} 1+b_i|\gamma_i^j(n)|^2 & (b_i-1)\bar{\gamma}_i^j(n) \\ (b_i-1)\gamma_i^j(n) & |\gamma_i^j(n)|^2+b_i \end{pmatrix} U_z^n$$

$$= \frac{1}{1+|\gamma_i^j(n)|^2}\begin{pmatrix} 1+b_i|\gamma_i^j(n)|^2 & (b_i-1)\gamma_i^j(n)z^n \\ (b_i-1)\bar{\gamma}_i^j(n)z^{-n} & |\gamma_i^j(n)|^2+b_i \end{pmatrix}.$$

Thus as $n \to +\infty$,

$$U_z^{-n}\tilde{B}_i^j U_z^n \to \begin{pmatrix} 1 & 0 \\ c_{i,j,+}(b_i-1) & b_i \end{pmatrix} \text{ for } j = 1, \ldots, k.$$

Since

$$A_+ = [U_z^{-n}\tilde{B}_i^1 U_z^n] \cdots [U_z^{-n}\tilde{B}_i^k U_z^n] W_n$$

for all n, we have

$$A_+(z_i) = \begin{pmatrix} 1 & 0 \\ -c_{i,1,+} & 0 \end{pmatrix} \cdots \begin{pmatrix} 1 & 0 \\ -c_{i,k,+} & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ a_1 & a_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

which contradicts to our assumption that $\mathrm{Im}A_+(z_i) \neq \langle \begin{pmatrix} 0 \\ 0 \end{pmatrix} \rangle$. Similarly, we can show that $V_- \neq \langle \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rangle$.

The following lemma characterize all rational half line data $(a_+, b_+) \in \mathbf{H}$ such that after pairing with a $(a_-, b_-) \in \mathbf{H}_0^*$, $(a_-, b_-)(a_+, b_+) = (a, 0)$ is a soliton solution.

Lemma 3.26. *Given rational $(a_+, b_+) \in \mathbf{H}$, there exists $(a_-, b_-) \in \mathbf{H}_0^*$ $(a_-, b_-)(a_k, b_k)$ is a soliton data $(a, 0)$ if and only if $b_+(\infty) = 0$. If $b_+(\infty) \neq 0$, then there is a finite k so that if the sequence of complex numbers $(F) = (F_0, F_1, \cdots, F_k)$ consists of the coefficients of the first k steps of layer stripping for (a_+, b_+), then $(\tilde{a}_+, \tilde{b}_+)$ defined by $(\tilde{a}_+, \tilde{b}_+ z^k) = (\widehat{(F)})^{-1}(a_+, b_+)$ is an element of $\mathbf{H}$ and $\tilde{b}_+(\infty) = 0$.*

Proof. First assume $(a_-, b_-)(a_+, b_+) = (a, 0)$ for some $(a_-, b_-) \in \mathbf{H}_0^*$. Then

$$a_-a_+ - b_-b_+^* = a \tag{3.15}$$

$$a_-b_+ + b_-a_+^* = 0\ . \tag{3.16}$$

Taking a linear combination to cancel b_- gives

$$a_-a_+a_+^* + a_-b_+b_+^* = aa_+^*\ .$$

By the determinant condition for (a_+, b_+) this gives

$$a_- = aa_+^*\ .$$

Evaluating at ∞ and noting that $a(\infty), a_-(\infty) > 0$ gives $a_+^*(\infty > 0$. Now evaluating (3.16) at ∞ and using that $b_-(\infty) = 0$ we obtain $b_+(\infty) = 0$.

Conversely, assume $b_+(\infty) = 0$. Using the determinant condition for a_+ at ∞ gives $a_+(\infty)a_+^*(\infty) = 1$ and hence $a_+^*(\infty) > 0$. Choose a^* to be the Blaschke product of minimal order so that a^*a_+ has no pole in D and set $a_-^* = a^*a_+$. Note that $a_-(\infty) > 0$ since the same is true for a and a_+^*. Now set $b_- = -b_+a$, then clearly we have (**??**). Moreover we have

$$a_-a_-^* + b_-b_-^* = a_+a_+^* + b_+b_+^* = 1$$

and using this one readily checks (3.15). It is then clear that (a_-, b_-) is the desired first factor of the soliton data $(a, 0)$.

Finally, assume that $b_+(\infty)$ is not equal to 0, hence it is finite of ∞. Note that the determinant condition

$$a_+a_+^* + b_+b_+^* = 1$$

evaluated at ∞ shows that $a_+^*(\infty)$ has a pole at most of the order of b_+ at ∞ because $a_+(\infty)$ is finite and non-zero and $b_+^*(\infty)$ is finite. One step of layer stripping produces new data $(\tilde{a}_+, \tilde{b}_+)$ with

$$\tilde{b}_+ = (1 + |F_0|^2)^{-1/2}(b_+ - F_0a_+^*)$$

and the previous discussion shows that $\tilde{b}_+$ has at most the same order of pole as b_+ at ∞. The potential sequence for $(\tilde{a}_+, \tilde{b}_+)$ vanbishes at 0, in other words $b_+(0) = 0$, and we may shift the sequence back obtaining the NLFT data $(\tilde{a}_+, \tilde{b}_+z^{-1})$. Note that $\tilde{b}_+z^{-1}$ has a pole of one order less than b_+, and it vanishes if b_+ was finite. Thus iterating the layer stripping and shifting one more time than the original order of pole of b_+ at ∞ we arrive at data in $\mathbf{H}$ which is the right factor of a factorization of soliton data. $\square$

$a = a_-/a_+^*$.

Bibliography

[1] M. Ablowitz, D. Kaup, A. Newell, H. Segur, *The inverse scattering transform-Fourier analysis for nonlinear problems.* Studies in Appl. Math. **53** (1974), no. 4, 249–315.

[2] M. Ablowitz, J. Ladik, *Nonlinear differential-difference equations and Fourier Analysis*, J. Math. Phys. **17** (1996), 1011–1018.

[3] R. Beals, R. Coifman, *Scattering and inverse scattering for first order systems.* Comm. Pure Appl. Math. **37** (1984), no. 1, 39–90.

[4] K. M. Case *Orthogonal polynomials II.* J. Math. Phys. **16** (1975), 1435-1440.

[5] M. Christ, A. Kiselev, *Absolutely continuous spectrum for one-dimensional Schrödinger operators with slowly decaying potentials: some optimal results*, J. Amer. Math. Soc. **11** (1998), 771–797.

[6] M. Christ and A. Kiselev, A. *Maximal functions associated to filtrations.* J. Funct. Anal. **179** (2001), no. 2, 409–425.

[7] D. Damanik and R. Killip, *Half-line Schrödinger operators with no bound states.* preprint, (2003).

[8] P. Deift, *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, Courant Lecture Notes in Mathematics **3**. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence RI, 1999.

[9] P. Deift, R. Killip, *On the absolutely continuous spectrum of one-dimensional Schrodinger operators with square summable potentials*, Comm. Math. Phys. **203** (1999), 341–347.

[10] S. Denisov, *Probability measures with reflection coefficients $\{a_n\} \in l^4$ and $\{a_{n+1} - a_n\} \in l^2$ are Erdös measures.* J. Approx. Theory **117** (2002), 42–54.

[11] P.Deift; E. Trubowitz, *Inverse scattering on the line*, Comm. Pure Appl. Math. **32** (1979), no. 2, 121–251.

[12] H. Dym, H.P. McKean, *Gaussian processes, function theory, and the inverse spectral problem*, Probability and Mathematical Statistics, Vol. 31. Academic Press Inc. [Harcourt Brace Jovanich Publishers]. New York-London, 1976.

[13] H. Dym and H. P. McKean, *Gaussian processes, function theory, and the inverse spectral problem.* Probability and Mathematical Statistics, Vol. 31. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, (1976).

[14] L.D. Faddeev and L.A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons.* Springer, Berlin, 1987

[15] A.S. Fokas, I.M. Gelfand, *Integrability of linear and non-linear evolution equations and the associated nonlinear Fourier transform*, Lett. Math. Phys. **32** (1994), 189–210.

[16] J. Garnett, *Bounded Analytic Functions*, Pure and Applied Mathematics **96**. Academic Press Inc. [Harcourt Brace Jovanich Publishers]. New York-London, 1981.

[17] G. Herglotz *Über Potenzreihen mit positivem reellen Teil im Einheitskreis*, Leipziger Berichte **63** (1911), 501–511.

[18] M. Hitrik, *Properties of the Scattering Transform on the Real Line.* J. Math. Anal. Appl. **258** (2001), 223–243.

[19] R. Killip, B. Simon, *Sum Rules for Jacobi Matrices and their Applications to Spectral Theory*, to appear, Ann. Math.

[20] P. Koosis, *The Logarithmic Integral I.*, Cambridge Studies in Advanced Mathematics **12**, Cambridge University Press, Cambridge 1988.

[21] G.S. Litvinchuk and I.M. Spitkovskii, *Factorization of Measurable Matrix Functions.* Birkhäuser Verlag 1987

[22] C. Muscalu, T. Tao, and C. Thiele, *A counterexample to a multilinear endpoint question by Christ and Kiselev.* Math. Res. Lett **10** (2003), 237–246.

[23] C. Muscalu, T. Tao, and C. Thiele, *A Carleson type theorem for a Cantor group model of the scattering transform.* Nonlinearity **16** (2003), 219–246.

[24] I. Schur, *Über Potenzreihen die im Innern des Einheitskreises beschränkt sind.* J. Reine Angew. Math. **147** (1917), 205–232.

[25] B. Simon, *A canonical factorization for meromorphic Herglotz functions on the unit disc and sum rules for Jacobi matrices* preprint, (2003)

[26] E. Stein, *Harmonic Analysis, Real variable methods, orthogonality, and oscillatory integrals.* Princeton Mathematical Series, 43, Princeton University Press, NJ (1993).

[27] J. Sylvester, D. Winebrenner, *Linear and nonlinear inverse scattering*, SIAM J. Appl. Math. **59** (1999), 669–699.

[28] G. Szëgo, *Orthogonal polynomials*, Fourth Edition, American Mathematical Society, Colloquium Publications Vol. XXIII, American Mathematical Society, Providence RI 1975.

[29] G. Szëgo, *Über der Entwieklung einer analytischen Funktion nach den Polynomen eines Orthogonalsystems*, Mathematische Annalen **82** (1921), 188–212.

[30] S. Tanaka, *Some Remarks on the Modified Korteweg- de Vries Equations.* Publ. RIMS, Kyoto Univ. **8** (1972/73), 429–437.

[31] T. Tao and C. Thiele, *IAS/Park City Mathematics Series Vol 13: Nonlinear Fourier Analysis.* (in preparation)

[32] S. Verblunsky, *On positive harmonic functions II.*, Proc. London Math. Soc. **40** (1936), 290–320.

[33] Verblunsky, S. *On positive harmonic functions II.* Proc. London Math. Soc. (2) **40** (1936), 290-320.

[34] A. Volberg, P. Yuditskii, *On the inverse scattering problem for Jacobi matrices with the spectrum on an interval, a finite system of intervals, or a Cantor set of positive length*, Comm. Math. Phys. **226** (2002), 567–605.

[35] K. Yajima, *The $W^{k,p}$ continuity of wave operators for Schrodinger operators*, J. Math. Soc. Japan **47** (1995), 551–581.