Pacific Journal of Mathematics

Volume 204 No. 2 June 2002

PACIFIC JOURNAL OF MATHEMATICS

http://www.pjmath.org

Founded in 1951 by

E. F. Beckenbach (1906–1982) F. Wolf (1904–1989)

See inside back cover or www.pjmath.org for submission instructions.

Regular subscription rate for 2006: $425.00 a year (10 issues). Special rate: $212.50 a year to individual members of supporting institutions. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163, U.S.A. Prior back issues are obtainable from Periodicals Service Company, 11 Main Street, Germantown, NY 12526-5635. The Pacific Journal of Mathematics is indexed by Mathematical Reviews, Zentralblatt MATH, PASCAL CNRS Index, Referativnyi Zhurnal, Current Mathematical Publications and the Science Citation Index.

The Pacific Journal of Mathematics (ISSN 0030-8730) at the University of California, c/o Department of Mathematics, 969 Evans Hall, Berkeley, CA 94720-3840 is published monthly except July and August. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices. POSTMASTER: send address changes to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163.

PUBLISHED BY PACIFIC JOURNAL OF MATHEMATICS
at the University of California, Berkeley 94720-3840
A NON-PROFIT CORPORATION
Typeset in LaTeX

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

APPLYING FUNCTIONAL IDENTITIES TO SOME LINEAR PRESERVER PROBLEMS

K.I. Beidar, M. Brešar, M.A. Chebotar, and Y. Fong

The theory of functional identities is used to obtain algebraic generalizations of some operator-theoretic results concerning commutativity and normal preserving linear maps between algebras with involution.

1. Introduction.

Over the last decades there has been a considerable interest in linear algebra and operator theory in the so-called linear preserver problems (see survey articles [**1**, **13**, **19**, **20**]). By a linear preserver we mean a linear map of algebras which, roughly speaking, preserve certain properties of some elements in an algebra. In the literature these algebras are usually algebras of matrices or algebras of bounded linear operators. The goal in the study of linear preservers is to find their form. It turns out that often the only solutions are just the most obvious ones, frequently (anti)isomorphisms or at least maps related to them.

It is our impression that some linear preserver problems could be solved in a more general setting using only ring-theoretic techniques. An example illustrating this general conjecture is a characterization of bijective linear maps of prime algebras that preserve commutativity, i.e., they map commuting pairs of elements into commuting pairs [**10**, Theorem 2] (see also [**2**, **4**] for some generalizations). This characterization was known before only for some special prime algebras which are studied in linear algebra and operator theory (see [**10**] for references). Moreover, it has turned out that one does not really need to assume that the map, say θ, preserves the commutativity of all elements, but only that $\theta(x)$ and $\theta(x^2)$ commute for every x. The fact that only this milder condition has to be asssumed has proved to be useful when this result was applied to another linear preserver problem, namely, the one concerning maps on the algebra of bounded linear operators on a Hilbert space that preserve normal operators [**12**].

The proof of [**10**, Theorem 2] was based on a characterization of commuting traces of biadditive maps [**10**, Theorem 1], which was one of the first results in the area which is now called the theory of functional identities in rings. Over the last few years this theory has been systematically developed. It is our goal in this paper to show that some of its most recent results

[**5**, **6**, **7**] can be used to obtain some further improvements in the study of linear maps preserving commutativity or normal elements.

First we introduce some terminology and fix the notation. A prime algebra A over a field F is said to be centrally closed over F if both the center and the extended centroid of A are equal to F. We will consider centrally closed prime algebras A, A' over F with involution $*$ (by an involution we mean an additive involution, that is, $*$ satisfies $(x+y)^* = x^* + y^*$, $(xy)^* = y^*x^*$ and $(x^*)^* = x$). We say that a linear map $\theta : A \to A'$ is $*$-linear if $\theta(x^*) = \theta(x)^*$ for all $x \in A$. Let $S = \{x \in A \mid x^* = x\}$ be the set of all symmetric elements in A, and $K = \{x \in A \mid x^* = -x\}$ be the set of all skew elements in A. Similarly, by S' and K' we denote the sets of all symmetric and skew elements in A', respectively. Next we set $F_s = F \bigcap S$. We say that the involution $*$ is of the first kind if $F = F_s$ (equivalently, $*$ is F-linear); otherwise we say that $*$ is of the second kind. Given a subset R of A, we write $\langle R \rangle$ for the subalgebra of A generated by R.

The concepts of the extended centroid and centrally closed prime algebras are explained in detail in the book [**8**]. Nevertheless, as some readers may be primarily interested in what is the meaning of our results in linear algebra and operator theory, let us just mention that the algebras of square matrices over a division ring, bounded linear operators on a Banach space (and moreover, all its subalgebras containing the identity and all finite rank operators) and prime unital C^*-algebras (in particular, von Neumann factors) are all examples of prime algebras centrally closed over their centers. Also, if one wants to restrict the attention to the case when A and A' are algebras consisting of linear operators on a Hilbert space H and x^* is the adjoint of the operator x, then $*$ is of the first kind when H is a real space, and $*$ is of the second kind when H is a complex space.

In Section 2 we gather together some results of the theory of functional identities that are needed later on.

In Section 3 we extend the treatment of commutativity-preserving maps of prime algebras [**10**] by considering maps from S onto S'. The result which we obtain is a ring-theoretic extension of the results on maps preserving commutativity of symmetric matrices (operators) [**15**, **14**, **24**]. Actually, as in [**10**], we do not really assume that the map θ preserves the commutativity of all elements in S, but only that $\theta(s)$ and $\theta(s^2)$ commute for every $s \in S$.

In Section 4 we consider bijective linear maps of algebras with involution of the second kind which preserve normal elements. As already mentioned, the special case when the algebras under consideration are algebras of all bounded linear operators on a complex Hilbert space was treated in [**12**] (see also [**15**, **16**]). It has turned out that Fuglede's theorem [**25**, Corollary 1.18], upon which the proof in [**12**] depends, can be avoided when treating this problem, and so we will be able to prove a ring-theoretic generalization of the result of [**12**].

The problem of describing normal-preserving maps is much more difficult when the involution is of the first kind. First of all, the involution is then a linear operator and so, for instance, a map of the form $x \mapsto \mu_1 x + \mu_2 x^*$, where $\mu_1, \mu_2 \in F$, is linear and preserves normal elements. Thus, we cannot expect the same result as in Section 4. Moreover, consider the following example.

Let $A = F\langle x, y\rangle$, where F is a field, be a free algebra in two indeterminates x and y (incidentally, A is a centrally closed prime algebra [**8**, Theorem 2.4.4]), and equip A with standard involution (given by $x^* = x$, $y^* = y$ and $\lambda^* = \lambda$, $\lambda \in F$). Let U be a linear span of all monomials in which both x and y appear, and V be a linear span of all monomials x^n, y^n with $n \geq 1$ (in particular, $V \subset S$). Then $A = F \oplus U \oplus V$ and note that a nonzero element in U never commutes with a nonzero element in V. Now let $T : V \to V$ be any bijective linear operator. Then the map $A \to A$ defined by $\lambda + u + v \mapsto \lambda + u + T(v)$ is $*$-linear, bijective and maps normal elements onto normal elements.

This example somehow indicates that it is almost impossible to obtain a definitive result for preservers of normal elements in the case the involution is of the first kind. Nevertheless, even in this example the map acts very simply on a rather large piece of A, namely on $F \oplus U$. In Section 5 we shall see that under some technical conditions (in particular, we have to assume that our map is $*$-linear) the action of normal-preservers can be described on $\langle K\rangle$, which can certainly be considered as a "large piece" of A. In particular, except in some very special case, it contains a nonzero ideal of A [**8**, Lemma 9.1.4 and Corollary 9.1.8]. Therefore, in simple algebras satisfying our technical assumptions, normal-preserving maps can be completely determined.

2. Preliminaries.

The aim of this section is to give a brief and self-contained outline of some parts the theory of functional identities, namely, those parts that shall really be needed in the subsequent sections. For a more detailed account on this theory we refer the reader to [**11**].

Throughout, the denotations F, F_s, A, K, S, A', K' and S' shall have the meaning already explained in the introduction. Though not always needed, we assume for simplicity that $\mathrm{char}(F) \neq 2$. Given $x, y \in A$, we set

$$[x, y] = xy - yx \quad \text{and} \quad x \circ y = xy + yx.$$

Next, by $\deg(x)$ we shall mean the degree of x over F (if x is algebraic over F) or ∞ (if x is not algebraic over F). Next we set

$$\deg(A) = \sup\{\deg(x) \mid x \in A\}.$$

For instance, $\deg(M_n(F)) = n$ for any field F. Moreover, from the structure theory of rings with polynomial identities [**23**, **26**] it can be deduced that $\deg(A) = n < \infty$ if and only if A is a subring of $M_n(\overline{F})$ such that $\overline{F}A = M_n(\overline{F})$, where $\overline{F}$ is the algebraic closure of F.

The goal in the study of functional identities is, roughly speaking, to describe the form of maps satisfying certain identities. The first functional identities that have been considered were those concerned with the so-called commuting maps, i.e., maps whose values commute with the variable. Let us now reword the basic result on commuting maps [**9**, Theorem 3.2] in the following somewhat unusual but useful form.

Theorem 2.1. *Let $f, \theta : A \to A'$ be linear maps such that $[f(x), \theta(x)] = 0$ for all $x \in A$. If θ is bijective, then there is $\tau \in F$ and a linear map $\zeta : A \to F$ such that $f(x) = \tau\theta(x) + \zeta(x)$ for all $x \in A$.*

Actually, [**9**] considers only the case when $A = A'$ and θ is the identity map. However, the seemingly more general condition treated in Theorem 2.1 can be reduced to that one by replacing the map f by the map $f\theta^{-1}$. We also remark that in [**9**] the result is stated for additive maps on rings and not linear maps on algebras, but the necessary modifications in the proof are obvious. The same remarks apply for the remaining results in this section. Moreover, in these three theorems such terms as linearity and vector space should be understood with respect to the field F_s rather than F.

A map $q : A \to A'$ is said to be a trace of a k-linear map if there is a map $B : A^k \to A'$, linear in each argument and such that $q(x) = B(x, \dots, x)$ for all $x \in A$ (by a trace of a 0-linear map we shall mean a constant). In the case when $\operatorname{char}(F) = 0$ or $\operatorname{char}(F) > k$, there is no loss of generality in assuming that this map B is symmetric (namely, otherwise replace $B(x_1, \dots, x_k)$ by $\frac{1}{k!}\sum_{\pi \in S_k} B(x_{\pi(1)}, \dots, x_{\pi(k)})$).

The next theorem follows from [**7**, Theorem 5.5] and [**5**, Lemma 2.2].

Theorem 2.2. *Let R be a vector subspace of A, and let R' be either S' or K'. Suppose that a trace of an n-linear map $q : R \to A'$ satisfies*

$$\sum_{i=0}^{m} \mu_i \theta(x)^i q(x) \theta(x)^{m-i} = 0 \quad \text{for all } x \in R,$$

where $\mu_0, \dots, \mu_m$ belong to F and not all of them are 0, and $\theta : R \to R'$ is a bijective linear map. Suppose that $\operatorname{char}(F) = 0$ or $\operatorname{char}(F) > n$ and $\deg(A') > 2(m+n)$. Then:

(i) *$q(x) = \sum_{k=0}^{n} \lambda_k(x)\theta(x)^{n-k}$, $x \in R$, where each $\lambda_k : R \to F$ is a trace of a k-linear map;*

(ii) *if $\sum_{i=0}^{m} \mu_i \neq 0$, then $q = 0$.*

Keeping the notation of Theorem 2.2, assume that $q(x)\theta(x) \in F$ for all $x \in R$, where $q(x) = B(x, \dots, x)$ and B is an n-linear map. A standard

approach, the so-called complete linearization, then shows that

$$\sum_{\pi\in S_{n+1}} B(x_{\pi(1)},\dots,x_{\pi(n)})\theta(x_{\pi(n+1)}) \in F.$$

Applying [**7**, Lemma 4.3] together with [**5**, Lemma 2.2] we then obtain:

Theorem 2.3. *Let R be a vector subspace of A, and let R' be either S' or K'. Suppose that $q : R \to A'$ is a trace of an n-linear map $B : R^n \to A'$ such that $q(x)\theta(x) \in F$ for all $x \in R$ (or $\theta(x)q(x) \in F$ for all $x \in R$), where $\theta : R \to R'$ is a bijective linear map. If $\deg(A') > 2n + 2$, then $\sum_{\pi\in S_n} B(x_{\pi(1)},\dots,x_{\pi(n)}) = 0$. Thus, if $\operatorname{char}(F) = 0$ or $\operatorname{char}(F) > n$, then $q = 0$.*

We conclude this section with a result which might appear somewhat strange. However, the conditions treated in this result really appear in the proof of Theorem 5.1.

Theorem 2.4. *Let $f(x_1,\dots,x_m)$ be a multilinear polynomial in noncommuting variables $x_1,\dots,x_m$ such that $f(k_1,\dots,k_m) \in K$ for all $k_1,\dots,k_m \in K$. Let ϕ be a linear map of K onto K' such that*

$$\phi(f(k_1,\dots,k_m)) = \lambda f(\phi(k_1),\dots,\phi(k_m)) \quad \text{for all } k_1,\dots,k_m \in K,$$

where λ is a nonzero element in F. Further, let a map $B : K \times K \to A'$ be such that

$$B(k,l) = -B(l,k)$$

for all $k, l \in K$, and

$$\begin{aligned} &B(f(k_1,\dots,k_m),l)\\ &= \sum_{i=1}^{m} \lambda f(\phi(k_1),\dots,\phi(k_{i-1}),B(k_i,l),\phi(k_{i+1}),\dots,\phi(k_m)) \end{aligned}$$

for all $k_1,\dots,k_m,l \in K$. If $\deg(A') > 4m + 1$, then there exists $\rho \in F$ such that

$$B(k,l) - \rho[\phi(k),\phi(l)] \in F \quad \text{for all } k,l \in K.$$

For $\lambda = 1$, Theorem 2.4 can be deduced at once from the statements of [**5**, Theorems 2.4] and [**6**, Theorem 2.9]. Almost the same proof, however, still works in the case when λ is any nonzero element in F.

We have seen that excluding algebras of "small" degree one can obtain definite results on functional identities. As a consequence, the proofs of our main results will work as long as the degree of the algebra will be big enough.

3. Commutativity-preservers on symmetric elements.

Having Theorems 2.2 and 2.3 in hand, the following result can be easily obtained just by modifying the proof of [**10**, Theorem 2].

Theorem 3.1. *Let A be A' be centrally closed prime algebras over a field F with involution. Let $\theta : S \to S'$ be a bijective F_s-linear map such that $\theta(s)$ and $\theta(s^2)$ commute for every $s \in S$. Suppose that* $\deg(A) > 4$, $\deg(A') > 8$ *and* $\mathrm{char}(F) \neq 2, 3$. *Then θ is of the form $\theta(s) = \alpha\phi(s) + \beta(s)$ where $\alpha \in F_s$, $\alpha \neq 0$, β is a F_s-linear map from S into F_s and $\phi : \langle S\rangle \to \langle S'\rangle$ is an F_s-algebra isomorphism.*

Proof. We have $\theta(s)\theta(s^2) - \theta(s^2)\theta(s) = 0$ for all $s \in S$. Clearly, the map $s \mapsto \theta(s^2)$ is a trace of a bilinear map and so Theorem 2.2 implies that

$$\theta(s^2) = \lambda\theta(s)^2 + \mu(s)\theta(s) + \nu(s) \tag{1}$$

where $\lambda \in F$, $\mu : A \to F$ is a linear and $\nu : A \to F$ is a trace of a bilinear map (again, the term linearity refers to F_s and not F). We claim that $\lambda \in F_s$ and $\mu(s), \nu(s) \in F_s$ for all $s \in S$. Indeed, since both $\theta(s)$ and $\theta(s^2)$ are symmetric for any $s \in S$, it follows that

$$\lambda\theta(s)^2 + \mu(s)\theta(s) + \nu(s) = \lambda^*\theta(s)^2 + \mu(s)^*\theta(s) + \nu(s)^*$$

and so

$$\{(\lambda - \lambda^*)\theta(s) + \mu(s) - \mu(s)^*\}\theta(s) \in F \quad \text{for all } s \in S.$$

Since $\deg(A')$ is, in particular, > 4, Theorem 2.3 first gives

$$(\lambda - \lambda^*)\theta(s) + \mu(s) - \mu(s)^* = 0 \quad \text{for all } s \in S,$$

which in turn implies, again by Theorem 2.3, that $\lambda = \lambda^*$ and $\mu(s) = \mu(s)^*$. But then also $\nu(s) = \nu(s)^*$ for any $s \in S$.

Next we claim that $\theta(1)$ is a central element, that is, it lies in F_s. Just as in [**10**, p. 535], substituting $s + 1$ for s in $[\theta(s^2), \theta(s)] = 0$ we arrive at $[\theta(s^2 + s), \theta(1)] = 0$, and then repeating the same substitution we get that $[\theta(s), \theta(1)] = 0$ for all $s \in S$. But then, since $\deg(A') > 2$, applying Theorem 2.2 again (or just referring to the standard theory of rings with involution) we infer that $\theta(1)$ is central.

Suppose that $\lambda = 0$. Then, since θ is F_s-linear, it follows that $\theta(s^2 - \mu(s)s) \in F_s$, which in turn implies, again using the F_s-linearity of θ together with $\theta(1) \in F_s$, that $(s - \mu(s))s \in F_s$ for all $s \in S$. Since $\deg(A)$ is assumed to be > 4, Theorem 2.3 yields $s - \mu(s) = 0$, which contradicts the assumption $\deg(A) > 2$. Therefore, $\lambda \neq 0$.

Now define $\varphi : S \to S'$ by

$$\varphi(s) = \lambda\theta(s) + \frac{1}{2}\mu(s), \quad s \in S.$$

Using (1) one can check that $\varphi(s^2) - \varphi(s)^2 \in F$ for all $s \in S$. Hence $\varphi(s \circ t) = \varphi(s) \circ \varphi(t) + \tau(s,t)$ for all $s,t \in S$, where $s \circ t = st + ts$ and $\tau(s,t) \in F$. Since $(s \circ s) \circ (s \circ s) = \{(s \circ s) \circ s\} \circ s$ for all $s \in S$, we have

$$\begin{aligned} 0 &= \varphi([(s \circ s) \circ (s \circ s) - \{(s \circ s) \circ s\} \circ s]) \\ &= \{\varphi(s) \circ \varphi(s) + \tau(s,s)\} \circ \{\varphi(s) \circ \varphi(s) + \tau(s,s)\} + \tau(s \circ s, s \circ s) \\ &\quad - \{\{\varphi(s) \circ \varphi(s) + \tau(s,s)\} \circ \varphi(s) + \tau(s \circ s, s)\} \circ \varphi(s) + \tau(\{s \circ s\} \circ s, s) \\ &= 4\tau(s,s)\varphi(s)^2 - 4\tau(s^2,s)\varphi(s) + 2\tau(s,s)^2 + 4\tau(s^2,s^2) - 4\tau(s^3,s) \end{aligned}$$

for all $s \in S$. Substituting $\lambda\theta(s) + \frac{1}{2}\mu(s)$ for $\varphi(s)$ we see that

$$[\lambda\tau(s,s)\theta(s) - (\tau(s^2,s) - \tau(s,s)\mu(s))]\theta(s) \in F$$

for all $s \in S$. Using Theorem 2.3 twice we conclude that $\tau(s,s) = 0$ for all $s \in S$ and so φ is a Jordan homomorphism. Let us show that φ is bijective. Basically we shall just repeat arguments given at the end of the proof of [**10**, Theorem 2]. Suppose that $\varphi(s) = 0$. Then $\theta(s) = -\frac{1}{2\lambda}\mu(s) \in F$. Since $0 \neq \theta(1) \in F_s$, $\theta(F_s) = F_s$, and so $s \in F_s$. Therefore the linearity of φ implies that $\varphi(1) = 0$. Since φ is a Jordan homomorphism, this yields

$$2\varphi(t) = \varphi(1 \circ t) = \varphi(1) \circ \varphi(t) = 0 \quad \text{for all } t \in S$$

forcing $\theta(S) \subseteq F_s$ and so $S' = F_s$, a contradiction. Hence φ is injective. Further, $2\varphi(1) = \varphi(1) \circ \varphi(1)$ and $\varphi(1) \in F_s$ together yield that $\varphi(1) = 1$. Since φ is linear, $\varphi(\sigma) = \sigma$ for all $\sigma \in F_s$. It is now straightforward to check that $\theta(s) = \varphi(\lambda^{-1}s - \frac{1}{2}\lambda^{-1}\mu(s))$ for all $s \in S$ and so φ is a Jordan isomorphism of S onto S'. Now it follows from [**17**] (see also [**21**, **22**]) that φ can be extended to a surjective $*$-linear homomorphism (which we also denote by φ) of associative F_s-algebras $\langle S \rangle$ and $\langle S' \rangle$. If $I = \ker(\varphi)$, then $I^* = I$ and $I \cap S = 0$. Therefore $x^* + x \in I \cap S = 0$ for all $x \in I$ and so $x^2 \in I \cap S = 0$ for all $x \in I$. On the other hand, the ring $\langle S \rangle$ is prime [**18**, Theorem 3.4], and so $I = 0$. That is, φ is an isomorphism.

Let us finally mention that the bound $\deg(A') > 8$ in the theorem is not the best possible. For example, one can lower it to $\deg(A') > 6$ arguing similarly as in the proof [**10**, Theorem 2]. However, this makes the proof considerably longer.

4. Normal-preservers: The case of involution of the second kind.

Theorem 4.1. *Let A be A' be centrally closed prime algebras over a field F with involution of the second kind. Suppose that* $\deg(A) > 2$, $\deg(A') > 2$, *and that* $\mathrm{char}(F) \neq 2, 3$. *Let $\theta : A \to A'$ be a bijective F-linear map with the property that $\theta(x)$ is normal whenever $x \in A$ is normal. Then θ is of the form $\theta(x) = \alpha\phi(x) + \beta(x)$ where $\alpha \in F$, $\alpha \neq 0$, $\beta : A \to F$ is a linear map and ϕ is either a $*$-isomorphism or a $*$-antiisomorphism of A onto A'.*

Proof. Let $\epsilon \in F$ be such that $\epsilon^* = -\epsilon$. Then $A = S + \epsilon S$.

First we show that $\theta(1) \in F$, that is, that $\theta(1)$ is a central element in A'. Since $s + \lambda$ is a normal element for every $s \in S$ and $\lambda \in F$, it follows that $\theta(s+\lambda) = \theta(s) + \lambda\theta(1)$ is normal, that is, $[\theta(s) + \lambda\theta(1), \theta(s)^* + \lambda^*\theta(1)^*] = 0$ and hence $\lambda[\theta(1), \theta(s)^*] + \lambda^*[\theta(s), \theta(1)^*] = 0$. First setting $\lambda = 1$ and then $\lambda = \epsilon$ it follows that $[\theta(1), \theta(s)^*] = 0$, which in turn implies that $\theta(1)$ is central for $A = S + \epsilon S$ and θ is bijective.

Next, $s^2 + \lambda s$ is normal for $s \in S$ and $\lambda \in F$, and so $\theta(s^2) + \lambda\theta(s)$ is normal which implies that $[\theta(s^2), \theta(s)^*] = 0$. Linearizing we get

$$[\theta(s^2), \theta(t)^*] + [\theta(s \circ t), \theta(s)^*] = 0 \quad \text{for all } s, t \in S.$$

Again using $A = S + \epsilon S$ it follows easily that $[\theta(x^2), \theta(x^*)^*] = 0$ for all $x \in A$. Replacing x by $x+1$ and using the fact that $\theta(1)$ is central it follows that $[\theta(x), \theta(x^*)^*] = 0$ for all $x \in A$. Now, using Theorem 2.1 we see that there is $\tau \in F$ and a map $\zeta : A \to F$ such that $\theta(x^*)^* = \tau\theta(x) + \zeta(x)$, $x \in A$. Of course, $\tau \neq 0$ for otherwise A' would be commutative, contrary to the assumption. Consequently, $[\theta(x^2), \theta(x)] = 0$ for all $x \in A$. Thus, all the requirements of [**10**, Theorem 2] are fulfilled, and so it follows that θ is of the form $\theta(x) = \alpha\phi(x) + \beta(x)$ where $\alpha \in F$, $\alpha \neq 0$, β is a linear map from A into the center of A' and ϕ is either an isomorphism or an antiisomorphism of A onto A'.

All it remains to show is that $\phi(x^*) = \phi(x)^*$, $x \in A$.

Assume that ϕ is an isomorphism. Then $\psi(x) = \phi(x^*)^*$ also defines an isomorphism. We want to show that $\phi = \psi$. We have

$$\alpha^*\psi(x) + \beta(x^*)^* = \theta(x^*)^* = \tau\theta(x) + \zeta(x) = \tau\alpha\phi(x) + \tau\beta(x) + \zeta(x).$$

Since $\alpha \neq 0$ and $\tau \neq 0$ it follows that $\nu(x) = \psi(x) - \gamma\phi(x) \in F$ for every $x \in A$, where $\gamma = \frac{\alpha\tau}{\alpha^*}$ is a nonzero element in F. Whence

$$\nu(xy) = \psi(x)\psi(y) - \gamma\phi(x)\phi(y) = \nu(x)\psi(y) + \gamma\phi(x)(\psi(y) - \phi(y)).$$

Commuting this expression with $\psi(y)$ it follows, since $\gamma \neq 0$ and $\psi(y) - \phi(y)$ commutes with $\psi(y)$, that $[\phi(x), \psi(y)](\psi(y) - \phi(y)) = 0$ for all $x, y \in A$. Replacing x by xz we get at once that $[A', \psi(y)]A'(\psi(y) - \phi(y)) = 0$ for every $y \in A$. Since A' is prime this shows that given $y \in A$, either $\psi(y)$ is central or $\psi(y) = \phi(y)$. Since a group cannot be the union of two proper subgroups and since A' is noncommutative, it follows that $\psi(y) = \phi(y)$ for all $y \in A$. Similarly we discuss the case when ϕ is an antiisomorphism.

5. Normal-preservers: The case of involution of the first kind.

Theorem 5.1. *Let A be A' be centrally closed prime algebras over a field F with involution of the first kind. Suppose that* $\deg(A) > 6$, $\deg(A') > 13$ *and* $\operatorname{char}(F) \neq 2, 3$. *Further, let $\theta : A \to A'$ be a bijective $*$-linear map with the property that $\theta(x)$ is normal whenever $x \in A$ is normal. Then there exist*

$\mu_1, \mu_2 \in F$, $\mu_1 \neq \pm\mu_2$, a linear map $\omega : \langle K\rangle \to F$, and a $*$-isomorphism ψ of $\langle K\rangle$ onto $\langle K'\rangle$ such that $\theta(x) = \psi(\mu_1 x + \mu_2 x^*) + \omega(x + x^*)$ for all $x \in \langle K\rangle$.

First note that, since θ is $*$-linear, the condition that θ maps normal elements into normal elements is equivalent to the condition that $\theta(s)$ and $\theta(k)$ commute whenever $s \in S$ and $k \in K$ commute. In particular, for any $k \in K$, k^2 is a symmetric element commuting with skew elements k and k^3, so that $[\theta(k^2), \theta(k)] = 0$ and $[\theta(k^2), \theta(k^3)] = 0$. One can note from the proof that this is essentially all that we need; more precisely, in the theorem we could replace the condition that θ preserves normal elements by a milder condition that θ satisfies these two identities and that $\theta(F) = F$.

The proof of Theorem 5.1 is broken up into a series of lemmas. We begin with:

Lemma 5.2. *There exists $\lambda_0 \neq 0$ in F such that the map $\phi = \lambda_0\theta$ satisfies $\phi(k^2) - \phi(k)^2 \in F$ for all $k \in K$.*

Proof. As already observed, $[\theta(k), \theta(k^2)] = 0$ for all $k \in K$. But then it follows from Theorem 2.2 that there exist $\lambda_0 \in F$ and a linear map $\mu_0 : K \to F$ such that $\theta(k^2) - \lambda_0\theta(k)^2 - \mu_0(k)\theta(k) \in F$ for every $k \in K$. However, since θ is $*$-linear and $*$ is of the first kind, μ_0 must be zero. Hence we see that $\phi = \lambda_0\theta$ indeed satisfies $\phi(k^2) - \phi(k)^2 \in F$, $k \in K$. Finally, assuming that λ_0 is zero we arrive at $\theta(k^2) \in F$; however, $\theta(F) = F$ for $\theta(1) \in F$ (namely, $\theta(1)$ commutes with $K' = \theta(K)$ and $\deg(A') > 2$ — cf. the proof of Theorem 2.2), and so k^2 lies in F for every $k \in K$. But Theorem 2.3 tells us that this is impossible. The lemma is thereby proved.

Of course, ϕ has the same properties as θ, that is, it is $*$-linear, bijective and preserves normal elements.

Define $\epsilon : K \times K \to F$ by

$$\epsilon(k, l) = \frac{1}{2}\{\phi(k \circ l) - \phi(k) \circ \phi(l)\}. \tag{2}$$

Clearly, ϵ is a bilinear symmetric map.

Lemma 5.3. *There exist $\lambda \neq 0$ in F and a symmetric bilinear map $\mu : K \times K \to F$ such that*

$$\phi(klk) = \lambda\phi(k)\phi(l)\phi(k) + \mu(k, l)\phi(k)$$

for all $k, l \in K$.

Proof. If $k \in K$, then $k^2 \in S$ and $k^3 \in K$, so that $[\phi(k^3), \phi(k^2)] = 0$. However, $\phi(k^2) = \phi(k)^2 + \epsilon(k, k)$ and so $[\phi(k^3), \phi(k)^2] = 0$ for every $k \in K$. Since $\deg(A') > 10$, Theorem 2.2, together with the fact that ϕ is $*$-linear and $*$ is of the first kind, shows that there are $\lambda \in F$ and a symmetric bilinear map $\mu : K \times K \to F$ such that

$$\phi(k^3) = \lambda\phi(k)^3 + \mu(k, k)\phi(k) \quad \text{for all } k \in K. \tag{3}$$

Note that $\lambda = 0$ yields $k^3 = \mu(k,k)k$ which is, since $\deg(A) > 6$, impossible by Theorem 2.3. Thus, $\lambda \neq 0$.

Linearizing (3) we get

$$\begin{aligned}(4)\qquad \phi(k^2l + klk + lk^2) &= \lambda\{\phi(k)^2\phi(l) + \phi(k)\phi(l)\phi(k) + \phi(l)\phi(k)^2\} \\ &\quad + \mu(k,k)\phi(l) + 2\mu(k,l)\phi(k).\end{aligned}$$

Next we compute $\phi(k^2lk + klk^2)$ in two different ways. First using (2) we get

$$\begin{aligned}&2\phi(k^2lk + klk^2) \\ &= \phi(k \circ \{k^2 \circ l + klk\}) - \phi(k^3 \circ l) \\ &= \phi(k) \circ \phi(k^2 \circ l + klk) + 2\epsilon(k, k^2 \circ l + klk) - \phi(k^3) \circ \phi(l) - 2\epsilon(k^3, l),\end{aligned}$$

so that

$$\begin{aligned}&2\phi(k^2lk + klk^2) - \phi(k)\phi(k^2l + klk + lk^2) \\ &\qquad - \phi(k^2l + klk + lk^2)\phi(k) + \phi(k^3)\phi(l) + \phi(l)\phi(k^3) \in F.\end{aligned}$$

Applying (3) and (4) it follows that

$$\begin{aligned}&2\phi(k^2lk + klk^2) \\ &\quad - \phi(k)\{\lambda\phi(k)^2\phi(l) + \lambda\phi(k)\phi(l)\phi(k) + \lambda\phi(l)\phi(k)^2 \\ &\qquad + \mu(k,k)\phi(l) + 2\mu(k,l)\phi(k)\} \\ &\quad - \{\lambda\phi(k)^2\phi(l) + \lambda\phi(k)\phi(l)\phi(k) + \lambda\phi(l)\phi(k)^2 \\ &\qquad + \mu(k,k)\phi(l) + 2\mu(k,l)\phi(k)\}\phi(k) \\ &\quad + \{\lambda\phi(k)^3 + \mu(k,k)\phi(k)\}\phi(l) + \phi(l)\{\lambda\phi(k)^3 + \mu(k,k)\phi(k)\} \in F,\end{aligned}$$

and hence

$$\phi(k^2lk + klk^2) - \lambda(\phi(k)^2\phi(l)\phi(k) + \phi(k)\phi(l)\phi(k)^2) - 2\mu(k,l)\phi(k)^2 \in F.$$

On the other hand, (2) implies that

$$\phi(k^2lk + klk^2) = \phi(k \circ \{klk\}) = \phi(k) \circ \phi(klk) + 2\epsilon(k, klk).$$

Comparing we obtain

$$\phi(k) \circ \{\phi(klk) - \lambda\phi(k)\phi(l)\phi(k) - \mu(k,l)\phi(k)\} \in F$$

for all $k, l \in K$. According to the statement (ii) of Theorem 2.2 we must then have $\phi(klk) - \lambda\phi(k)\phi(l)\phi(k) - \mu(k,l)\phi(k) = 0$ for all $k, l \in K$, and the lemma is proved.

We shall need the conclusion of Lemma 5.3 in the following form

$$\phi(k_1lk_2 + k_2lk_1) = \lambda\{\phi(k_1)\phi(l)\phi(k_2) + \phi(k_2)\phi(l)\phi(k_1)\} + \mu(k_1, l)\phi(k_2) + \mu(k_2, l)\phi(k_1) \tag{5}$$

for all $k_1, k_2, l \in K$.

Lemma 5.4. *$\mu(k, l) = 0$ for all $k, l \in K$.*

Proof. The proof is based on computing $\phi(kl_1kl_2k + kl_2kl_1k)$, where k, l_1, l_2 are arbitrary elements in K, in two different ways. First, applying (5) we get

$$\begin{aligned}
&\phi(kl_1kl_2k + kl_2kl_1k)\\
&= \phi((kl_1k)l_2k + kl_2(kl_1k))\\
&= \lambda\{\phi(kl_1k)\phi(l_2)\phi(k) + \phi(k)\phi(l_2)\phi(kl_1k)\} + \mu(kl_1k, l_2)\phi(k)\\
&\quad + \mu(k, l_2)\phi(kl_1k)\\
&= \lambda^2\{\phi(k)\phi(l_1)\phi(k)\phi(l_2)\phi(k) + \phi(k)\phi(l_2)\phi(k)\phi(l_1)\phi(k)\}\\
&\quad + 2\lambda\mu(k, l_1)\phi(k)\phi(l_2)\phi(k) + \lambda\mu(k, l_2)\phi(k)\phi(l_1)\phi(k)\\
&\quad + \{\mu(kl_1k, l_2) + \mu(k, l_2)\mu(k, l_1)\}\phi(k).
\end{aligned}$$

However, l_1 and l_2 appear symmetrically in the expression $kl_1kl_2k + kl_2kl_1k$ and so, on the other hand, we must have

$$\begin{aligned}
&\phi(kl_1kl_2k + kl_2kl_1k)\\
&= \lambda^2\{\phi(k)\phi(l_2)\phi(k)\phi(l_1)\phi(k) + \phi(k)\phi(l_1)\phi(k)\phi(l_2)\phi(k)\}\\
&\quad + 2\lambda\mu(k, l_2)\phi(k)\phi(l_1)\phi(k) + \lambda\mu(k, l_1)\phi(k)\phi(l_2)\phi(k)\\
&\quad + \{\mu(kl_2k, l_1) + \mu(k, l_1)\mu(k, l_2)\}\phi(k).
\end{aligned}$$

Comparing both relations we obtain

$$\lambda\phi(k)\{\mu(k, l_2)\phi(l_1) - \mu(k, l_1)\phi(l_2)\}\phi(k) = \{\mu(kl_1k, l_2) - \mu(kl_2k, l_1)\}\phi(k)$$

for all $k, l_1, l_2 \in K$. Now using Theorem 2.3 twice it follows that

$$\mu(k, l_2)\phi(l_1) - \mu(k, l_1)\phi(l_2) = 0 \quad \text{for all } k, l_1, l_2 \in K,$$

which readily implies the assertion of the lemma.

Thus, (5) now reduces to

$$\phi(k_1lk_2 + l_2kl_1) = \lambda\{\phi(k_1)\phi(l)\phi(k_2) + \phi(k_2)\phi(l)\phi(k_1)\}. \tag{6}$$

Lemma 5.5. *There exists $\rho \in F$ such that $\rho^2 = \lambda$ and $\phi([k,l]) = \rho[\phi(k), \phi(l)]$ for all $k, l \in K$.*

Proof. We have arrived at the situation when Theorem 2.4 can be applied. Taking a polynomial f to be $f(x_1, x_2, x_3) = x_1x_2x_3 + x_3x_2x_1$ and a map B to be equal to $B(k,l) = \phi([k,l])$, we see, by making use of (6), that all the conditions of this theorem are fulfilled (this is the place when the condition $\deg(A') > 13$ is used). It follows that there exists $\rho \in F$ such that $\phi([k,l]) - \rho[\phi(k), \phi(l)] \in F$ for all $k, l \in K$; however, since $*$ is of the first kind, this clearly yields $\phi([k,l]) = \rho[\phi(k), \phi(l)]$. It remains to show that $\rho^2 = \lambda$. We have

$$\phi([[k_1, k_2], k_3]) = \rho[\phi([k_1, k_2]), \phi(k_3)] = \rho^2[[\phi(k_1), \phi(k_2)], \phi(k_3)].$$

On the other hand, using (6), we get

$$\begin{aligned}\phi([[k_1, k_2], k_3]) &= \phi(k_1k_2k_3 + k_3k_2k_1) - \phi(k_2k_1k_3 + k_3k_1k_2) \\ &= \lambda[[\phi(k_1), \phi(k_2)], \phi(k_3)].\end{aligned}$$

Whence $(\rho^2 - \lambda)[[K', K'], K'] = 0$. Suppose that $[[K', K'], K'] = 0$. Then applying [**8**, Theorem 9.1.13] we get that $\deg(A') \leq 2$, a contradiction. Therefore $\rho^2 = \lambda$ and the lemma is proved.

Lemma 5.6. *There exist a $*$-isomorphism ψ of an algebra $\langle K \rangle$ onto an algebra $\langle K' \rangle$ and a linear map $\tau : K \circ K \to F$ such that $\psi(k) = \rho\phi(k)$ for all $k \in K$ and $\psi(s) = \lambda\phi(s) - \tau(s)$ for all $s \in K \circ K$.*

Proof. We first define ψ on K by $\psi(k) = \rho\phi(k)$, $k \in K$. Since $\phi = \lambda_0\theta$ and θ is a $*$-linear map, $\psi(K) = \theta(K) = K'$. It follows from Lemmas 5.3, 5.4 and 5.5 together that

$$\psi([k,l]) = [\psi(k), \psi(l)] \quad \text{and} \quad \psi(k^3) = \psi(k)^3 \quad \text{for all } k, l \in K. \tag{7}$$

Now both (7) and [**8**, Lemma 9.4.5] imply that ψ can be uniquely extended to an isomorphism (which we also denote by ψ) of associative rings $\langle K \rangle$ and $\langle K' \rangle$. Since $\psi|_K$ is a linear map and K generates $\langle K \rangle$, ψ is an isomorphism of algebras. Clearly

$$\psi(K) = K' \quad \text{and} \quad \psi(K \circ K) = \psi(K) \circ \psi(K) = K' \circ K'. \tag{8}$$

According to [**8**, Lemma 9.1.5], $\langle K\rangle = K + K \circ K$ and $\langle K'\rangle = K' + K' \circ K'$. Obviously K (respectively, $K \circ K$) is the set of skew (respectively, symmetric) elements of the algebra $\langle K\rangle$. It now follows from (8) that ψ is a $*$-isomorphism.

Now define a linear map τ on $K \circ K$ by $\tau(s) = \lambda\phi(s) - \psi(s)$. We claim that $\tau(s)$ lies in F for any $s \in K \circ K$. Indeed, clearly the vector space $K \circ K$ is spanned by the set $\{k^2 \mid k \in K\}$. Given $k \in K$, we have

$$\begin{aligned}\tau(k^2) &= \lambda\phi(k^2) - \psi(k^2) = \lambda\phi(k^2) - \psi(k)^2 \\ &= \lambda\phi(k^2) - \{\rho\phi(k)\}^2 = \lambda\{\phi(k^2) - \phi(k)^2\} \in F\end{aligned}$$

by Lemma 5.2 which proves our claim.

Finally, invoking the definition of ϕ we see from Lemma 5.6 that for any $x \in \langle K\rangle$ we have

$$\begin{aligned}\theta(x) = \lambda_0\phi(x) &= \frac{\lambda_0}{2}\phi(x - x^*) + \frac{\lambda_0}{2}\phi(x + x^*) \\ &= \frac{\lambda_0\rho^{-1}}{2}\psi(x - x^*) + \frac{\lambda_0\lambda^{-1}}{2}\psi(x + x^*) + \frac{\lambda_0\lambda^{-1}}{2}\tau(x + x^*).\end{aligned}$$

Now set $\mu_1 = \frac{1}{2}\lambda_0(\lambda^{-1} + \rho^{-1})$, $\mu_2 = \frac{1}{2}\lambda_0(\lambda^{-1} - \rho^{-1})$, $\omega(x - x^*) = 0$, $\omega(x + x^*) = \frac{1}{2}\lambda_0\lambda^{-1}\tau(x + x^*)$ and note that the desired conclusion holds true.

Acknowledgement. The paper has been written while the second and the third author were visiting the National Cheng-Kung University. They would like to express their deep gratitude to the University for its hospitality and for the financial support of their visit.

The authors would also like to thank the referee for careful reading of the paper and for some useful suggestions.

References

[1] *A survey of linear preserver problems*, Linear and Multilinear Algebra, **33**(**1**–**2**) (1992), Gordon and Breach, Yverdon, 1992, 1-129, MR 96c:15043.

[2] R. Banning and M. Mathieu, *Commutativity preserving mappings on semiprime rings*, Comm. Algebra, **25** (1997), 247-265, MR 97j:16033, Zbl 0865.16015.

[3] K.I. Beidar, *On functional identities and commuting additive mappings*, Comm. Algebra, **26** (1998), 1819-1850, MR 99f:16023, Zbl 0901.16011.

[4] K.I. Beidar, S.-C. Chang, M.A. Chebotar and Y. Fong, *On functional identities in left ideals of prime rings*, Comm. Algebra, **28** (2000), 3041-3058, MR 2001c:16045, Zbl 0971.16014.

[5] K.I. Beidar and M.A. Chebotar, *On functional identities and d-free subsets of rings* I, Comm. Algebra, **28** (2000), 3925-3952, MR 2001j:16046.

[6] ———, *On functional identities and d-free subsets of rings* II, Comm. Algebra, **28** (2000), 3953-3972, MR 2001j:16046.

[7] K.I. Beidar and W.S. Martindale 3rd, *On functional identities in prime rings with involution*, J. Algebra, **203** (1998), 491-532, MR 99f:16024, Zbl 0904.16012.

[8] K.I. Beidar, W.S. Martindale 3rd and A.V. Mikhalev, *Rings with Generalized Identities*, Marcel Dekker, Inc., New York-Basel-Hong Kong, 1996, MR 97g:16035, Zbl 0847.16001.

[9] M. Brešar, *Centralizing mappings and derivations in prime rings*, J. Algebra, **156** (1993), 385-394, MR 94f:16042, Zbl 0773.16017.

[10] ———, *Commuting traces of biadditive mappings, commutativity-preserving mappings and Lie mappings*, Trans. Amer. Math. Soc., **335** (1993), 525-546, MR 93d:16044, Zbl 0791.16028.

[11] ———, *Functional identities: A survey*, Contemporary Math., **259** (2000), 93-109, MR 2001h:16023, Zbl 0967.16011.

[12] M. Brešar and P. Šemrl, *Normal-preserving linear mappings*, Can. Math. Bull., **37** (1994), 306-309, MR 96b:47040, Zbl 0816.47018.

[13] ———, *Linear preservers on $\mathcal{B}(X)$*, Banach Center Publ., **38** (1997), 49-58, MR 99c:47044, Zbl 0939.47031.

[14] G.H. Chan and M.H. Lim, *Linear transformations on symmetric matrices that preserve commutativity*, Linear Algebra Appl., **47** (1982), 11-22, MR 83k:15020, Zbl 0492.15006.

[15] M.D. Choi, A.A. Jafarian and H. Radjavi, *Linear maps preserving commutativity*, Linear Algebra Appl., **87** (1987), 227-242, MR 88f:15003, Zbl 0615.15004.

[16] C.M. Kunicki and R.D. Hill, *Normal-preserving linear transformations*, Linear Algebra Appl., **170** (1992), 107-115, MR 93j:15004, Zbl 0751.15002.

[17] L.A. Lagutina, *Jordan homomorphisms of associative algebras with involution*, Algebra and Logic, **27** (1988), 250-260, MR 90k:16013, Zbl 0697.16013.

[18] C. Lanski, *On the relationship of a ring and the subring generated by its symmetric elements*, Pacific J. Math., **44** (1973), 581-592, MR 48 #331, Zbl 0252.16004.

[19] C.-K. Li and N.-K. Tsing, *Linear preserver problems: A brief introduction and some special techniques*, Linear Algebra Appl., **162-164** (1992), 217-235, MR 93b:15003, Zbl 0762.15016.

[20] M. Marcus, *Linear operations on matrices*, Amer. Math. Monthly, **69** (1962), 837-847, MR 26 #5007, Zbl 0108.01104.

[21] W.S. Martindale 3rd, *Jordan homomorphisms onto nondegenerate Jordan algebras*, J. Algebra, **133** (1990), 500-511, MR 91m:17045, Zbl 0708.17027.

[22] K. McCrimmon, *The Zelmanov approach to Jordan homomorphisms of associative algebras*, J. Algebra, **123** (1989), 457-477, MR 90j:17053, Zbl 0675.17016.

[23] C. Procesi, *Rings with Polynomial Identities*, Marcel Decker, Inc., 1973, MR 51 #3214, Zbl 0262.16018.

[24] H. Radjavi, *Commutativity-preserving operators on symmetric matrices*, Linear Algebra Appl., **61** (1984), 219-224, MR 85h:15007, Zbl 0547.15007.

[25] H. Radjavi and P. Rosenthal, *Invariant Subspaces*, Ergeb. Math. Grenzeb., **77**, Springer-Verlag, New York, 1973, MR 51 #3924, Zbl 0269.47003.

[26] L.H. Rowen, *Polynomial Identities in Ring Theory*, Academic Press, 1980, MR 82a:16021, Zbl 0461.16001.

Received September 25, 2000 and revised January 10, 2001. The second author was partially supported by a grant from the Ministry of Science of Slovenia.

Department of Mathematics
National Cheng-Kung University
Tainan
Taiwan
E-mail address: beidar@mail.ncku.edu.tw

Department of Mathematics, PF
University of Maribor
Maribor
Slovenia
E-mail address: bresar@uni-mb.si

Department of Mechanics and Mathematics
Tula State University
Tula
Russia
E-mail address: mchebotar@tula.net

Department of Mathematics
National Cheng-Kung University
Tainan
Taiwan
E-mail address: fong@mail.ncku.edu.tw

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

FLOW EQUIVALENCE OF SHIFTS OF FINITE TYPE VIA POSITIVE FACTORIZATIONS

Mike Boyle

Together with M. Boyle and D. Huang (2000), this paper gives an alternate development of the Huang classification of shifts of finite type up to flow equivalence, and provides additional functorial information, used to analyze the action of the mapping class group of the mapping torus of a shift of finite type on the "isotopy futures" group, which is introduced here. For a shift of finite type σ_A, this group is isomorphic to the Bowen-Franks group $\mathrm{cok}(I-A)$. The action on the isotopy futures group of a subshift is the flow equivalence analogue of the dimension group representation.

1. Introduction.

Shifts of finite type (SFTs) are the fundamental building blocks of symbolic dynamics, with applications to hyperbolic dynamics, ergodic theory, topological dynamics, matrix theory and other areas [**Bow**, **DGS**, **Ki**, **LM**, **Rob**, **S**]. Any SFT is conjugate to an SFT σ_A defined by a matrix A with nonnegative integer entries. A fundamental question about SFTs, when are they flow equivalent, is important also for the study of certain C^*-algebras [**C**, **CK**, **H2**, **H3**, **R**]. This question was solved in the irreducible case by Franks [**F**], extending earlier work of Parry and Sullivan [**PS**] and Bowen and Franks [**BowF**], and then in the general case by Huang [**H4**, **H5**], following earlier work on more tractable special cases [**H1**, **H3**]. Huang [**H4**, **H5**] developed complete algebraic invariants (defined in terms of the given matrix A) for flow equivalence of SFTs.

This paper has three main features.

(1) Taken together with [**BH**], the paper gives a self-contained alternate development of the Huang classification of SFTs up to flow equivalence. This development separates algebraic and positivity issues, and provides additional functorial information.

(2) We introduce the isotopy futures group $\mathcal{F}_S$ of the mapping torus Y_S of a subshift S, and when S is an SFT σ_A we construct an isomorphism of $\mathcal{F}_S$ and the Bowen-Franks group $\mathrm{cok}(I-A)$, and analyze the induced action of the mapping class group of Y_S on $\mathcal{F}_S$.

(3) We integrate the study of flow equivalence of SFTs into the "positive K-theory" framework for classification problems in symbolic dynamics.

We now discuss these features in more detail.

1. To study reducible SFTs, we work with certain infinite block triangular integral matrices with block rows and columns indexed by a finite poset $\mathcal{P}$: If $i \not\leq j$ in $\mathcal{P}$, then the ij block of the matrix must be zero. The elements of $\mathcal{P}$, and their ordering, correspond to the irreducible components of the SFT, and their asymptotic transitions; the isomorphism class of the poset $\mathcal{P}$ is an invariant of flow equivalence. We say two such matrices B, B' are $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalent if there are matrices U, V satisfying the same zero-subblock conditions, and with all diagonal blocks having determinant 1, such that $UBV = B'$. After fixing a choice of $\mathcal{P}$, and allowing a permutation of $\mathcal{P}$, we show that A, A' define flow equivalent SFTs if and only if the matrices $I - A$ and $I - A'$ are $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalent by an equivalence which is "positive on cycle components"(a technical condition which may be removed after reduction to a standard form, see Theorem 3.4). The key to this result is the Factorization Theorem 3.3, which gives necessary and sufficient conditions for an $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence to be a composition of "positive" elementary equivalences (which induce flow equivalences). Complete algebraic invariants for $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence are contained in the joint work [**BH**] with Danrun Huang. (The proofs in the current paper are very different from those of Huang [**H4**, **H5**], but the algebraic sequel [**BH**] depends completely on the ideas introduced by Huang in [**H4**, **H5**].)

In Huang's development, the proofs involve creating positive matrix models realizing given isomorphisms of an associated "K-web" of exact sequences of associated groups; the difficult positivity and algebraic issues are intertwined. By interposing the $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence relation between the SFTs and the complicated K-web algebraic invariants, we separate the positivity issues (which we address in this paper) from purely algebraic issues (which are addressed in [**BH**]). This clarifies the meaning of the invariants and the structure of the problem. It also facilitates the application of algebraic results.

2. The analysis of the induced action on $\mathrm{cok}(I-A)$ uses the Factorization Theorem 3.3 together with purely algebraic results from [**BH**] on $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence. There is a plausible program (7.15) for extending these ideas to obtain more information.

It seems to be nontrivial to construct a functor which attaches isomorphisms of Bowen-Franks groups to isotopy classes of flow equivalences of SFTs. (For example, we do not know if it is possible to construct a homomorphism from $\check{H}^1(Y_{\sigma_A})$ onto the Bowen-Franks group $\mathrm{cok}(I - A)$ such that the natural action of the mapping class group on $\check{H}^1(Y_{\sigma_A})$ induces an

automorphism of $\operatorname{cok}(I - A)$; and we suspect there is no such homomorphism.) An alternate approach using work of Badoian is discussed at the end of Section 7. Another possible approach would be to extend ideas of Bowen and Franks, who computed $\operatorname{cok}(I - A)$ as a relative homology group in the context of basic sets of Smale flows ([**BowF**, **F**]).

3. In the framework of positive K-theory (a term introduced by Wagoner), some class of matrices A presents some category of dynamical system, and multiplication of $I - A$ by elementary matrices satisfying some positivity condition induces isomorphisms of the system presented by A. This framework was born in [**KRW2**, **KRW3**], where matrices over $t\mathbb{Z}_+[t]$ presented SFTs, and multiplication of $I - A$ by certain elementary matrices over $\mathbb{Z}[t]$ gave a completely new method of constructing topological conjugacies, which allowed the solution of a difficult and important open problem. This framework for SFT's is developed or exploited further in [**BW**, **B1**, **KR1**, **W2**]; in the last reference [**W2**], the K-theory connection is more than a formal analogy and gives new counterexamples to Williams' shift equivalence conjecture. In [**G**], the matrix entries lie in a certain ring of formal power series, and the elementary matrix multiplications induce good finitary isomorphisms of Markov chains. In this paper and in [**Ba1**], the matrices have integer (or zero-one) entries, and the elementary multiplications induce flow equivalences. There is a passage from the topological conjugacy case to the flow equivalence case by "setting t equal to 1" (applying the coinvariants functor), as described in [**B1**]. The positive K-theory approach gives a unified and useful framework for classification problems in symbolic dynamics, and we view this paper as a significant piece of the theory for the case of flow equivalence of SFTs. It is possible that the methods of this paper may be suggestive for the case of topological conjugacy of SFTs.

Some of our results on flow equivalence have alternate proofs based on the work of Leslie Badoian [**Ba1**], who develops for irreducible SFTs a flow equivalence theory analogous to the theory created by Wagoner for topological conjugacies of SFTs. At the end of Section 7, we summarize the main results of [**Ba1**], and discuss those alternate proofs.

Now some words on the structure of the paper. In Section 2 we give some definitions and technical results necessary for the statement of the main results in Section 3. The proof of the Factorization Theorem is carried out in Sections 4-5 and the Appendix. Shifts of finite type and the relation of flow equivalence to the matrix results are addressed in Section 6. The isotopy futures group and connections to flow equivalence are studied in Section 7. For the simple general statement of our Factorization Theorem for matrices, we need preliminary technical arguments to reduce our matrices to a nondegenerate form. These preliminaries are complicated, and we banish them to the Appendix.

The basic approach of this paper, and the Factorization Theorem in the "no cycle components" case under additional technical assumptions since removed, were announced in [**B1**].

I thank Danrun Huang for many helpful comments, and for a very satisfying collaboration in our sequel paper [**BH**]. Also, without his earlier work, this paper would not exist.

2. Definitions.

2.1. Poset blocked matrices. For the rest of the paper, we let $\mathcal{P} = \{1, \dots, N\}$ denote a finite poset (partially ordered set). We describe the order with a relation $\prec$ satisfying the following conditions (in which $<$ refers to the usual order on $\mathbb{N}$) for all i, j, k in $\mathcal{P}$:

$$i \prec j \implies i < j,$$
$$i \prec j \prec k \implies i \prec k.$$

We write $i \preceq j$ to mean that $i \prec j$ or $i = j$. We can visualize the poset as an acyclic directed graph with vertex set $\{1, \dots, N\}$ and transitions $i \to j$ iff $i \prec j$.

We say that a matrix (or a block in a matrix) is *square* if its rows and columns are indexed by the same set (which may be finite or countably infinite). Suppose that $n_1, \dots, n_N$ lie in the set $\{1, 2, \dots, \infty\}$. Let $\mathbf{n} = (n_1, \dots, n_N)$. We say a square matrix M is "$\mathbf{n}$-blocked" if it splits into blocks $M_{ij}, 1 \le i, j \le N$, where M_{ij} denotes the intersection of the ith block row and the jth block column, and has size $n_i \times n_j$. (We will also use the notation $M\{i, j\}$ to denote M_{ij}.) Given an $\mathbf{n}$-blocked matrix M, we let $\mathcal{I}_j$ denote the set of indices for rows/columns through the block M_{jj}.

Definition 2.1. $\mathfrak{M}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ is the set of $\mathbf{n}$-blocked matrices with entries in $\mathbb{Z}$ satisfying the following conditions:

(1) For $1 \le i \le N$, the block M_{ii} equals the identity matrix in all but finitely many entries.
(2) For $1 \le i, j \le N$ and $i \ne j$, the block M_{ij} is zero in all but finitely many entries.
(3) If $i \not\preceq j$, then the block M_{ij} is zero.

The matrices in the semiring $\mathfrak{M}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ are block upper triangular and in addition certain blocks above the diagonal must be zero. $\mathfrak{M}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ is closed under addition and (because $\prec$ is transitive) under matrix multiplication.

A nonnegative matrix A is *irreducible* if it is square with all entries nonnegative, and for every (i, j) there exists $n > 0$ such that $A^n(i, j) > 0$. (In particular, for us a zero matrix is not irreducible.) A square matrix is *essentially irreducible* if it has a unique principal submatrix which is irreducible and which is contained in no larger irreducible principal submatrix.

Definition 2.2. $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n},\mathbb{Z})$ is the set of $\mathbf{n}$-blocked nonnegative integral matrices with only finitely many nonzero entries, satisfying the following conditions:

(1) Each diagonal block M_{ii} is essentially irreducible.
(2) If $i \not\preceq j$, then the block M_{ij} is zero.
(3) If $i \prec j$, then there is an index i' occuring on a cycle of M_{ii} and an index j' occuring on a cycle of M_{jj} and a positive integer n, such that $A^n(i',j') > 0$.

Definition 2.3. $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ is the set of matrices M in $\mathfrak{M}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ such that $\det(M_{ii}) = 1$ for $1 \le i \le N$.

Abbreviations 2.4. $\mathfrak{M}_{\mathcal{P}}(\mathbb{Z})$, $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$ and $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ denote the sets $\mathfrak{M}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$, $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n},\mathbb{Z})$ and $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ for which $\mathbf{n} = (n_1,\dots,n_N)$ with every $n_i = \infty$. Whenever any such matrix family appears with no subscript $\mathcal{P}$, it means that $\mathcal{P} = \{1\}$ (the block structure is trivial).

We say two matrices B, B' in $\mathfrak{M}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ are $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$-*equivalent in* $\mathfrak{M}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ if there are matrices U, V in $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ such that $UBV = B'$, and we write this as $(U,V)\colon B \to B'$. We say a matrix is a basic elementary matrix if it equals the identity matrix except in at most one offdiagonal entry. It is not difficult to check that $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ is a group under multiplication which is generated by basic elementary matrices [**BH**]. Given $\mathbf{n} \le \mathbf{r}$, we have natural truncation and embedding maps between $\mathbf{n}$-blocked and $\mathbf{r}$-blocked matrices,

$$\mathrm{tru}_{\mathbf{n}}\colon \mathfrak{M}_{\mathcal{P}}(\mathbf{r},\mathbb{Z}) \to \mathfrak{M}_{\mathcal{P}}(\mathbf{n},\mathbb{Z}),$$
$$\iota_{\mathbf{r}}\colon \mathfrak{M}_{\mathcal{P}}(\mathbf{n},\mathbb{Z}) \to \mathfrak{M}_{\mathcal{P}}(\mathbf{r},\mathbb{Z}).$$

The truncation map replaces an ij block with its $n_i \times n_j$ upper left corner. The embedding map embeds an ij block as the upper left corner of an ij block. If $i \ne j$, then the image ij block is zero outside this embedded left corner; if $i = j$, it is the identity outside this left corner. We will use A_∞ to abbreviate $\iota_{\mathbf{n}}(A)$ in the case that every $n_i = \infty$. We will also use $\mathrm{tru}_{\mathbf{n}}$, $\iota_{\mathbf{r}}$ and A_∞ for matrix families other than $\mathfrak{M}_{\mathcal{P}}$. The only potentially ambiguous point, which should be clear from context, is whether the embedded block corners should be extended as above with $\mathfrak{M}_{\mathcal{P}}$ to match the identity matrix, or should be extended to match the zero matrix (e.g., when the range is $\mathfrak{M}^{\circ}_{\mathcal{P},+}$).

2.2. Positive equivalence. Suppose for some (i,j) that E is a basic elementary matrix with offdiagonal entry $E(i,j) = 1$, $A \in \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n},\mathbb{Z})$, and $A(i,j) > 0$. Then we say that each of the equivalences

$$(E,I)\colon (I-A) \to E(I-A), \qquad (E^{-1},I)\colon E(I-A) \to (I-A),$$
$$(I,E)\colon (I-A) \to (I-A)E, \qquad (I,E^{-1})\colon (I-A)E \to (I-A)$$

is a *basic positive equivalence* in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$. Note, $E \in \mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$. We say that an $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalence is a *positive* equivalence in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$ if it is a composition of basic positive equivalences in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$.

To understand the meaning of a basic positive equivalence, suppose $A \in \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$ with $A(i,j) > 0$ and E is basic elementary matrix with offdiagonal entry $E(i,j) = 1$. We first discuss the case $(E, I)\colon (I - A) \to E(I - A)$. Define A' by the requirement $E(I - A) = (I - A')$. Then A and A' agree except perhaps in row i, where

$$\begin{aligned} A'(i,k) &= A(i,k) + A(j,k) \qquad \text{if } j \neq k, \qquad \text{and} \\ A'(i,j) &= A(i,j) + A(j,j) - 1. \end{aligned}$$

View A as the adjacency matrix of a directed graph $\mathcal{G}_A$ with edge set $\mathcal{E}_A$ and vertex set given by the $n_1 + \cdots + n_N$ indices for the rows/columns of A. (There can be edges joining only finitely many of those vertices.) We can describe a directed graph $\mathcal{G}_{A'}$ which has A' as its adjacency matrix as follows. $\mathcal{G}_{A'}$ has the same vertex set as $\mathcal{G}_A$. Now pick an edge e which runs from vertex i to vertex j in $\mathcal{G}_A$ (e exists because by assumption $A(i,j) > 0$). The edge set $\mathcal{E}_{A'}$ will be derived from $\mathcal{E}_A$ as follows: Remove e from $\mathcal{E}_A$; and then for every vertex k, for every edge f in $\mathcal{E}_A$ from j to k add in a new edge (named $[ef]$) from i to k. It is easy to verify that with this edge set $\mathcal{E}_{A'}$, the directed graph $\mathcal{G}_{A'}$ has adjacency matrix A'.

With this notation, now define a map $\gamma\colon \mathcal{E}_{A'} \to (\mathcal{E}_A)^*$ by $\gamma\colon f \mapsto f$ and $\gamma\colon [ef] \mapsto ef$. Then γ induces an injective map (also called γ), from the set $\Sigma_{A'}$ of biinfinite paths through $\mathcal{G}_{A'}$ to the set Σ_A of biinfinite paths through $\mathcal{G}_A$, sending x' to x by the rule

$$\gamma\colon \quad \ldots x'_{-2}x'_{-1}|x'_0x'_1\ldots \;\mapsto\; \ldots\gamma(x'_{-2})\gamma(x'_{-1})|\gamma(x'_0)\gamma(x'_1)\ldots$$

(in which the placement of the vertical bar indicates the indexing for x, e.g., $x_0x_1\cdots = \gamma(x'_0)\gamma(x'_1)\ldots$). Briefly: We get x from x' by replacing each A' edge $[ef]$ with ef.

The injective map $\gamma\colon \Sigma_{A'} \to \Sigma_A$ is not surjective precisely because the image will not contain points x for which $x_{-1} = e$ (the image will contain the shifted point $\sigma^{-1}x$ which is defined by $(\sigma^{-1}x)_i = x_{i-1}$). However, although γ is generally not a bijection, it does induce a bijection of orbits (under the shift) in Σ_A and $\Sigma_{A'}$. Also, γ induces a bijection of finite orbits: That is, γ induces a bijection (also called γ) of cycles in $\mathcal{G}_A$ and $\mathcal{G}_{A'}$ (which need not respect the cycle length). If $1 \leq t \leq N$ and c is a cycle for the block A_{tt}, then $\gamma(c)$ is a cycle for A'_{tt}, because if i and j are not indices for the same component then γ is the identity on cycles. Also, if x in Σ_A is backwardly asymptotic (under the shift) to a cycle c and forwardly asymptotic to a cycle $\widetilde{c}$, then $\gamma(x)$ is backwardly asymptotic to $\gamma(c)$ and forwardly asymptotic to $\gamma(\widetilde{c})$. It follows that the matrix A' satisfies the conditions of Definition 2.2 and lies in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$.

The discussion of the case $(I, E)\colon (I - A) \to (I - A)E = (I - A')$ is much the same. Let f be an A-edge from i to j. To form the A' graph from the A graph in this case, delete f, and add a new edge $[ef]$ for each edge e with terminal vertex i. Then define γ as before.

The following lemma, fundamental to the sequel, is implicit in Franks' paper [**F**].

Lemma 2.5. *Suppose $A \in \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, E is a basic elementary matrix in $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ whose nonzero offdiagonal entry is $E(i, j) = 1$, and there is a positive integer k such that $A^k(i, j) > 0$.*

(1) *If $(E(I - A))(i, j) \leq 0$, then $(E, I)\colon (I - A) \to E(I - A)$ is a positive equivalence in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$.*
(2) *If $((I - A)E)(i, j) \leq 0$, then $(I, E)\colon (I - A) \to (I - A)E$ is a positive equivalence in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$.*

Proof. We will consider the claim for the first equivalence (E, I); the other case is similar. By assumption, there is a list $i = i_0, i_1, \ldots, i_k = j$ (which we take to be of minimal length, so the indices $i_0, i_1, \ldots, i_k$ are distinct) such that for $0 \leq t < k$ we have $A(i_t, i_{t+1}) > 0$. If $k = 1$, then the equivalence is a basic positive equivalence (and we know a basic positive equivalence takes a matrix in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$ to a matrix in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$). So suppose $k > 1$. Let F_t be the elementary matrix whose which acts to add row i_t to row i. Let $F = F_{k-1} \ldots F_1$. Then we have basic positive equivalences

$$\begin{aligned}(I - A) \;&\to\; F_1(I - A) \;\to\; F_2F_1(I - A) \;\to\; \cdots \;\to\; (F_{k-1} \cdots F_2F_1)(I - A)\\ &=\; F(I - A) \;\to\; EF(I - A) \;\to\; (F_{k-1})^{-1}EF(I - A)\\ &\to\; \cdots \;\to\; (F_1)^{-1} \cdots (F_{k-2})^{-1}(F_{k-1})^{-1}EF(I - A)\\ &=\; F^{-1}EF(I - A) = E(I - A).\end{aligned}$$

□

2.3. Cycle components. The technical discussion of this subsection is only required for the case when the matrix A in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$ has a diagonal block whose maximal irreducible submatrix is a permutation matrix.

Lemma 2.6. *Suppose A is an $\mathcal{S} \times \mathcal{S}$ nonnegative integral matrix which has as its unique irreducible submatrix a cyclic permutation matrix. Then the cokernel group $\mathrm{cok}(I - A) = \mathbb{Z}^{\mathcal{S}}/(I - A)\mathbb{Z}^{\mathcal{S}}$ is isomorphic to $\mathbb{Z}$. Let $\mathcal{I}$ denote the set of indices involved in the cyclic permutation. Then the canonical basis vectors satisfy the following conditions:*

(1) *$[e_i]$ is a generator of $\mathrm{cok}(I - A)$ if $i \in \mathcal{I}$.*
(2) *$[e_i] = [e_j]$ if i and j are in $\mathcal{I}$.*
(3) *$[e_i] = 0$ if $i \notin \mathcal{I}$.*

Proof. (3) If $i \notin \mathcal{I}$, then for large n, $A^n e_i = 0$, and $e_i = (I - A^n)e_i = (I - A)(I + A + \cdots + A^{n-1})e_i$. Then $[e_i] = 0$ in $\mathrm{cok}(I - A)$.

(2) Let π denote the given permutation and suppose i and j are in $\mathcal{I}$. Then there exists $n > 0$ such that $A^n e_i = e_j$, so in $\mathrm{cok}(I - A)$ we have $[e_i] - [e_j] = [e_i - e_j] = [(I - A^n)e_i] = 0$.

(1) Clearly now, if $i \in \mathcal{I}$, then $[e_i]$ generates $\mathrm{cok}(I - A)$. Also, because $\det(I-A) = 0$, $\mathbb{Z}$ is a subgroup of $\mathrm{cok}(I-A)$. Therefore $\mathrm{cok}(I-A) \cong \mathbb{Z}$. $\square$

For a matrix A satisfying the hypotheses of the lemma, we make $\mathrm{cok}(I-A)$ an ordered group by declaring its positive set to be the collection of those $[w]$ such that (in the notation of the statement of the lemma) $\sum_{i\in\mathcal{I}} w_i \geq 0$. (This sum does not depend on the representative w of $[w]$.) We say an isomorphism between two such cokernel groups is *positive* if it takes the positive set in the domain to the positive set in the range.

If $A \in \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, then for $1 \leq i \leq N$ the diagonal block A_{ii} contains a unique maximal irreducible principal submatrix. If this matrix is a permutation matrix, then we say that i is a *cycle component* of A. We let $\mathcal{C}_A$ denote the set of cycle components of A. For each i in $\mathcal{C}_A$, we make the cokernel group

$$\mathrm{cok}(I - A)_{ii} = \mathbb{Z}^{n_i}/(I - A)_{ii}Z^{n_i} \;\cong\; \mathbb{Z}$$

an ordered group as described above. For A and A' in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, if (U, V) is an $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence from A to A', then for $1 \leq i \leq N$ the equivalence (U, V) induces an $\mathrm{SL}(\mathbb{Z})$ equivalence (U_{ii}, V_{ii}) from A_{ii} to A'_{ii}, and this induces an isomorphism from $\mathrm{cok}(I - A)_{ii}$ to $\mathrm{cok}(I - A')_{ii}$ by the rule $[x] \mapsto [U_{ii}x]$. We say that the $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence (U, V) is *positive on cycle components* if this induced isomorphism of the ith component cokernel groups is positive whenever i is a cycle component for both A and A'. For example, if

$$A_{ii} = A'_{ii} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad U_{ii} = V_{ii} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$

then (U, V) is not positive on cycle components.

Proposition 2.7. *Suppose (U, V) is a positive $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalence from $(I - A)$ to $(I - A')$ in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$. Then:*

(1) *A and A' have the same cycle components, and*
(2) *(U, V) is positive on cycle components.*

Proof. It suffices to consider the case $(U, V) = (E, I)$ where E is a basic elementary matrix with offdiagonal entry $E(i_1, j_1) = 1$ such that i_1 and j_1 index rows through A_{ii} and i is a cycle component of A.

(1) It is clear from the earlier discussion on positive equivalence that the ith component of A has a unique cycle iff the ith component of A' has a unique cycle.

(2) For any canonical basis vector e_s, the vector Ee_s is nonnegative because E is nonnegative. It follows that (E, I) must be positive on components. □

3. Statement of results.

In this section we state the main results (Theorem 3.1 and Theorem 3.3) which do not involve the mapping class group. We also give Theorems 3.4 and 3.5, which clarify computational issues. The definition of flow equivalence is given in Section 6, and all discussion of the mapping class group results is deferred to Section 7.

We need a little more notation. Given $\mathcal{P}$, we will use the same index set $\mathcal{I}^{\mathcal{P}}$, a disjoint union of countably infinite sets $\mathcal{I}^{\mathcal{P}}_p$, $p \in \mathcal{P}$, for all matrices with $\mathcal{P} \times \mathcal{P}$ blocking into infinite subblocks. Given finite posets $\mathcal{P}, \mathcal{P}'$, let $\mathrm{Iso}[\mathcal{P}, \mathcal{P}']$ be the set of poset isomorphisms from $\mathcal{P}$ to $\mathcal{P}'$. For each ν in $\mathrm{Iso}[\mathcal{P}, \mathcal{P}']$, fix an infinite permutation matrix $P = P_\nu$ such that

$$P(i,j) = 1 \text{ and } j \in \mathcal{I}^{\mathcal{P}}_p \;\Rightarrow\; i \in \mathcal{I}^{\mathcal{P}'}_{\nu(p)}.$$

Informally, a block $P\{p, q\}$ is zero if $q \neq \nu(p)$ and is the (infinite) identity matrix if $q = \nu(p)$.

Theorem 3.1 (Classification Theorem)**.** *Suppose A is in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$ and A' is in $\mathfrak{M}^{\circ}_{\mathcal{P}',+}(\mathbb{Z})$. The following are equivalent:*

1. *The* SFTs *σ_A and $\sigma_{A'}$ are flow equivalent.*
2. *For some $\nu \in \mathrm{Iso}[\mathcal{P}, \mathcal{P}']$, with $P = P_\nu$: there exists a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence from $(I - A)$ to $(I - P^{-1}A'P)$ in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$.*
3. *For some $\nu \in \mathrm{Iso}[\mathcal{P}, \mathcal{P}']$, with $P = P_\nu$: A and $P^{-1}A'P$ have the same cycle components, and there exists an $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence from $(I - A)$ to $(I - P^{-1}A'P)$ which is positive on cycle components.*

Remarks 3.2.

1. There are only finitely many automorphisms $\nu\colon \mathcal{P} \to \mathcal{P}'$, and they are easily computed. So, we can decide (3) in Theorem 3.1 if we can decide it in the case where $P = I$ and $\mathcal{P} = \mathcal{P}'$.
2. The content of Theorem 3.1 is contained in [**H4**, **H5**]. We will prove the equivalence (1) $\iff$ (2) in Section 6. The implication (2) $\implies$ (3) is trivial. The implication (3) $\implies$ (2) follows from the main contribution of this paper, which is the next theorem.

Theorem 3.3 (Factorization Theorem)**.** *Suppose A and A' are in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$, and $(U, V)\colon (I - A) \to (I - A')$ is an $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence. The following are equivalent:*

1. *$(U, V)\colon (I - A) \to (I - A')$ is a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$.*

(2) *A and A' have the same cycle components, and (U, V) is positive on cycle components.*

Below, given a matrix A in any $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, we let A_∞ denote its embedding in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$.

Theorem 3.4. *Suppose A and A' are in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, where $\mathbf{n} = (n_1, \ldots, n_N)$ and the following hold for $1 \le i \le N$:*

- *$n_i = 1 \Leftrightarrow i$ is a cycle component of $A \Leftrightarrow i$ is a cycle component of A',*
- *$n_i = 1$ or $n_i = \infty$.*

Then the following are equivalent:

(1) *There exists a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence from $(I - A_\infty)$ to $(I - A'_\infty)$ in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$.*

(2) *$(I - A)$ and $(I - A')$ are $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalent.*

Proof of Theorem 3.4. (2) $\implies$ (1) Suppose (U, V) is the $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalence. If $n_i = 1$, then $U_{ii} = V_{ii} = 1$ because $\{U, V\} \subset \mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$. So, the embeddings of U and V in $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ give an $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence $(I - A_\infty) \to (I - A'_\infty)$ in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$ which is positive on cycle components.

(1) $\implies$ (2) This follows from Lemmas A.3 and A.7. □

The point of Theorem 3.4 is to give a flow equivalence criterion in terms of just $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalence (which is characterized in [**BH**]), without a "positive on cycle components" condition. Given matrices A_1 and A'_1 in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, Lemmas A.1 and A.2 give us positive equivalences, from $I - A_1$ to $I - A$ and from $I - A'_1$ to $I - A'$, such that A_∞ and A'_∞ are of the form described in Theorem 3.4.

Theorem 3.5 ([**BH**])**.** *Suppose B and B' are matrices in $\mathfrak{M}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ such that for each diagonal block in B or B', the greatest common divisor of the entries of the block is* 1*. Suppose $\mathbf{n} \le \mathbf{r}$, and let ι be the embedding of $\mathfrak{M}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ into $\mathfrak{M}_{\mathcal{P}}(\mathbf{r}, \mathbb{Z})$.*

Then B is $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalent to B' if and only if ιB is $\mathrm{SL}_{\mathcal{P}}(\mathbf{r}, \mathbb{Z})$ equivalent to $\iota B'$.

Theorem 3.5, taken from the Stabilization result in [**BH**], reduces the problem of $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence of the matrices $(I - A)$ and $(I - A')$ in Theorem 3.4 to an equivalence problem for finite matrices.

4. Factorization: The proof.

To begin, we describe a matrix class in which our positivity considerations will be simplified.

Definition 4.1. Given a subset $\mathcal{C}$ of $\{1, \dots, N\}$, and a vector $\mathbf{n}$ with positive integer entries such that $n_i = 1$ if $i \in \mathcal{C}$, define $\mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$ to be the set of $\mathbf{n}$-blocked integral matrices M whose blocks M_{ij} satisfy the following conditions:

- $M_{ii} = 0$ if $i \in \mathcal{C}$,
- $M_{ij} = 0$ if $i \neq j$ and $i \not\prec j$,
- $M_{ij} > 0$ otherwise.

(So, each block of M has all entries zero or all entries greater than zero, $M_{ii} = 0$ when $i \in \mathcal{C}$, and otherwise $M_{ij} > 0$ if and only if $i \preceq j$. If $-M = I - A$, then $\mathcal{C}$ is the set of cycle components of A.)

Definition 4.2. An elementary positive equivalence in $\mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$ is an $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalence $(U, V)\colon B \to B' = UBV$ such that $\{B, B'\} \subset \mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$; one of U, V equals Id; and the other is a basic elementary matrix. A positive equivalence in $\mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$ is a composition of elementary positive equivalences in $\mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$. For such an equivalence (U, V), we use notations such as

$$(U, V)\colon B \underset{+}{\longrightarrow} B' \quad \text{or} \quad B \overset{(U,V)}{\underset{+}{\longrightarrow}} B' \quad \text{or} \quad B \underset{+}{\longrightarrow} B'.$$

Observation 4.3. Suppose $B = (A - I)$, $B' = (A' - I)$ and

$$(U, V)\colon B \underset{+}{\longrightarrow} B'.$$

Then $(U, V)\colon (I - A) \to (I - A')$ is a positive equivalence in $\mathfrak{M}_{\mathcal{P},+}^{\circ}(\mathbf{n}, \mathbb{Z})$.

Outline of the proof. Now we can give an outline of the proof of the Factorization Theorem (3.3), which we break into four steps.

Step 1 of the proof ("block positive reduction") is to reduce it to proving the following theorem:

Theorem 4.4. *Suppose $B = A - I$ and $B' = A' - I$, satisfying*

- *B and B' are in $\mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$*
- *$(U, V)\colon B \to B'$ is an $\mathrm{SL}(\mathbf{n}, \mathbb{Z})$ equivalence*
- *if $i \notin \mathcal{C}$, then $\dim(\ker(A_{ii})) \geq 2$.*

Then

$$(U, V)\colon B \underset{+}{\longrightarrow} B'.$$

Step 2 ("the positive case") is to prove Theorem 4.4 in the case B and B' are positive (i.e., $\mathcal{P} = \{1\}$ and $\mathcal{C} = \emptyset$). This step is the heart of the proof and it is carried out in Section 5. This is the only step which uses the condition $\dim(\ker(A_{ii})) \geq 2$.

Step 3 ("the unipotent case") is to prove Theorem 4.4 in the case that U and V lie in $\mathcal{U}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$, where $\mathcal{U}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ denotes the set of matrices M in $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ such that $M_{ii} = I$ for all i in $\mathcal{P}$.

Step 4 ("the general case") is to finish the proof of Theorem 4.4.

Step 1: *Block positive reduction.* We will accomplish this step by proving the following proposition. Let $\mathcal{C}_A = \mathcal{C}$ denote the set of cycle components of A. For each cycle component i, let $\mathcal{C}_i^{\rm sec}$ denote the set of indices for rows/columns through A_{ii} such that i does not lie on a cycle, and let $\mathcal{C}_i^{\rm prim}$ denote the set of indices i for rows/columns through A_{ii} which lie on the unique cycle in A_{ii}. Let $\mathcal{C}^{\rm prim} = \cup \mathcal{C}_i^{\rm prim}$ and $\mathcal{C}^{\rm sec} = \cup \mathcal{C}_i^{\rm sec}$.

Proposition 4.5. *Suppose* $\{A, A'\} \subset \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$*;* $\mathcal{C}_A = \mathcal{C}_{A'} = \mathcal{C}$*;* $(U, V) : (I - A) \to (I - A')$ *is an* $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ *equivalence which is positive on cycle components; and* $\mathbf{n} = (n_1, \ldots n_N)$ *has positive integer entries. Then there is a commuting diagram of* $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ *equivalences*

$$\begin{array}{ccc} (I - A) & \longrightarrow & (I - \overline{A}) \\ {\scriptstyle (U,V)}\downarrow & & \downarrow{\scriptstyle (\overline{U},\overline{V})} \\ (I - A') & \longrightarrow & (I - \overline{A}') \end{array}$$

such that:

(1) *The horizontal arrows are positive* $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ *equivalences in* $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$.
(2) *For both* $\overline{A}$ *and* $\overline{A}'$*, the principal submatrix indexed by the complement of* $\mathcal{C}^{\rm sec}$ *is strictly positive wherever the* $\mathcal{P}$ *ordering permits a nonzero entry, and in addition the diagonal blocks* $(\overline{A} - I)_{tt}$ *and* $(\overline{A}' - I)_{tt}$ *are strictly positive whenever* $t \notin \mathcal{C}$.
(3) *For both* $\overline{A}$ *and* $\overline{A}'$*,* $\mathcal{C}^{\rm prim}$ *is the set of indices* ℓ *such that for some* $i \in \mathcal{C}$*,* (ℓ, ℓ) *indexes the upper left corner of the* ii *block.*
(4) $\overline{A}(i, j) = \overline{A'}(i, j) = 0$ *whenever* $\{i, j\} \cap \mathcal{C}^{\rm sec} \neq \emptyset$.
(5) $\overline{U}(i, j) = \overline{V}(i, j) = \delta_{ij}$ *whenever* $\{i, j\} \cap \mathcal{C}^{\rm sec} \neq \emptyset$.

For matrices $\overline{A}, \overline{A}'$ in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, we say an $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalence $(\overline{U}, \overline{V}) : (I - \overline{A}) \to (I - \overline{A}')$ is **nondegenerate** if it satisfies Conditions (2), (3), (4) and (5) of Proposition 4.5. Note Condition (3) implies that $\overline{A}(\ell, \ell) = 1$ if $\ell \in \mathcal{C}^{\rm prim}$.

Let us see that Proposition 4.5 reduces the proof of (2) $\implies$ (1) in the Factorization Theorem 3.3 to the proof of Theorem 4.4. Given $(I - A)$, $(I - A')$ and (U, V) satisfying (2) in the statement of Theorem 3.3, pick a vector $\mathbf{n}$ with positive integer entries large enough that:

- For all i, j in $\mathcal{P}$, the ij blocks of $U, V, I - A$ and $I - A'$ agree with $\delta_{ij} I$ outside the upper left $n_i \times n_j$ corner, and
- if $i \notin \mathcal{C}$, then the upper left $n_i \times n_i$ corners of A_{ii} and A'_{ii} have kernels of dimension at least two.

Replace A, A', U and V with their trunctations to $\mathbf{n}$-blocked matrices. Then it suffices to prove that $(U,V)\colon (I-A) \to (I-A')$ is a positive $\mathrm{SL}(\mathbf{n},\mathbb{Z})$ equivalence in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n},\mathbb{Z})$. To do this, first apply Proposition 4.5 to the matrices U, V, A, A'. Then truncate the resulting $\overline{U}, \overline{V}, \overline{A}, \overline{A'}$ by removing all rows and columns indexed by $\mathcal{C}^{\mathrm{sec}}$, and call the resulting matrices U, V, A, A'. To finish the proof of the Factorization Theorem 3.3, it suffices to show $(U,V)\colon (I-A) \to (I-A')$ is a positive equivalence, and this now follows by an application of Theorem 4.4 and Observation 4.3.

We want Proposition 4.5 in order to have a completely general result about factoring equivalences into positive equivalences, and in order to see the main arguments more clearly in the less technical setting of $\mathfrak{M}^{++}_{\mathcal{P}}(\mathcal{C},\mathbf{n},\mathbb{Z})$. Because the proof of Proposition 4.5 is tedious (almost entirely on account of technicalities involving cycle components), we relegate the proof of Proposition 4.5 to Appendix A.

Below, we use $\mathcal{U}_{\mathcal{P}}$ to denote $\mathcal{U}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ and we use $\mathfrak{M}^{++}_{\mathcal{P}}$ to denote $\mathfrak{M}^{++}_{\mathcal{P}}(\mathcal{C}, \mathbf{n},\mathbb{Z})$. For i, j in $\mathcal{P}$ and B a matrix with a $\mathcal{P}$-indexed block structure, we let $B\{i,j\}$ denote the ij block of B.

Step 2: *The positive case.* This is carried out in Section 5.

Step 3: *The unipotent case.*

Lemma 4.6. *Suppose U and V are matrices in $\mathcal{U}_{\mathcal{P}}$, B and B' are in $\mathfrak{M}^{++}_{\mathcal{P}}$, and $UBV = B'$. Then*

$$B \xrightarrow[+]{(U,V)} B'.$$

Proof. Write U as a product of matrices in $\mathcal{U}_{\mathcal{P}}$, $U = U_n \cdots U_1$, where for each U_t there is an associated pair (i_t, j_t), such that the following hold:

- $U_t = I$, except in the block $U_t\{i_t, j_t\}$, and
- if $s \neq t$, then $(i_s, j_s) \neq (i_t, j_t)$.

Note, whenever i_s is an immediate predecessor of j_s in $\mathcal{P}$ and $B\{i_s, i_s\} = 0$, these conditions imply

$$(U_s B)\{i_s, j_s\} = B'\{i_s, j_s\}. \tag{4.7}$$

We claim there are nonnegative matrices $Q_1, \ldots, Q_n$ in $\mathcal{U}_{\mathcal{P}}$ such that (with $Q = Q_1 \cdots Q_n$)

$$B \xrightarrow[+]{(U_1,Q_1)} \cdot \xrightarrow[+]{(U_2,Q_2)} \cdots \xrightarrow[+]{(U_n,Q_n)} U_n \cdots U_1 B Q_1 \cdots Q_n = UBQ. \tag{4.8}$$

To show (4.8), first we will produce Q_1 such that

$$B \xrightarrow[+]{(U_1,Q_1)} U_1 B Q_1.$$

Denote (i_1, j_1) as (i, j). Factor U_1 as $U_1 = U_1^- U_1^+$, where U_1^- and U_1^+ equal I outside the $\{i, j\}$ block, $U_1^+\{i, j\}$ is the nonnegative part of $U_1\{i, j\}$, and $U_1^-\{i, j\}$ is the nonpositive part of $U_1\{i, j\}$. Clearly

$$(U_1^+, I)\colon B \underset{+}{\longrightarrow} U_1^+ B.$$

For U_1^- there are two cases.

Case I: $B\{i, i\} > 0$. We have $U_1^-(U_1^+ B) = U_1^+ B$ outside blocks $\{i, k\}$ such that $i \prec j \preceq k$. Because $(U_1^+ B)\{i, i\} = B\{i, i\} > 0$, we can pick Q_1 in $\mathcal{U}_\mathcal{P}$, with sufficiently large nonnegative entries in such blocks $\{i, k\}$, to put $U_1^-(U_1^+ B)Q_1$ into $\mathfrak{M}_\mathcal{P}^{++}$. Then

$$U_1^+ B \xrightarrow[+]{(I,Q_1)} U_1^+ BQ_1 \xrightarrow[+]{(U_1^-,I)} U_1^- U_1^+ BQ_1 = U_1 BQ_1.$$

Case II: $B\{i, i\} = 0$. Again, $U_1^-(U_1^+ B) = U_1^+ B$ outside blocks $\{i, k\}$ such that $i \prec j \preceq k$. Because $B\{i, j\} > 0$, we can choose Q_1 nonnegative in $\mathcal{U}_\mathcal{P}$ such that for all k satisfying $i \prec j \prec k$, we have $U_1^-(U_1^+ B)Q_1\{i, k\} > 0$. (A positive entry in the block $Q\{j, k\}$ acts here to add a multiple of a column through the $\{i, j\}$ block to a column in the $\{i, k\}$ block.) If there is some h such that $i \prec h \prec j$, then suitable positive entries in $Q_1\{h, k\}$ will also achieve $U_1^-(U_1^+ B)Q_1\{i, j\} > 0$. If there is no such h, then i is an immediate predecessor of j in $\mathcal{P}$, and by appeal to (4.7) we have

$$\begin{aligned}\big(U_1^-(U_1^+ B)Q_1\big)\{i, j\} &= (U_1 B)\{i, j\} \\ &= B'\{i, j\} > 0.\end{aligned}$$

Therefore

$$U_1^+ B \xrightarrow[+]{(I,Q_1)} U_1^+ BQ_1 \xrightarrow[+]{(U_1^-,I)} U_1 BQ_1$$

as required.

Thus in either case we have

$$B \xrightarrow[+]{(U_1,Q_1)} U_1 BQ_1 \in \mathfrak{M}_\mathcal{P}^{++}.$$

An easy induction on the argument gives (4.8), with

$$B \xrightarrow[+]{(U,Q)} UBQ \in \mathfrak{M}_\mathcal{P}^{++},$$

with Q a product of nonnegative elementary matrices in $\mathcal{U}_\mathcal{P}$. The transposed argument gives a matrix P in $\mathcal{U}_\mathcal{P}$ such that P is a product of nonnegative elementary matrices such that

$$B' \xrightarrow[+]{(P,V^{-1})} PB'V^{-1} \in \mathfrak{M}_\mathcal{P}^{++}.$$

Then

$$B \xrightarrow[+]{(U,Q)} UBQ \xrightarrow[+]{(P,I)} PUBQ = PB'V^{-1}Q \xleftarrow[+]{(I,Q)} PB'V^{-1} \xleftarrow[+]{(P,V^{-1})} B'$$

so

$$(P^{-1}PU, QQ^{-1}V) = (U,V)\colon B \underset{+}{\longrightarrow} B'$$

as required. □

Step 4: *The general case.*

Lemma 4.9. *Suppose $i \notin \mathcal{C}$, E is a basic elementary matrix in $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$, $E\{j,k\} = (\mathrm{Id})\{j,k\}$ when $(j,k) \neq (i,i)$, $\{B, B'\} \subset \mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$ and*

$$(E\{i,i\}, \mathrm{Id})\colon B\{i,i\} \underset{+}{\longrightarrow} B'\{i,i\}.$$

Then there exists V in $\mathcal{U}_{\mathcal{P}}$ such that

$$(E,V)\colon B \underset{+}{\longrightarrow} B'.$$

Similarly, if

$$(\mathrm{Id}, E\{i,i\})\colon B\{i,i\} \underset{+}{\longrightarrow} B'\{i,i\}$$

then there exists U in $\mathcal{U}_{\mathcal{P}}$ such that

$$(U,E)\colon B \underset{+}{\longrightarrow} B'.$$

Proof. We will consider the equivalence (E,I), the other case is similar. Let $E(s,t)$ be the nonzero offdiagonal entry of E. If $E(s,t) = 1$, then set $V = \mathrm{Id}$. Now suppose $E(s,t) = -1$, so E acts from the the left to subtract row t from row s. Then possibly there are nonpositive entries in blocks $(EB)\{i,j\}$ where $i \prec j$. To correct for this, pick r an index for a column through the ii block; note that $B(s,r) > B(t,r)$ because $(EB)\{i,i\} > 0$ by assumption; consider a positive integer M; and let V be the matrix in $\mathcal{U}_{\mathcal{P}}$ which acts from the right to add column r to column q, M times, for every q indexing a column through an ij block for which $i \prec j$. For these q,

$$(EBV)(s,q) = M(B(s,r) - B(t,r)) + B(s,q) - B(t,q).$$

So, if M is large enough, then this gives

$$B \xrightarrow[+]{(I,V)} BV \xrightarrow[+]{(E,I)} EBV$$

as required. □

Proof of the general case. Now let $(U,V)\colon B \to B'$ be the $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ equivalence, with $\{B,B'\} \subset \mathfrak{M}_{\mathcal{P}}^{++}(\mathcal{C},\mathbf{n},\mathbb{Z})$. By Step 2 (Theorem 5.1), for each $i \in \mathcal{P} \setminus \mathcal{C}$, we have

$$(U\{i,i\},V\{i,i\})\colon B\{i,i\} \underset{+}{\longrightarrow} B'\{i,i\}.$$

So, we may find a string of elementary equivalences, say $(E_1,F_1),\dots(E_t,F_t)$, with every $E_t\{i,j\} = F_t\{i,j\} = (\mathrm{Id})\{i,j\}$ unless $i=j \notin \mathcal{C}$, which accomplishes the elementary positive equivalence decomposition inside the diagonal blocks. By Lemma 4.9, we may find $(U_1,V_1),\dots,(U_t,V_t)$ with each U_s and V_s in $\mathcal{U}_{\mathcal{P}}$, such that

$$B \xrightarrow[+]{(U_1,F_1)} \cdot \xrightarrow[+]{(E_1,V_1)} \cdots \xrightarrow[+]{(U_t,F_t)} \cdot \xrightarrow[+]{(E_t,V_t)} B''.$$

Let $X = E_tU_t\cdots E_2U_2E_1U_1$. Let $Y = F_1V_1F_2V_2\cdots F_tV_t$. Then for all i in $\mathcal{P}$, $X\{i,i\} = U\{i,i\}$ and $Y\{i,i\} = V\{i,i\}$, so $UX^{-1} \in \mathcal{U}_{\mathcal{P}}$ and $Y^{-1}V \in \mathcal{U}_{\mathcal{P}}$. It follows from Step 3 (Lemma 4.6) that

$$B'' \xrightarrow[+]{(UX^{-1},Y^{-1}V)} B'.$$

Thus $(U,V)\colon B \to B'$ is the composition

$$B \xrightarrow[+]{(X,Y)} B'' \xrightarrow[+]{(UX^{-1},Y^{-1}V)} B'$$

and this finishes the proof. □

5. Factorization: The positive case.

In this section, all matrices are $K \times K$, where K is a positive integer and $K > 1$. We let $\mathfrak{M}_+$ denote the set of $K \times K$ matrices with strictly positive integer entries.

We say an equivalence $(U,V)\colon B \to B'$ is a positive equivalence through $\mathfrak{M}_+$ if it can be given as a chain of positive elementary equivalences

$$B = B_0 \to B_1 \to B_2 \to \cdots \to B_n = B'$$

in which every B_i is in $\mathfrak{M}_+$.

The purpose of this section is to prove the following theorem.

Theorem 5.1. *Suppose U and V are in $\mathrm{SL}(K,\mathbb{Z})$, and B and UBV are in $\mathfrak{M}_+$. Suppose also that B is $\mathrm{SL}(K,\mathbb{Z})$ equivalent to a diagonal matrix in which at least two entries equal 1.*

Then $(U,V)\colon B \to UBV$ is a positive equivalence through $\mathfrak{M}_+$.

Remark 5.2. The "two entries" technical assumption may be excessive, but is harmless for our applications. Except for the final argument which addresses the possibility that UB is nonpositive, we only use the weaker assumption that B has rank greater than one.

The proof of Theorem 5.1 rests on three lemmas. We begin the preparations.

By a *signed transposition matrix*, we mean a matrix which is the matrix of a transposition, but with one of the off-diagonal 1's replaced by -1. By a *signed permutation matrix* we mean a product of signed transposition matrices. Since $K > 1$, any $K \times K$ permutation matrix with determinant 1 is a signed permutation matrix. A $K \times K$ matrix S is a signed permutation matrix if and only if $\det S = 1$ and the matrix $|S|$ is a permutation matrix (where $|S|(i,j) := |S(i,j)|$).

Lemma 5.3. *Suppose $B \in \mathfrak{M}_+$, E is a basic elementary matrix with nonzero offdiagonal entry $E(i,j)$, and the ith row of EB is not the zero row.*

Then in $\mathrm{SL}(K,\mathbb{Z})$ *there are a nonnegative matrix Q and a signed permutation matrix S such that $(SE,Q)\colon B \to SEBQ$ is a positive equivalence through $\mathfrak{M}_+$.*

Proof. If $E(i,j) = 1$, then let $Q = I = S$. Now suppose $E(i,j) = -1$, so E acts from the left by subtracting row j from row i, and the rows i and j of B are not equal.

Case I: For some k, $B(i,k) > B(j,k)$.

Here we may repeatedly add column k of B to other columns, until we have a matrix B' with $B'(i,m) > B'(j,m)$ for all m. This B' is BQ for some Q which is a product of nonnegative basic elementary matrices. Now $(E,Q)\colon B \to EBQ$ is the composition of positive equivalences, $(I,Q)\colon B \to BQ$ followed by $(E,I)\colon BQ \to EBQ$. Let $S = I$.

Case II: For every k, $B(i,k) \le B(j,k)$.

Because the rows i and j of B are not equal, after multiplying from the right by a suitable Q we can assume in this Case that $0 < B(i,k) < B(j,k)$ for all k. Now $(I,Q)\colon B \to BQ$ in $\mathfrak{M}_+$, so for notational simplicity from here we may assume $Q = I$.

For concreteness of notation, let $(i,j) = (1,2)$. For the rest of this Case, for simplicity we will restrict what we write to these two rows, e.g.,

$$E = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

where B_1 and B_2 denote the first and second rows of B, and we have $B_1 < B_2$. Let $S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. Then

$$(SE)B = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = \begin{pmatrix} B_2 \\ B_2 - B_1 \end{pmatrix}$$

and the latter matrix is positive. Let $E' = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and let $E'' = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$, then

$$SE = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix} = E'E''.$$

Now $(E'', I)\colon B \to E''B$ is a positive equivalence in $\mathfrak{M}_+$, since row 2 of B is positive and greater than row 1; and $(E', I)\colon E''B \to E'E''B$ is also a positive equivalence in $\mathfrak{M}_+$. □

Lemma 5.4. *Suppose B is a $K \times K$ integral matrix of rank at least 2, and U is in $\mathrm{SL}(K, \mathbb{Z})$, and no row of B or UB is the zero row. Then U is the product of basic elementary matrices, $U = E_k \cdots E_1$, such that for $1 \le j \le k$ the matrix $E_j E_{j-1} \cdots E_1 B$ has no zero row.*

Proof. Without loss of generality, assume $K \ge 3$ and U is not the identity. Let $\mathcal{E}(i)$ denote the set of integral matrices which equal I both on the diagonal and outside of row i. Let $\mathcal{U}$ be the set of factorizations $U = U_n \cdots U_1$ such that for $1 \le h \le n$, the matrix U_h is not the identity and there is an index i_h such that $U_h \in \mathcal{E}(i_h)$. Given such a factorization $U = U_n \cdots U_1$, let

$$z = \#\{h\colon 1 \le h \le n \text{ and row } i_h \text{ of } U_h \cdots U_1 B \text{ is the zero row}\}.$$

Step 1. We will produce an element of $\mathcal{U}$ for which $z = 0$.

By induction, it suffices to begin with a factorization $U = U_n \cdots U_1$ from $\mathcal{U}$ for which $z > 0$, and produce another factorization from $\mathcal{U}$ with reduced z. Pick s minimal such that row i_s of $U_s \cdots U_1 B$ is zero, and let t be minimal such that $t > s$ and $i_t = i_s$. (This t exists because row i_s of UB is nonzero.) We will change the factorization by replacing the subword $U_t \cdots U_s$ with a suitable word $U'_T \cdots U'_s$, to be defined recursively.

First pick $j_s \ne i_s$ such that row j_s of $U_{s-1} \cdots U_1 B$ is nonzero ($U_{s-1} \cdots U_1 B$ just denotes B in the case that $s = 1$). Choose F_s an elementary matrix which acts to add a multiple of row j_s to row i_s, such that (for notational simplicity) $F_s^{-1} U_s \ne I$. Define $U'_s = F_s^{-1} U_s \in \mathcal{E}(i_s)$. Now $U_t \cdots U_s = U_t \cdots U_{s+1} F_s U'_s$ and row i_s of $U'_s U_{s-1} \cdots U_1 B$ is not zero.

Now we give the recursive step. Suppose $s < m \le t$ and we have produced $U_t \cdots F_{m-1} U'_r \cdots U'_s = U_t \cdots U_s$ such that there is a nonzero integer c_{m-1} and an index $j_{m-1} \ne i_s$ such that $F_{m-1}(i_s, j_{m-1}) = c_{m-1}$ and otherwise $F_{m-1} = I$. We will replace $U_m F_{m-1}$ with new terms. There are three cases.

Case 1: *$m < t$ and $j_{m-1} \ne i_m$.* Set $F_m = F_{m-1}$ and $U'_{r+1} = F_m^{-1} U_m F_m$. For example, if $K = 3$ and $(i_s, i_m, j_{m-1}) = (1, 2, 3)$, then we would have for

some a, b, c that

$$U'_{r+1} = F_m^{-1} U_m F_m = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -c & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & b \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ c & 0 & 1 \end{pmatrix}$$
$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -c & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ a+bc & 1 & b \\ c & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ a+bc & 1 & b \\ 0 & 0 & 1 \end{pmatrix}.$$

Now $U'_{r+1} \in \mathcal{E}(i_m)$ and $F_m U'_{r+1} = U_m F_{m-1}$ and row i_m of $U'_{r+1} U'_r \cdots U'_s U_{s-1} \cdots U_1 B$ equals row i_m of $U_m \cdots U_1 B$.

Case 2: $m < t$ *and* $j_{m-1} = i_m$. Choose an index j_m such that $j_m \notin \{i_m, i_s\}$ and row j_m of $U'_r \cdots U'_s U_{s-1} \cdots U_1 B$ is not zero. This is possible because rows i_s and j_{m-1} of $U'_r \cdots U'_s U_{s-1} \cdots U_1 B$ are linearly dependent (since row i_s of $F_m U'_r \cdots U'_s U_{s-1} \cdots U_1 B$ equals row i_s of $U_m \cdots U_1 B$ which is the zero row) and $\operatorname{rank}(B) \geq 2$. Pick F_m with $F_m(i_s, j_m) = 1$ and otherwise $F_m = I$. Set $U'_{r+1} = F_m^{-1} F_{m-1}$ and $U'_{r+2} = F_m^{-1} U_m F_m$. Now

- $F_m U'_{r+2} U'_{r+1} = F_m (F_m^{-1} U_m F_m)(F_m^{-1} F_{m-1}) = U_m F_{m-1}$,
- $U'_{r+1} \in \mathcal{E}(i_s)$ and row i_s of $U'_{r+1} \cdots U'_s U_{s-1} \cdots U_1 B$ is not zero,
- $U'_{r+2} \in \mathcal{E}(i_m)$ and row i_m of $U'_{r+2} \cdots U'_s U_{s-1} \cdots U_1 B$ equals row i_m of $U_m \cdots U_1 B$.

Case 3: $m = t$. If $U_t F_{t-1} \neq I$, then set $U'_T = U'_{r+1} = U_t F_{t-1} \in \mathcal{E}(i_s)$: Row i_s is the same in the matrices $U_m \cdots U_1 B$ and $U'_T \cdots U'_s U_{s-1} \cdots U_1 B$. If $U_t F_{t-1} = I$, then simply delete $U_t F_{t-1}$, so $U'_T = U'_r$.

The new factorization has z reduced. This concludes Step 1.

Step 2. Suppose we have the factorization from $\mathcal{U}$ with $z = 0$, $U = U_n \cdots U_1$, with $U_h \in \mathcal{E}(i_h)$. For $1 \leq h \leq n$, we will replace U_h with a suitable product of elementary matrices in $\mathcal{E}(i_h)$. The argument will be clear from the case $h = 1$. For notational simplicity, suppose $i_1 = 1$. Write U_1 as a product $U_1 = E_k \cdots E_1$ of basic elementary matrices which agree with I outside row 1. Now, choose a row $m > 1$ of B which is not a rational multiple of row 1 of $U_1 B$ (such a row m exists because $\operatorname{rank}(B) > 1$). Let E_0 be the elementary matrix which adds row m to row 1: If $s > 0$, then $(E_0)^s B$ has row 1 nonzero. Choose a nonnegative integer M large enough that for $1 \leq j \leq k$, row 1 of $[E_j \cdots E_1 (E_0)^M] B$ is nonzero. Then for $0 \leq s \leq M$,

$$[E_0^{-s}][E_k \cdots E_1 (E_0)^M] B = [E_0^{M-s}][E_k \cdots E_1] B$$
$$= [E_0^{M-s}] U_1 B$$

and therefore row 1 of $[E_0^{-s}][E_k \cdots E_1 (E_0)^M] B$ cannot be zero. Thus the factorization $U_1 = (E_0)^{-M} E_k \cdots E_1 (E_0)^M$ has the required properties. □

Lemma 5.5 (Key Lemma)**.** *Suppose B and B' are in $\mathfrak{M}_+$, U and W are in $\mathrm{SL}(K,\mathbb{Z})$, the matrix UB has at least one strictly positive entry, and $UB = B'W$. Then the equivalence $(U, W^{-1})\colon B \to B'$ is a positive equivalence through $\mathfrak{M}_+$.*

Proof. Step 1: Reduction to the case $UB > 0$.

Consider an entry $(UB)(i,j) > 0$. We can repeatedly add column j to other columns until row i of UB has all entries strictly positive. This corresponds to multiplying from the left by a nonnegative matrix Q in $\mathrm{SL}(K,\mathbb{Z})$, giving $UBQ = B'WQ$. Then we can repeatedly add row i of UBQ to other rows until all entries of UBQ are positive. This corresponds to multiplying from the left by a matrix P in $\mathrm{SL}(K,\mathbb{Z})$, giving

$$(PU)(BQ) = (PB')(WQ) > 0$$

with positive equivalences in $\mathfrak{M}_+$ given by

$$(I,Q)\colon B \to BQ, \qquad (P,I)\colon B' \to PB'.$$

Therefore, after replacing (U, B, B', W) with (PU, BQ, PB', WQ), we may assume without loss of generality that $UB > 0$.

Step 2: Reducing the length of an elementary factorization.

By Lemma 5.4, we can write U has a product of basic elementary matrices, $U = E_k \cdots E_1$, such that for $1 \le j \le k$, the matrix $B_j = E_j \cdots E_1 B$ has no zero row. By Lemma 5.3, given the pair (E_1, B), there is a nonnegative Q_1 in $\mathrm{SL}(K,\mathbb{Z})$ and a signed permutation matrix S_1 such that

$$(S_1E_1, Q_1)\colon B \to S_1E_1BQ_1$$

is a positive equivalence in $\mathfrak{M}_+$. We observe that

$$UBQ_1 = S_1^{-1}[S_1E_kS_1^{-1}]\cdots[S_1E_2S_1^{-1}][S_1E_1]BQ_1.$$

Now, for $2 \le j \le k$, the matrix $S_1E_jS_1^{-1}$ is again a basic elementary matrix E_j', and the matrix $E_j' \cdots E_2'(S_1E_1BQ_1)$ has no zero rows.

Again using Lemma 5.3, for the pair $([S_1E_2S_1^{-1}], [S_1E_1BQ_1])$ choose a signed permutation matrix S_2 and nonnegative Q_2 producing a positive equivalence in $\mathfrak{M}_+$

$$(S_2[S_1E_2S_1^{-1}], Q_2)\colon S_1E_1BQ_1 \to S_2[S_1E_2S_1^{-1}]S_1E_1BQ_1Q_2$$

so that we get a positive equivalence in $\mathfrak{M}_+$

$$([S_2S_1E_2S_1^{-1}][S_1E_1], Q_1Q_2)\colon B \to [S_2S_1E_2E_1BQ_1Q_2]$$

and we observe that

$$UBQ_1Q_2 = S_1^{-1}S_2^{-1}[S_2S_1E_kS_1^{-1}S_2^{-1}]\cdots \\ [S_2S_1E_3S_1^{-1}S_2^{-1}][S_2S_1E_2S_1^{-1}][S_1E_1]BQ_1Q_2.$$

Continue this, to obtain a signed permuation matrix $S = S_k \cdots S_1$ and nonnegative $Q = Q_1 \cdots Q_k$ such that

$$\begin{aligned} UBQ &= S^{-1}[S_k \cdots S_1 E_k S_1^{-1} \cdots S_{k-1}^{-1}] \cdots [S_2 S_1 E_2 S_1^{-1}][S_1 E_1]BQ \\ &= S^{-1}(SUBQ) \end{aligned}$$

and $(SU, Q)\colon B \to SUBQ$ is a positive equivalence in $\mathfrak{M}_+$.

Step 3: Realizing the permutation.

We continue from Step 2. It remains to show that

$$(S, I)\colon UBQ \to SUBQ$$

is a positive equivalence in $\mathfrak{M}_+$. Since S is a product of signed transposition matrices, it may be described as a permutation matrix in which some rows have been multiplied by -1. Since UBQ and $SUBQ$ are strictly positive, it must be that S is a permutation matrix. Also, $\det(S) = 1$, so if $S \neq I$ then S is the matrix of a permutation which is a product of 3-cycles. So it is enough to realize the positive equivalence in $\mathfrak{M}_+$ in the case that S is the matrix of a 3-cycle. For this we write the matrix

$$C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

as the following product $C_0 C_1 \cdots C_5$:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

For $0 \leq i \leq 5$, the matrix $C_i C_{i+1} \cdots C_5$ is nonnegative. Therefore the equivalence $(C, I)\colon B \to CB$ is a positive equivalence through $\mathfrak{M}_+$ whenever $B \in \mathfrak{M}_+$. □

We can now complete the proof of Theorem 5.1. It only remains to address the technical point that in the equivalence $(U, V)\colon B \to B'$, all the entries of UB might be nonpositive. (For example, with K even we could have $(U, V) = (-I, -I)$.)

Proof of Theorem 5.1. By assumption there are X, Y in $\mathrm{SL}(K, \mathbb{Z})$ such that $XBY = D$, where D is diagonal and has the block form $D = \begin{pmatrix} I & 0 \\ 0 & F \end{pmatrix}$, where I is 2×2. For any H in $\mathrm{SL}(2, \mathbb{Z})$, the $K \times K$ matrix $G = G_H = \begin{pmatrix} H & 0 \\ 0 & I \end{pmatrix}$ yields a self equivalence $(X^{-1}GX,\ YG^{-1}Y^{-1})\colon B \to B$.

For a matrix Q, we let $Q\{12;*\}$ denote the submatrix consisting of the first two rows. The matrix $(XBY)\{12;*\} = D\{12;*\}$ has rank two, so the matrix $(XB)\{12;*\}$ has rank two, and we may choose $H' \in \mathrm{SL}(2,\mathbb{Z})$ such that the first row r of $H'[(XB)\{12;*\}]$ has both a positive entry and a negative entry. For $M \in \mathbb{N}$, let $H_M = \begin{pmatrix} M & -1 \\ 1 & 0 \end{pmatrix}$, $H = H_M H'$, and $G = G_H$. Let c denote the first column of X^{-1}. Since c is not the zero vector, the $K \times K$ matrix cr has a positive entry and a negative entry.

If M is sufficiently large, then the entries of the two matrices $X^{-1}GXB$ and Mcr will have the same sign wherever the entries of Mcr are nonzero, and $X^{-1}GXB$ will have a positive entry. Then the Key Lemma 5.5 shows that $(X^{-1}GX, YG^{-1}Y^{-1})$ gives a positive equivalence in $\mathfrak{M}_+$ from B to B.

Similarly, for large enough M the entries of $UX^{-1}GXB$ will agree in sign with the entries of $UMcr$ wherever the entries of the latter matrix are nonzero. Because U is nonsingular, the matrix Ucr is nonzero, and then contains positive and negative entries because r does.

So, using M sufficiently large, we obtain $(U,V)\colon B \to B'$ as a positive equivalence in $\mathfrak{M}_+$, the inverse of $(X^{-1}GX, YG^{-1}Y^{-1})$ followed by $(UX^{-1}GX, YG^{-1}Y^{-1}V)$. □

6. Flow equivalence.

The purpose of this section is to prove the claims of Theorem 3.1 involving flow equivalence. As sketched in [**B1**] (see also [**Ba1**]), the positive K-theory framework is most natural for this. Because a complete development of this connection has not yet appeared, for brevity we will make no direct use of it below.

We begin with some background. For S a selfhomeomorphism of a compact metric space X, the mapping torus Y_S of S is the quotient space $(X \times \mathbb{R})/\sim$ where $(x, n+t) \sim (S^n x, t)$ if $n \in \mathbb{Z}$. Y_S admits a natural flow,

$$Y_S \times \mathbb{R} \to Y_S$$
$$([(x,t)], s) \mapsto [(x, s+t)].$$

This flow has the copy $X_0 = \{[(x,0)]\colon x \in X\}$ of X as a cross section, and the return map to X_0 under the flow (given by $[(x,0)] \mapsto [(Sx,0)]$) is obviously topologically conjugate to S. Let T be another selfhomeomorphism of a compact metric space. Then S and T are *flow equivalent* if and only if there is a homeomorphism $Y_S \to Y_T$ which takes flow lines onto flow lines and respects the direction of the associated flows. (Equivalently: S and T are conjugate to return maps of cross sections of a common flow.)

For example, consider $S = \sigma_A$, $T = \sigma_{A'}$ and the map γ arising from a basic positive equivalence in Subsection 2.2. It is not difficult to see that γ

is the restriction of a homeomorphism $Y_S \to Y_T$ which takes flow lines onto flow lines and respects the direction of the associated flows, and therefore σ_A and $\sigma_{A'}$ are flow equivalent.

Now fix A in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$ and A' in $\mathfrak{M}^{\circ}_{\mathcal{P}',+}(\mathbb{Z})$. Let F and F' be finite matrices such that $F_\infty = A$ and $F'_\infty = A'$. Let $\sigma_A = \sigma_F$ and $\sigma_{A'} = \sigma_{F'}$ be the associated SFTs. (So, for example σ_A is the left shift on the path space Σ_A, which is given the natural zero dimensional metrizable topology.) Parry and Sullivan [**PS**] showed that σ_F and $\sigma_{F'}$ are flow equivalent if and only if σ_F is topologically conjugate to some SFT which after a time change is topologically conjugate to $\sigma_{F'}$. It follows ([**PS**]) that σ_F and $\sigma_{F'}$ are flow equivalent if and only if F' can be obtained from F by a finite sequence of basic flow moves, which are state splittings and stretchings and their inverses. The inverse of a splitting is called an amalgamation. We will describe the splitting and stretching moves now.

Let B and B' be finite square matrices. B' is obtained from B by an elementary row amalgamation if there exist indices i_1, i_2 and i such that the columns i_1 and i_2 of A' are equal, and A is obtained from A' as follows: Add row i_1 to row i_2, then remove the row and column indexed by i_1. The reverse move is that B is obtained from B' by a row splitting. Analogously there are column splittings and amalgamations. By state splittings we mean row splittings and column splittings.

We say B' is obtained from B by a state stretching if for some indices i, j the following hold: $B'(i,j) = 1$, the other entries of row i and column j are zero, and B is the matrix obtained from B' by adding column i to column j and then removing row i and column i.

We are now ready for the proof. Suppose A is in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$ and A' is in $\mathfrak{M}^{\circ}_{\mathcal{P}',+}(\mathbb{Z})$. We will show the following are equivalent:

(1) σ_A and $\sigma_{A'}$ are flow equivalent.
(2) There exists $\nu \in \mathrm{Iso}\,[\mathcal{P}, \mathcal{P}']$ such that for $P = P_\nu$, there exists a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence from $(I - A)$ to $(I - P^{-1}A'P)$ in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$.

Proof. Given (2), it follows from Lemma 2.5 that there is a chain of basic positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalences from $(I - A)$ to $(I - P^{-1}A'P)$. Each basic positive equivalence gives rise to a flow equivalence as discussed above. It follows that (2) implies (1).

Now we assume (1) and will deduce (2). Let F and F' be finite matrices such that $F_\infty = A$ and $F'_\infty = A'$. After using Lemmas A.1 and A.2 to pass to flow equivalent SFTs, we may assume that for each $i \in \mathcal{P}$, the diagonal blocks F_{ii} and F'_{ii} are strictly positive.

From [**PS**] we are given a sequence of basic moves through finite matrices, $F = F_0 \to F_1 \to \cdots \to F_m = F'$. We may regard $\mathcal{P}$ and $\mathcal{P}'$ as the posets of irreducible components of F and F' respectively, where e.g., $i \preceq j$ in $\mathcal{P}$ when there exists a transition from i to j (by which we mean that there exists a

point in the SFT Σ_A forwardly asymptotic to a cycle from component j and backwardly asymptotic to a cycle from component i). Each move $F_i \to F_{i+1}$ induces a bijection of irreducible components, respecting transitions, and thus the composition induces a poset isomorphism $\nu\colon \mathcal{P}(A) \to \mathcal{P}(A')$. After replacing A' with $P^{-1}A'P$, where $P = P_\nu$, we may assume $\mathcal{P} = \mathcal{P}'$ and $\nu = \mathrm{Id}$.

Next, for $1 \le i \le n$, we will associate to F_i a matrix A_i in $\mathfrak{M}^\circ_{\mathcal{P},+}(\mathbb{Z})$ such that (modulo permutations of indices) $\mathrm{tru}(A_i) = F_i$. We must take a little care with the indices, to be able to lift each of the moves $F_i \to F_{i+1}$ to a (positive) $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence $(I - A_i) \to (I - A_{i+1})$. Let $\mathrm{Ind}(B)$ denote the set indexing the rows and columns of a square matrix B. For each F_j, we will define an injection $\tau_j\colon \mathrm{Ind}(F_j) \to \mathcal{I}^{\mathcal{P}}$, and then define $A_j = \iota(F_j)$ by setting

$$\begin{aligned} A_j(s,t) &= F(s',t') && \text{if } (s,t) = (\tau_j(s'), \tau_j(t')) \\ &= 0 && \text{otherwise.} \end{aligned}$$

The maps τ_j will be defined recursively. For $j = 0$, we set $A_0 = A$ and take τ_j to be compatible with the embedding of F as a principal submatrix of A. Now suppose $0 \le j < n$ and τ_j and A_j are given. The transition $F_j \to F_{j+1}$ is given by a basic flow move, and under such a move, every element of $\mathrm{Ind}(F_{j+1})$ is naturally related to one or two elements of $\mathrm{Ind}(F_j)$. (An element i of $\mathrm{Ind}(F_{j+1})$ is related to two elements i_1, i_2 of $\mathrm{Ind}(F_j)$ when the move $F_{j+1} \to F_j$ is a splitting or stretching of the state i into the states i_1, i_2.) In any case, for each i in $\mathrm{Ind}(F_{j+1})$, fix a related vertex $\mathrm{rel}(i)$ in $\mathrm{Ind}(F_j)$. Then choose any map $\tau_{j+1}\colon \mathrm{Ind}(F_{j+1}) \to \mathcal{I}^{\mathcal{P}}$ such that $\tau_j(\mathrm{rel}(i))$ and $\tau_{j+1}(i)$ lie in $\mathcal{I}^{\mathcal{P}}_p$ for the same element p of $\mathcal{P}$. (When i is related to two indices, this p may depend on the choice for $\mathrm{rel}(i)$.) This defines the matrices $A = A_0, A_1, \ldots, A_m$.

Next we will show that each elementary flow move $F_j \to F_{j+1}$ gives rise to a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence $(I - A_j) \to (I - A_{j+1})$. Each of the equivalences we give will be accomplished by elementary matrices which must lie in $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ on account of our choices of indices.

First we show how an elementary row splitting gives rise to a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence. The general construction can be understood from the example

$$B = \begin{pmatrix} a & b & 0 \\ c_1 + c_2 & d_1 + d_2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \to \begin{pmatrix} a & b & b \\ c_1 & d_1 & d_1 \\ c_2 & d_2 & d_2 \end{pmatrix} = B'.$$

Here the positive equivalence $(I-B) \to (I-B')$ is accomplished as follows:

$$
\begin{aligned}
(I-B) &= \begin{pmatrix} 1-a & -b & 0 \\ -(c_1+c_2) & 1-(d_1+d_2) & 0 \\ 0 & 0 & 1 \end{pmatrix} \\
&\to \begin{pmatrix} 1-a & -b & 0 \\ -(c_1+c_2) & 1-(d_1+d_2) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -c_2 & -d_2 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1-a & -b & 0 \\ -(c_1+c_2) & 1-(d_1+d_2) & 0 \\ -c_2 & -d_2 & 1 \end{pmatrix} \\
&\to \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1-a & -b & 0 \\ -(c_1+c_2) & 1-(d_1+d_2) & 0 \\ -c_2 & -d_2 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1-a & -b & 0 \\ -c_1 & 1-d_1 & -1 \\ -c_2 & -d_2 & 1 \end{pmatrix} \\
&\to \begin{pmatrix} 1-a & -b & 0 \\ -c_1 & 1-d_1 & -1 \\ -c_2 & -d_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} 1-a & -b & -b \\ -c_1 & 1-d_1 & -d_1 \\ -c_2 & -d_2 & 1-d_2 \end{pmatrix}.
\end{aligned}
$$

The positive equivalence for a column splitting is constructed similarly.

Next we show that a state stretching gives rise to a positive equivalence. The general construction can be understood from the example

$$
B = \begin{pmatrix} 0 & 0 & 0 \\ 0 & a & b \\ 0 & c & d \end{pmatrix} \to \begin{pmatrix} 0 & 1 & 0 \\ a & 0 & b \\ c & 0 & d \end{pmatrix} = B'.
$$

Here the positive equivalence $(I-B) \to (I-B')$ is accomplished as follows.

$$
\begin{aligned}
(I-B) &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-a & -b \\ 0 & -c & 1-d \end{pmatrix} \\
&\to \begin{pmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ -c & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1-a & -b \\ 0 & -c & 1-d \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -a & 1-a & -b \\ -c & -c & 1-d \end{pmatrix} \\
&\to \begin{pmatrix} 1 & 0 & 0 \\ -a & 1-a & -b \\ -c & -c & 1-d \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -a & 1 & -b \\ -c & 0 & 1-d \end{pmatrix}.
\end{aligned}
$$

At this point we have a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence $(I - A) \to (I - A'_n)$, where there is a permutation matrix Q such that $Q^{-1}A'_nQ = A'$. Because $A' = F'_\infty$ and F' has all diagonal blocks positive, if $A'(i,j) > 0$ then i lies on an A' cycle and j lies on an A' cycle. Therefore the permutation given by Q can be chosen compatible with the poset isomorphism $\nu = \mathrm{Id}$, and the matrix Q is a block diagonal matrix in $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$. It remains to check that $I - A'_n \to Q^{-1}A'_nQ$ is accomplished by a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence. Because Q is block diagonal and we can use compositions, it is enough to give the equivalence in the case that Q is the transposition matrix for indices i, j which lie in some $\mathcal{I}_p^{\mathcal{P}}$. Choose indices α, β in $\mathcal{I}_p^{\mathcal{P}}$ such that A'_n is identically zero in the rows and columns indexed by α and β. Let P be the permutation matrix for the product of transpositions $(i,j)(\alpha,\beta)$. Then P is in $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ and $PA'_nP = Q^{-1}A'_nQ$. This finishes the proof. □

7. The mapping class group.

In this section the symbols S, T denote subshifts (e.g., T is the restriction of some full shift $\sigma_{[n]}$ to a closed shift-invariant subset, which we also call T). As in Section 6, let Y_S denote the mapping torus of S. We regard Y_S as an oriented space, in the sense that the associated flow gives an orientation to each of its orbits (i.e., to each connected component of Y_S). A *flow equivalence* from a subshift S to a subshift T is an orientation preserving homeomorphism $\varphi\colon Y_S \to Y_T$ (where "orientation preserving"means orientation preserving on each orbit). Two such homeomorphisms φ_0, φ_1 are *isotopic* ($\varphi_0 \sim \varphi_1$) if there is a continuous map $t \mapsto \varphi_t, 0 \le t \le 1$, which connects them in the metrizable space of homeomorphisms from Y_S to Y_T. Let $\mathrm{Is}(S,T)$ denote the set of isotopy classes of flow equivalences from S to T. We let $\mathrm{Is}(S,S) = \mathrm{Is}(S)$ and call this the mapping class group of the oriented space Y_S.

The isotopy futures group of Y_S.

Given S, $x \in S$ and $n \in \mathbb{Z}$, define

$$r(x,n) = \{[(w,0)] \in Y_S\colon w \in S, w_i = x_i \text{ for } i \le n\}.$$

We call such a set a *ray* in Y_S. We say two sets E, E' in Y_S are *isotopic* ($E \sim E'$) if there is a homeomorphism $\varphi\colon Y_S \to Y_S$ such that $\varphi(E) = E'$ and φ is isotopic to the identity. An *isotopy ray* is a set isotopic to a ray. A *beam* is a disjoint union of finitely many rays. An *isotopy beam* is a set isotopic to a beam. Let $\mathcal{B} = \mathcal{B}(S)$ denote the set of isotopy beams of Y_S.

We define $\mathcal{F}(S)$, the *isotopy futures group of* S, to be $\mathbb{Z}\mathcal{B}/K$, where $\mathbb{Z}\mathcal{B}$ is the free abelian group with generating set $\mathcal{B} = \mathcal{B}(S)$, and $K = K(S)$ is

the subgroup of $\mathbb{Z}\mathcal{B}$ generated by all elements of the following forms:

$$b - b', \qquad \text{if } \{b, b'\} \subset \mathcal{B} \text{ and } b \sim b', \tag{7.1}$$

$$b - \sum_{j=1}^{k} b_j, \qquad \text{if } \{b, b_1, \dots, b_k\} \subset \mathcal{B},\ k \in \mathbb{N}, \text{ and } b = \dot{\cup}_j b_j. \tag{7.2}$$

For S a subshift, let $\mathcal{P}_n(S)$ denote the partition of S into clopen sets of the form $C(x,n) = \{w \in S\colon w_i = x_i \text{ if } |i| \le n\}$.

Lemma 7.3. *Suppose S, T are subshifts and $\varphi\colon Y_S \to Y_T$ is an orientation preserving homeomorphism and $b \in \mathcal{B}(S)$. Then $\varphi(b) \in \mathcal{B}(T)$.*

Proof. Exploiting the zero dimensionality of S as in [**PS**], after postcomposing φ with a suitable map isotopic to the identity we may assume that there is a positive integer M such that for any C in $\mathcal{P}_M(S)$ there is a constant $h = h_C$ and a homeomorphism $f = f_C$ from C to a clopen subset D of T such that $\varphi([(x,0)]) = [(f(x), h)]$, for all x in C.

Because φ respects disjoint union and pushes $\mathrm{Is}(S)$ forward to $\mathrm{Is}(T)$ (by the rule $[h] \mapsto [\varphi h \varphi^{-1}]$), it suffices to consider the case that b is a ray $r(x,n)$ with $n \ge M$. Let $C' = \{w \in S\colon w_i = x_i \text{ if } i \le n\} \subset C \in \mathcal{P}_M(S)$, with $h = h_C$ and $f = f_C$. Choose $k \in \mathbb{N}$ such that for all x in C, the sequence $x(-\infty, n]$ determines $(fx)(-\infty, n-k]$ and the sequence $(fx)(-\infty, n+k]$ determines $x(-\infty, n]$. So, if $w \in f(C')$, then $\{z \in T\colon z_i = w_i \text{ if } i \le n+k\} \subset f(C')$. Let $\mathcal{W}$ be the (finite) set of words $\{w[n-k, n+k]\colon w \in f(C')\}$. Then $\varphi(b) = \varphi(r(x,n)) = \cup_{W \in \mathcal{W}}\{[(z,h)]\colon z(-\infty, n-k-1] = (fx)(-\infty, n-k-1]$ and $z[n-k, n+k] = W\}$, so $\varphi(b)$ is an isotopy beam. □

The following proposition follows easily from the lemma.

Proposition 7.4. *Suppose S and T are subshifts and $\varphi\colon Y_S \to Y_T$ is an orientation preserving homeomorphism. Then the mapping of isotopy beams $b \mapsto \varphi(b)$ induces an isomorphism $\varphi_*\colon \mathcal{F}(S) \to \mathcal{F}(T)$.*

Let $\mathrm{Iso}(\mathcal{F}(S), \mathcal{F}(T))$ denote the set of group isomorphisms from $\mathcal{F}(S)$ to $\mathcal{F}(T)$. Let $\mathrm{Aut}(\mathcal{F}(S)) = \mathrm{Iso}(\mathcal{F}(S), \mathcal{F}(S))$. The next proposition is now obvious.

Proposition 7.5. *The rule $\varphi \mapsto \varphi_*$ induces a group homomorphism $\rho : \mathrm{Is}(S) \to \mathrm{Aut}(\mathcal{F}(S))$.*

Remark 7.6. The construction of $\mathcal{F}_S$ is one of several variations on the dimension group construction introduced by Krieger [**Kr1**, **Kr2**]; our construction was influenced also by [**LM**] and [**BFF**]. The construction of $\mathcal{F}_S$ is a flow equivalence analogue of Krieger's construction of a dimension group from a subshift S. The map $\rho\colon \mathrm{Is}(Y_S) \to \mathrm{Aut}(\mathcal{F}_S)$ is the analogue for flow equivalence of the dimension representation of the automorphism group of a subshift.

The isomorphism $\beta\colon \mathcal{F}(\sigma_A) \to \mathbf{cok}(I-A)$.

Suppose A is a matrix in $\mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$. Let $\mathcal{I}$ denote the index set of the rows and columns of A. Let $\mathbb{Z}^{\mathcal{I}}$ be the group of (infinite) row vectors indexed by $\mathcal{I}$, with all but finitely many entries zero. For a symbol/edge x_n of σ_A, let $\tau(x_n)$ denote the terminal vertex of the edge x_n (so, $\tau(x_n) \in \mathcal{I}$).

The group $\mathrm{cok}(I-A)$ is the cokernel of the map $\mathbb{Z}^{\mathcal{I}} \to \mathbb{Z}^{\mathcal{I}}$ given by $v \mapsto v(I-A)$ (i.e., $\mathrm{cok}(I-A) = \mathbb{Z}^{\mathcal{I}}/\mathrm{image}(I-A)$). Given a ray $r = r(x,n)$ with $i = \tau(x_n)$, let e_i be the ith canonical basis vector in $\mathbb{Z}^{\mathcal{I}}$, and define $\beta(r) = [e_i] \in \mathrm{cok}(I-A)$.

First note, given $k \in \mathbb{Z}$ and a ray $r = r(x,n)$, if we set r' equal to $\{[(w,k)]\colon [(w,0)] \in r(x,n)\}$, then r' is again a ray,

$$r' = r(\sigma^k x, n-k) \quad \text{and} \quad \beta(r') = \beta(r). \tag{7.7}$$

Here the equality of sets follow from the manipulations

$$\begin{aligned}\{[(w,k)]\colon [(w,0)] \in r(x,n)\} &= \{[(\sigma^k w,0)]\colon w(-\infty,n] = x(-\infty,n]\}\\ &= \{[(z,0)]\colon z(-\infty,n-k] = x(-\infty,n]\}\\ &= \{[(z,0)]\colon z(-\infty,n-k] = (\sigma^k x)(-\infty,n-k]\}\end{aligned}$$

and then $\beta(r) = \beta(r')$ because the edges x_n and $(\sigma^k x)_{n-k}$ are equal.

Next, given $x \in \sigma_A$, $n \in \mathbb{Z}$ and $k \in \mathbb{N}$, for each σ_A-word $W = W_1 \cdots W_k$ which can follow x_n, choose a point $y = y_W$ such that $y(-\infty,n] = x(-\infty,n]$ and $y[n+1, n+k] = W$. Then the equality

$$\beta(r(x,n]) = \sum_W \beta(r(y_W, n+k)) \in \mathrm{cok}(I-A) \tag{7.8}$$

follows for $k=1$ by direct computation and for $k>1$ by induction.

Given a beam b which is a disjoint union of finitely many rays $r(x^{(i)}, n^{(i)})$, we now define

$$\beta(b) = \sum_i \beta(r(x^{(i)}, n^{(i)})).$$

(We will use the symbol β for various maps derived from the map β on rays.) To see that this definition is independent of the particular choice of rays, suppose b is also the union of rays $r(w^{(j)}, m^{(j)})$. Choose $M \geq \max_{i,j}\{n^{(i)}, m^{(j)}\}$. Then b is the disjoint union of rays $r(z^{(k)}, M)$, each of the $r(x^{(i)}, n^{(i)})$ and $r(w^{(j)}, m^{(j)})$ is a union of some of the rays $r(z^{(k)}, M)$, and by (7.8) we have

$$\sum_i \beta(r(x^{(i)}, n^{(i)})) = \sum_k \beta(r(z^{(k)}, M)) = \sum_j \beta(r(w^{(j)}, m^{(j)})).$$

Therefore $\beta(b)$ is well-defined.

We will write Y_A for the mapping torus of σ_A.

Lemma 7.9. *If b and b' are beams in Y_A such that $b \sim b'$, then $\beta(b) = \beta(b')$.*

Proof. Without loss of generality, choose $M \in \mathbb{N}$ and a finite set E of Y_A such that b is the disjoint union of rays $r(x, M)$, $x \in E$. Let $\{\varphi_t\}$ be an isotopy such that $\varphi_0 = \mathrm{id}$ and $\varphi_1(b) = b'$. Because $\varphi_1 \sim \mathrm{id}$, there is a continuous function $k(x)$ such that for all $[(x,0)]$ in b, $\varphi_1 \colon [(x,0)] \mapsto [(x, k(x))]$. Because $\varphi(b)$ is a beam, the function k is integer valued. Possibly after increasing our choice of M, we may assume that k is constant on each ray $r(x, M)$. By (7.7), φ_1 takes each ray $r = r(x, M)$ onto a ray r' such that $\beta(r) = \beta(r')$, b' is the disjoint union of these rays r', and

$$\beta(b) = \sum_r \beta(r) = \sum_{r'} \beta(r') = \beta(b').$$

□

An isotopy beam b is isotopic to some beam b'. Define $\beta(b) = \beta(b')$. It follows from the lemma that $\beta(b)$ does not depend on the choice of b'. Likewise we have a well-defined homomorphism of groups

$$\begin{aligned} \beta \colon \quad \mathbb{Z}\mathcal{B} &\to \operatorname{cok}(I - A), \\ \sum n_i b_i &\mapsto \sum n_i \beta(b_i). \end{aligned} \tag{7.10}$$

Proposition 7.11. *The kernel of the map β in* (7.10) *is the subgroup K with generators* (7.1, 7.2)*. So, there is an induced isomorphism of groups*

$$\beta_A \colon \mathcal{F}(\sigma_A) \to \operatorname{cok}(I - A).$$

Proof. First we show $K \subset \operatorname{Ker}\beta$ by showing that β vanishes on the generators of K. For (7.1), suppose $b \sim b'$; then $\beta(b - b') = 0$ by Lemma 7.9. For (7.2), suppose b is an isotopy beam and b is the disjoint union of finitely many isotopy beams b_i. Without loss of generality, suppose b is a beam. The b_i are a finite collection of disjoint compact sets, so for sufficiently large m, for any C in $\mathcal{P}_m(S)$ such that $b \cap C \neq \emptyset$, the set $(b \cap C)$ will be contained in one of the b_i. If m is large enough, then $b \cap C$ if nonempty will be a ray. Thus, taking sums over C in $\mathcal{P}_m$, and for notational convenience defining β to be zero on the empty set, we get

$$\begin{aligned} \beta(b) &= \sum_C \beta(C \cap b) \\ &= \sum_j \sum_C \beta(C \cap b_j) = \sum_j \beta(b_j). \end{aligned}$$

Now we show $\ker\beta \subset K$. Suppose $g = \sum n_j b_j \in \ker\beta$. There exists $M \geq 0$ such that for each j, there are rays $r(x^{(jk)}, M)$ such that $b_j - \sum_k r(x^{(jk)}, M) \in K$. so $g = \sum_{jk} n_j r(x^{(jk)}, M) \pmod K$. For any x, $\beta(r(\sigma^M x, 0)) = \beta(r(x, M))$; also, $r(x,0) - r(x',0) \in K$ if x_0 and x'_0 have the same terminal vertex i. So, we may choose for each i an element $x^{(i)}$

such that $(x^{(i)})_0$ has terminal vertex i, for each $x^{(jk)}$ replace $x^{(jk)}$ with the appropriate $x^{(i)}$, and after reindexing obtain integers m_i such that

$$g = \sum_i m_i r(x^{(i)}, 0) \text{ (mod } K).$$

Because $\beta(g) = 0$ and $K \subset \ker\beta$, there is an integral row vector w such that $\sum_i m_i e_i = w(I - A)$, and therefore

$$g = \sum_i w_i \left(\left[r\left(x^{(i)}, 0\right) \right] - \sum_j A_{ij} \left[r\left(x^{(j)}, 0\right) \right] \right) \text{ (mod } K).$$

For each i, we have $r(\sigma_A x^{(i)}, -1) - r(x^{(i)}, 0) \in K$, and also

$$r(\sigma_A x^{(i)}, -1) - \sum_j A_{ij} r(x^{(j)}, 0) \in K.$$

It follows that $g = 0$ (mod K). □

In our definition of $\mathcal{F}(S)$ and $\text{cok}(I - A)$ we used sets $r(x, n)$ and row vectors. (So, $\text{cok}(I - A) = \text{rowcok}(I - A)$.) In the same way, using sets

$$r^+(x, n) := \{[(w, 0)] \in Y_S \colon w \in S \text{ and } w_i = x_i \text{ for } i \geq n\}$$

we may define the pasts group $\mathcal{P}(S)$; and using column vectors, we obtain an isomorphism $\mathcal{P}(\sigma_A) \to \text{colcok}(I - A)$. For a flow equivalence $\varphi \colon Y_A \to Y_B$, the isomorphism $\varphi_* \colon \mathcal{F}(\sigma_A) \to \mathcal{F}(\sigma_B)$ given by Proposition 7.4 induces the isomorphism

$$\varphi_*^{\text{row}} := (\beta_B)\varphi_*(\beta_A)^{-1} \colon \text{ rowcok}(I - A) \to \text{rowcok}(I - B).$$

Likewise, the action of φ on $\mathcal{P}(\sigma_A)$ induces an isomorphism

$$\varphi_*^{\text{col}} \colon \text{colcok}(I - A) \to \text{colcok}(I - B).$$

The action of $\text{Is}(\sigma_A)$ on $\text{cok}(I - A)$.

For a flow equivalence $\varphi \colon Y_A \to Y_A$, we have group homomorphisms

$$\begin{array}{ccc} \text{Is}(\sigma_A) \to \text{Aut}(\text{rowcok}(I - A)) & \text{and} & \text{Is}(\sigma_A) \to \text{Aut}(\text{colcok}(I - A)) \\ \varphi \mapsto \varphi_*^{\text{row}} & & \varphi \mapsto \varphi_*^{\text{col}}. \end{array}$$

As described in Subsection 2.2, if (U, V) is a basic positive $\text{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalence from $(I - A)$ to $(I - B) = U(I - A)V$, and B plays the role of A' in Subsection 2.2, then there is an associated map γ from σ_A to σ_B, and it is easy to see that this map is the restriction (to the cross section σ_A) of an orientation preserving homeomorphism $Y_A \to Y_B$. More generally, if (U, V) is the composition of basic positive $\text{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalences (U_i, V_i), and φ is the corresponding composition of the flow equivalences associated to the (U_i, V_i), then we will write $\varphi = \varphi_{(U,V)}$. This is an abuse of notation in that we are not claiming that (U, V) determines φ (the map φ may depend on

the particular factorization of (U,V)); we are only indicating that φ arises via some factorization of (U,V).

Proposition 7.12. *Suppose $A \in \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$. Suppose $(U,V)\colon (I-A) \to (I-B)$ is a positive $\mathrm{SL}_{\mathcal{P}}$ equivalence, and $\varphi_{(U,V)}$ is an associated flow equivalence.*

Then the induced map $\varphi^{\mathrm{row}}_*\colon \mathrm{rowcok}(I-A) \to \mathrm{rowcok}(I-B)$ is given by the rule $[w] \mapsto [wV]$, and the induced map $\varphi^{\mathrm{col}}_*\colon \mathrm{colcok}(I-A) \to \mathrm{colcok}(I-B)$ is given by the rule $[w] \mapsto [Uw]$.

Proof. We will check the proposition in the case that $(U,V) = (E,I)$ and E is a basic elementary matrix with unique offdiagonal entry $E(i,j) = 1$. The argument for (I,E) is similar and then the proposition follows by composition. For concreteness, suppose $E(1,2) = 1$ (in other entries $E = I$). Let $\varphi = \varphi_{(E,I)}$ be defined via the map γ and edge e described in Subsection 2.2.

Suppose $x \in \sigma_A$ and x_{-1} has terminal vertex i. Then the edge $(\gamma x)_{-1}$ has terminal vertex i and φ maps $r(x,-1)$ onto $r(\gamma x,-1)$. It follows that the diagram

$$\begin{array}{ccc} \mathcal{F}(\sigma_A) & \xrightarrow{\varphi_*} & \mathcal{F}(\sigma_B) \\ {\scriptstyle \beta_A}\downarrow & & {\scriptstyle \beta_B}\downarrow \\ \mathrm{rowcok}(I-A) & \xrightarrow{\mathrm{Id}} & \mathrm{rowcok}(I-B) \end{array}$$

commutes; that is, $\varphi^{\mathrm{row}}_* = \mathrm{Id}$.

If $x \in \sigma_B$ and x_0 has initial vertex not equal to 2, then φ maps $r^+(x,0)$ onto $r^+(\gamma x,0)$. Thus the map $(\varphi^{\mathrm{col}}_*)^{-1}$ sends $[e_i]$ in $\mathrm{colcok}(I-A)$ to $[e_i]$ in $\mathrm{colcok}(I-B)$ whenever $i \neq 2$. If the initial vertex of x_0 is 2, then φ^{-1} sends $r^+(x,0)$ to the set of all points $(w,0)$ in $r^+(\gamma^{-1}x,0)$ such that $w_{-1} \neq e$. Consequently, if y is a point in σ_A such that $y_i = (\gamma x)_i$ if $i \geq 0$ and $y_{-1} = e$, then

$$\varphi^{-1}\big(r^+(x,0)\big) = r^+(\gamma^{-1}x,0) \setminus r^+(y,-1).$$

We also have

$$\begin{aligned} \beta^{\mathrm{col}}_B\colon &\quad r^+(x,0) \mapsto [e_2] \in \mathrm{colcok}(I-B), \\ \beta^{\mathrm{col}}_A\colon &\ r^+(\gamma^{-1}x,0) \mapsto [e_2] \in \mathrm{colcok}(I-A), \\ \beta^{\mathrm{col}}_A\colon &\quad r^+(y,-1) \mapsto [e_1] \in \mathrm{colcok}(I-A). \end{aligned}$$

Therefore $(\varphi^{\mathrm{col}}_*)^{-1}\colon [e_2] \mapsto [e_2]-[e_1]$, hence for all integral column vectors v we have $(\varphi^{\mathrm{col}}_*)^{-1}\colon [v] \mapsto [E^{-1}v]$ as required. □

Theorem 7.13. *Suppose $A \in \mathfrak{M}^{\circ}_{+}(\mathbb{Z})$ and the mapping torus of σ_A is not a circle. Then the induced map $\mathrm{Is}(Y_A) \to \mathrm{Aut}(\mathrm{cok}(I-A))$ is surjective.*

Remark 7.14. Of course, the theorem is true for colcok as well as for rowcok. In the case that the mapping torus of σ_A is a circle (i.e., A has a unique irreducible component, and this component is a permutation matrix), any orientation preserving homeomorphism from Y_A to Y_A is isotopic to the identity, but $\mathrm{cok}(I-A) \cong \mathbb{Z}$ and $\mathrm{Aut}(\mathrm{cok}(I-A)) \cong \mathbb{Z}/2$, so the map $\mathrm{Is}(Y_A) \to \mathrm{Aut}(\mathrm{cok}(I-A))$ is not surjective. Theorem 7.13 says that apart from this case, every automorphism of the isotopy futures group of an irreducible shift of finite type is induced by a flow equivalence.

Proof of Theorem 7.13. It is proved in [**BH**] that any automorphism of $\mathrm{rowcok}(I-A)$ or $\mathrm{colcok}(I-A)$ is induced by an $\mathrm{SL}(\mathbb{Z})$ equivalence (by the rules described in the statement of Proposition 7.12). By the Factorization Theorem 3.3, such an equivalence is a positive equivalence. By Proposition 7.12, a flow equivalence associated to this positive equivalence has the desired action on the cokernel group. □

From the view of symbolic dynamics, Theorem 7.13 stands in contrast to the Kim-Roush-Wagoner result [**KRW1**] that the dimension representation of a mixing shift of finite type is not in general surjective. (The contrast is meaningful because the invariants are related by "setting t equal to 1" [**B1**].)

When $A \in \mathfrak{M}_+^\circ(\mathbb{Z})$ (i.e., A is essentially irreducible) and σ_A is not a circle, the flow equivalence class of σ_A is given by the $\mathrm{SL}(\mathbb{Z})$ equivalence class of $I-A$, for which $\det(I-A)$ and $\mathrm{cok}(I-A)$ give complete invariants. When $\mathcal{P}$ is nontrivial and $A \in \mathfrak{M}_{\mathcal{P},+}^\circ(\mathbb{Z})$ (i.e., the SFT σ_A is reducible), the flow equivalence class of A (modulo a permutation of $\mathcal{P}$) is given by its positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence class, and the complete algebraic invariants (introduced by Huang) are more subtle, involving the "K-web" of the matrix $I-A$, denoted $K(I-A)$. The K-web is a diagram of exact sequences of certain kernel and cokernel groups of submatrices of $I-A$. The K-web invariants are completely analyzed in [**BH**], which also characterizes the automorphisms of $K(A)$ which can be induced by an $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence. We believe that the type of analysis carried out to describe the action of $\mathrm{Is}(\sigma_A)$ on $\mathrm{cok}(I-A)$ in the irreducible case can be extended to describe the possible actions of $\mathrm{Is}(\sigma_A)$ on the more complicated algebraic structure of the K-web which classifies in the reducible case. Specifically, we expect that the following program can be carried out. Together with [**BH**], this program would give a complete description of the possible actions of $\mathrm{Is}(\sigma_A)$ on the K-web.

Program 7.15. For A, B in $\mathfrak{M}_{\mathcal{P}}(\mathbb{Z}_+)$, we conjecture the following.

(1) The K-web data for $I-A$ can be described in terms of isotopy beams of subsystems of Y_A, and the map on isotopy beams by an orientation preserving homeomorphism $\varphi\colon Y_A \to Y_B$ induces an isomorphism $K(A) \to K(B)$.

(2) For a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence (U, V) from A to B, the isomorphism $K(A) \to K(B)$ induced by $\varphi_{(U,V)}$ is the natural isomorphism induced by (U, V) as described in [**BH**].
(3) If φ is an orientation preserving homeomorphism from Y_A to Y_B, then there is a positive $\mathrm{SL}_{\mathcal{P}}(\mathbb{Z})$ equivalence (U, V) such that φ is isotopic to $\varphi_{(U,V)}$.

The most fundamental of the three steps above is the last one, and a version of this has already been carried out in the irreducible case (i.e., $\mathcal{P} = \{1\}$) by Badoian [**Ba1**], as we discuss below.

The work of Badoian.

We'll describe some of the work [**Ba1**] of Leslie Badoian, which gives alternate proofs of some of our results. The work [**Ba1**] is too extensive for a full summary here; roughly speaking, Badoian carries out for irreducible shifts of finite type a flow equivalence version of the strong shift equivalence theory Wagoner [**W1**] built on the foundation laid by Williams [**Wi**].

Badoian builds an infinite oriented CW complex, denoted FK. A zero-cell for FK is an equivalence class of infinite, essentially irreducible, finitely supported zero-one matrices, where two matrices are equivalent iff their unique maximal irreducible principal submatrices are equal. A one-cell $[A] \to [B]$ corresponds to an elementary equivalence $(I - B) = U(I - A)V$ satisfying certain conditions. Two-cells are also defined, by certain matrix relations. The two main results of [**Ba1**] are the following:

- *Classification Theorem.* σ_A and σ_B are flow equivalent if and only if A and B lie in the same connected component of FK.
- *Flow Equivalence Theorem.* $\pi_1(FK_A) \cong \mathrm{Is}(\sigma_A)$. (I.e., a path along one-cells gives rise to a flow equivalence, and two paths give rise to isotopic flow equivalences if and only if the paths are homotopic in FK.)

The elementary equivalences of [**Ba1**] are not the same as our elementary positive equivalences, but Badoian has found short arguments [**Ba2**] which show directly that that her elementary equivalences and ours generate the same set of flow equivalences up to isotopy. With this fact and some technical remarks, the results of Section 6 for irreducible shifts of finite type follow directly from Badoian's Classification Theorem (which in turn rests on Parry-Sullivan [**PS**] and Williams [**Wi**], as does our Section 6).

The Flow Equivalence Theorem gives an alternate route in the irreducible case to the representation $\mathrm{Is}(\sigma_A, \sigma_B) \to \mathrm{Iso}(\mathrm{cok}(I - A), \mathrm{cok}(I - B))$: We could take the natural definition along an edge (given by the associated flow equivalence), compose along paths of edges, and consult the definition of two-cells in FK to verify that the definition only depends on the homotopy class of the path of edges. All of this is parallel to the development of the dimension representation in Wagoner's strong shift equivalence theory [**W1**].

We have not relied in proofs on citation of [**Ba1**], for a few reasons. Although there should be no fundamental problem with extending Badoian's approach to reducible shifts of finite type, the results in [**Ba1**] are only for irreducible shifts of finite type. We also wanted self-contained and reasonably brief arguments. (The long work [**Ba1**] deals with a fundamental difficult problem which we avoid: We do not try to understand when two paths give rise to the same flow equivalence up to isotopy.) Finally, although the CW complex approach has rather spectacularly proved its worth [**KR2**, **KRW1**, **W1**], the Krieger-style construction remains important, and its more earthy definition (by actions on sets) makes sense directly for general subshifts. Matsumoto [**Ma**] has a far reaching extension of Williams' theory to general subshifts, and this offers hope for some analogue of Wagoner's strong shift equivalence theory for general subshifts; but there is no such theory yet.

Appendix A. Reduction to nondegenerate form.

This appendix is devoted to the proof of Proposition 4.5.

We will prove Proposition 4.5 by composition in a larger commuting diagram (to be assembled in three stages):

$$\begin{array}{ccccccc}
(I-A) & \longrightarrow & (I-\overline{A_1}) & \longrightarrow & (I-\overline{A_2}) & \longrightarrow & (I-\overline{A}) \\
{\scriptstyle (U,V)}\downarrow & & {\scriptstyle (U_1,V_1)}\downarrow & & {\scriptstyle (U_2,V_2)}\downarrow & & \downarrow{\scriptstyle (\overline{U},\overline{V})} \\
(I-A') & \longrightarrow & (I-\overline{A_1'}) & \longrightarrow & (I-\overline{A_2'}) & \longrightarrow & (I-\overline{A'})\,.
\end{array}$$

The horizontal arrows will be positive equivalences and the vertical equivalences to the right of (U,V) will be defined from them by composition (then the diagram will commute). Stage I will produce the left square with $\overline{A_1}$ and $\overline{A_1'}$ satisfying Conditions (2), (3) and (4) of Proposition 4.5. Stage II will produce the middle square, with $(U_2)_{ii} = (V_2)_{ii} = \mathrm{Id}$ for $i \in \mathcal{C}$, and with $\overline{A_2}$ and $\overline{A_2'}$ still satisfying Conditions (2), (3) and (4) of Proposition 4.5. Stage III will produce the right square to finish the proof. The individual stages will follow from several lemmas.

Lemma A.1. *Suppose $A \in \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n},\mathbb{Z})$. Then there is a positive $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ equivalence in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n},\mathbb{Z})$ from $I-A$ to a matrix $I-C$ such that for all $i \in \mathcal{C}_A$ the following hold:*

1. *The block C_{ii} has its upper left corner entry equal to 1, and C_{ii} has no other nonzero entry.*
2. *Let (ℓ,ℓ) be the entry of C which is the upper left corner of C_{ii}. Then for $j \neq i$, every row of a block C_{ij} other than row ℓ is zero, and every column of a block C_{ji} other than column ℓ is equal to zero.*

Proof. Suppose $i \in \mathcal{C}_A$. Let $i_1, \dots, i_k$ be nonrepeated indices such that $A_{ii}(i_t, i_{t+1}) = 1$, $1 \le t < k$, and $A_{ii}(i_k, i_1) = 1$.

Cycle-shortening construction. Suppose $k > 1$. Let $A = A^{(0)}$. For $1 \le j < k$, define $A^{(j)}$ by the equation $I - A^{(j)} = E_j(I - A^{(j-1)})$, where E_j denotes the basic elementary matrix which acts to add row i_{k-j+1} to row i_{k-j}. Each $A^{(j)}$ is nonnegative. Then add the columns $i_2, \dots, i_k$ of $A^{(k)}$ to column i_1 of $A^{(k)}$. By Lemma 2.5, each step in this process gives a positive equivalence in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, and in the last matrix A', the block A'_{ii} has as its unique cycle the 1-cycle (i_1). Below is an example of the process, with $(i_1, i_2, i_3, i_4) = (1, 2, 3, 4)$, viewed in the principal submatrices on indices $1, 2, 3, 4$:

$$\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \to$$
$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix} \to \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Now without loss of generality, we suppose $k = 1$ with $A(i_1, i_1) = 1$. Because (i_1) is the unique A_{ii} cycle, if A_{ii} is nonzero at any entry other than (i_1, i_1), then it is nonzero at some entry (j, l) such that row l of A is zero or column j of A is zero. In the former case, let E be the elementary matrix which acts from the left to add row l to row j, then $(E, I)\colon (I - A) \to E(I - A) = (I - A'')$ is a positive equivalence in which $A'' = A$ except that $A'(j, l) = 0$. The latter case is treated similarly, by adding column j to column l. Iterating, we produce a positive equivalence in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$ from $I - A$ to a matrix $I - A'$ such that $A'(i_1, i_1)$ is the only nonzero entry of A'_{ii}.

Next, given i in $\mathcal{C}_{A'}$ with $A'(i_1, i_1) = 1$, we may for each $j \prec i$ add rows of $I - A'$ through the ii block to rows through the ji block (never adding row i_1) until every column of the block $(I - A')_{ji}$ except column i_1 is zero. We do this for all the cycle components i, for i in decreasing order, so that no block zeroed out for some i is made nonzero by subsequent operations. Then similarly, taking i in $\mathcal{C}_A$ in increasing order, we add columns through the ii block to columns through the ji blocks with $i \prec j$, to end with a matrix C' which satisfies the statement of the lemma (with C' in place of C), except that the distinguished indices i_1 might not be the corner indices ℓ.

So, suppose i is a cycle component for which $\ell \ne i_1$. We apply four basic positive equivalences to give $(I - C') \to (I - C'')$, as viewed below

in principal submatrices on indices $\{i_1, \ell, k\}$ (where k is any index not in $\{i_1, \ell\}$). (In the very special case that C' is 2×2, there can be no third index k and these principal submatrices should be restricted to indices $\{i_1, \ell\}$.) For concreteness we use $(\ell, i_1, k) = (1, 2, 3)$:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & x \\ 0 & y & z \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & x \\ 0 & y & z \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & x \\ 0 & y & z \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & x \\ 0 & y & z \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & x \\ 0 & y & z \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & x \\ 0 & y & z \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & x \\ y & y & z \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & x \\ y & y & z \end{pmatrix} = \begin{pmatrix} 0 & 0 & x \\ 0 & 1 & x \\ y & y & z \end{pmatrix}.$$

If above for any k we have $x \neq 0$, then $x < 0$ and the (i_1, k) entry lies in an ij block with $i \prec j$; then $y = 0$ and it is a positive equivalence to add column i_1 of C'' $|x|$ times to column k. Doing this as needed, and dealing similarly with nonzero entries y using rows in place of columns, we produce another version of C'' which enjoys the additional property that $i_1 = \ell$ for the cycle component i. Then we repeat until $i_1 = \ell$ for every cycle component i. The resulting matrix C satisfies the statement of the lemma. □

Lemma A.2. *Suppose* $A \in \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbf{n}, \mathbb{Z})$, *and* $\mathbf{n} = (n_1, \ldots, n_N)$ *is a vector with positive integer entries. Then there is a positive* $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ *equivalence in* $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$ *from* $(I - A)$ *to a matrix with Properties* (2), (3) *and* (4) *of Proposition* 4.5.

Proof of Lemma A.2. We will describe a sequence of row and column operations (corresponding, by repeated tacit appeal to Lemma 2.5, to positive $\mathrm{SL}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$ equivalences in $I - \mathfrak{M}^{\circ}_{\mathcal{P},+}(\mathbb{Z})$) which put the matrix $I - A$ into the required form. To simplify notation, rather than renaming $I - A$ after an equivalence, we will discuss changing properties of $I - A$. We begin with a matrix A with the properties stated (for C) in Lemma A.1, i.e., A satisfies Properties (3) and (4) of Proposition 4.5.

Our first goal will be, given $t \in \mathcal{P}$ which is not a cycle component, to arrange that the block $(I - A)_{tt}$ be strictly negative. Recall $\mathcal{I}_t$ denotes the index set for rows/columns of A_{tt}. Let $\mathcal{S}$ denote the index set for the unique maximal irreducible submatrix of A_{tt}, let $\mathcal{S}'$ denote the complement of $\mathcal{S}$ in $\mathcal{I}_t$, and e.g., let $A\{\mathcal{S}\}$ denote the principal submatrix of A on index set $\mathcal{S}$. We will arrange (in order) the following properties (after each stage keeping the

properties achieved at earlier stages, and not changing entries in any block A_{ss} with $s \neq t$, and not losing Properties (3) and (4) of Proposition 4.5).

(1) $\exists i \in \mathcal{S}$ such that $(I - A)(i, i) \leq 0$.
(2) $\{i, j\} \cap \mathcal{S}' \neq \emptyset \implies (I - A)(i, j) = \delta_{ij}$.
(3) If $\mathcal{S}' \neq \emptyset$, then $|\mathcal{S}| > 1$.
(4) $\mathcal{S}' = \emptyset$.
(5) The block $(I - A)_{tt}$ is strictly negative.

(1) If necessary achieve this with the initial row operations of the cycle-shortening construction of the Lemma A.1.

(2) First suppose this condition does not hold for some $\{i, j\} \subset \mathcal{I}_t$. Then pick some $i \in \mathcal{S}'$ and $j \in \mathcal{I}_t$ such that $j \neq i$ and one of the following hold:

- $(I - A)(i, j) \neq 0$ and column i of A_{tt} is zero, or
- $(I - A)(j, i) \neq 0$ and row i of A_{tt} is zero.

In the former case, add column i of $(I - A)$ to other columns j where $(I - A)(i, j) < 0$, until $(I - A)(i, j) = \delta_{ij}$ for all $j \in \mathcal{I}_t$. In the latter case, similarly use row additions to achieve $(I - A)(j, i) = \delta_{ij}$ for all $j \in \mathcal{I}_t$. This procedure reduces the cardinality of the set of entries in $(I - A)_{tt}$ at which Condition (2) fails, and it may be repeated until Condition (2) holds for $\{i, j\} \subset \mathcal{I}_t$. We then add rows and columns indexed by $\mathcal{S}'$ to others as needed until (2) holds in general.

(3) Suppose (for concreteness) that $\mathcal{S} = \{1\}$ and $2 \in \mathcal{S}'$. Then we must have $A(1, 1) = k > 1$ (since t is not a cycle component). Now, subtract row 2 of $(I - A)$ from row 1; then subtract column 2 from column 1. The effect of these moves is to enlarge $\mathcal{S} = \{1\}$ to $\mathcal{S} = \{1, 2\}$. The moves are summarized below in principal submatrices on indices $\{1, 2, 3\}$, where 3 is an arbitrary additional index:

$$\begin{pmatrix} 1-k & 0 & w \\ 0 & 1 & 0 \\ x & 0 & z \end{pmatrix} \to \begin{pmatrix} 1-k & -1 & w \\ 0 & 1 & 0 \\ x & 0 & z \end{pmatrix} \to \begin{pmatrix} -(k-2) & -1 & w \\ -1 & 1 & 0 \\ x & 0 & z \end{pmatrix}.$$

(4) Suppose $\mathcal{S}' \neq \emptyset$. By (1) and (3), we may pick i_1, j_1 in $\mathcal{S}$ such that $i_1 \neq j_1$, $(I - A)(i_1, i_1) \leq 0$, and $(I - A)(i_1, j_1) \leq -1$. Add row i_1 of $(I - A)$ to row j_1, $(|\mathcal{S}'| + 1)$ times, producing $(I - A)(j_1, j_1) \leq -|\mathcal{S}'|$. For each j in $\mathcal{S}'$, subtract row j of $(I - A)$ from row j_1. Then subtract each $\mathcal{S}'$ column from column j_1. This produces A with $\mathcal{S}' = \emptyset$.

(5) With i_1, j_1 as in (4): Add row i_1 to row j_1 (now $(I - A)(j_1, j_1) < 0$); for each i in $\mathcal{S}$ with $i \neq j_1$, add column j_1 to column i (now row j_1 of $(I - A)$ is negative); and for each i in $\mathcal{S}$ with $i \neq j_1$, add row j_1 to row i. We now have $(I - A)_{tt}$ strictly negative as required.

After applying a positive equivalence, then, we may assume that $(I - A)_{ii} < 0$ for every noncycle component i. Consequently, if $i \prec j$, and i or j is not a cycle component, then for large n the block $(A^n)_{ij}$ is strictly positive.

We can then get a positive equivalence to $(I-A)$ whose block $(I-A)_{ij}$ is strictly negative, by adding columns through i to columns through j (if $i \notin \mathcal{C}_A$) or by adding rows through j to rows through i (if $j \notin \mathcal{C}_A$). Similarly, for every noncycle component j and cycle component i, with $\mathcal{C}_i^{\text{prim}} = \{\ell\}$, add a j-row to row ℓ if $i \prec j$, and add a j-column to column ℓ if $j \prec i$.

Note, if $i \prec j$ and $\{i,j\} \subset \mathcal{C}$, with say $\mathcal{C}_i^{\text{prim}} = \{\ell_i\}$ and $\mathcal{C}_j^{\text{prim}} = \{\ell_j\}$, then $A(\ell_i, \ell_j) > 0$, because the block A_{ij} is not the zero block (because $A \in \mathfrak{M}^o_{\mathcal{P},+}(\mathcal{C}, \mathbf{n}, \mathbb{Z})$) and the only possible nonzero entry is $A(\ell_i, \ell_j) > 0$. Finally, whenever $(I-A)_{ij} < 0$ with $\{i,j\} \subset \mathcal{C}_A$ and $i \prec k \prec j$ for some k in $\mathcal{P}$, pick k such that $i \prec k \prec j$, and add columns of $I-A$ through component k to columns through component j. The resulting matrix satisfies the statement of the lemma. □

Lemmas A.1 and A.2 finish the proof for Stage I. We now shift our focus to the form of the equivalence (U,V). The next lemma gives the proof for Stage II.

Lemma A.3. *Suppose* $(U,V)\colon (I-A) \to (I-A')$ *is an* $\mathrm{SL}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ *equivalence which is positive on cycle components, and* A, A' *satisfy Conditions* (2), (3) *and* (4) *of Proposition* 4.5. *Then there is a commuting diagram*

$$\begin{array}{ccc} (I-A) & \longrightarrow & (I-\widetilde{A}) \\ {\scriptstyle (U,V)}\downarrow & & \downarrow{\scriptstyle (\widetilde{U},\widetilde{V})} \\ (I-A') & \longrightarrow & (I-\widetilde{A'}) \end{array}$$

in which the horizontal arrows are positive equivalences; $\widetilde{A}$ *and* $\widetilde{A'}$ *still satisfy Conditions* (2), (3) *and* (4)*; and for each* $i \in \mathcal{C}$, $\widetilde{U}_{ii} = \widetilde{V}_{ii} = \mathrm{Id}$.

Proof. Suppose i is a cycle component for which $n_i > 1$ (otherwise there is nothing to prove). Then $(I-A)_{ii} = (I-A')_{ii} = Q$, where $Q = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$, in which I is $(n_i-1)\times(n_i-1)$. Considering blocks of $U_{ii}Q = QV_{ii}^{-1}$, we see U_{ii} and V_{ii} have the corresponding block forms $U_{ii} = \begin{pmatrix} a & 0 \\ x & Z \end{pmatrix}$ and $V_{ii} = \begin{pmatrix} b & y \\ 0 & Z^{-1} \end{pmatrix}$. The positive on cycle components assumption implies $a = 1$. Then $\det(U) = 1$ implies $\det(Z) = 1$. Then $\det(Z^{-1}) = 1 = \det V$ implies $b = 1$. So we have

$$U_{ii} = \begin{pmatrix} 1 & 0 \\ x & Z \end{pmatrix} \quad \text{and} \quad V_{ii} = \begin{pmatrix} 1 & y \\ 0 & Z^{-1} \end{pmatrix} \tag{A.4}$$

for some Z in $\mathrm{SL}(n_i - 1, \mathbb{Z})$.

Now suppose E is a basic elementary matrix with offdiagonal entry $E(j,k) = 1$, where j,k index rows of the ii block other than the first row. Then

$$(I-A) \xrightarrow{(I,E^{-1})} \cdot \xrightarrow{(E,I)} (I-A) \tag{A.5}$$

gives a factorization of $(E,E^{-1})\colon (I-A) \to (I-A)$ into basic positive equivalences. For example, if rows 1,2,3 run through Q and $Q(1,1) = 0$, then in the principal submatrix on indices 1,2,3 we could have $E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$, and (A.5) would become

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{(I,E^{-1})} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{(E,I)} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Now we can factor (U,V) as

$$(I-A) \xrightarrow{(I,E^{-1})} \cdot \xrightarrow{(E,I)} (I-A) \xrightarrow{(UE^{-1},EV)} (I-A'). \tag{A.6}$$

Because Z is a composition of elementary matrices, and Conditions (2), (3) and (4) are not disturbed by this move, we can repeat this move to obtain a positive equivalence $(G,G^{-1})\colon (I-A) \to (I-A)$ such that the (U,V) equals (G,G^{-1}) followed by (UG^{-1},GV) where $(UG^{-1})_{ii}$ and $(GV)_{ii}$ have the forms (A.4) with $Z = I$. After doing this as needed for every cycle component i, we can assume for each $i \in \mathcal{C}$ with $n_i > 1$ that we have the forms $U_{ii} = \begin{pmatrix} 1 & 0 \\ x^{(i)} & I \end{pmatrix}$ and $V_{ii} = \begin{pmatrix} 1 & y^{(i)} \\ 0 & I \end{pmatrix}$.

Let D and D' be the block diagonal matrices equal to Id except in cycle component diagonal blocks, where $D_{ii} = U_{ii}$ and $D'_{ii} = V_{ii}^{-1}$. We will produce matrices P,Q in $\mathcal{U}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ such that $(D,Q)\colon (I-A) \to D(I-A)Q$ and $(P,D')\colon (I-A') \to P(I-A')D'$ are positive equivalences, and the matrices $D(I-A)Q$ and $P(I-A')D'$ satisfy Conditions (2), (3) and (4). Then the lemma will follow by defining $(\widetilde{U},\widetilde{V})$ by requiring the following diagram to commute:

$$\begin{array}{ccc} (I-A) & \xrightarrow{(D,Q)} & D(I-A)Q \\ {\scriptstyle (U,V)}\downarrow & & \downarrow{\scriptstyle (\widetilde{U},\widetilde{V})} \\ (I-A') & \xrightarrow[(P,D')]{} & P(I-A')D' \end{array}.$$

We will prove the first claim, for (D,Q); the proof of the second claim is similar. Let $i_1 < i_2 < \cdots < i_k$ be the elements of $\mathcal{C}$. (Recall, $i \prec j \implies i < j$.) To begin, let $i = i_k$ and let $1,2,\ldots m$ index the rows through U_{ii}. For $2 \le j \le m$, let R_j be the elementary matrix which acts from the right

to subtract column j from column 1. Let $R = R_2 \cdots R_m$ and let $\mathbf{1}$ denote a vector with every entry equal to 1, then

$$((I-A)R)_{ii} = \begin{pmatrix} 0 & 0 \\ -\mathbf{1} & I \end{pmatrix}, \quad \text{and}$$
$$((I-A)R)_{rs} = (I-A)_{rs} \qquad \text{if } rs \neq ii,$$

and we get a positive equivalence

$$(I-A) \xrightarrow{(I,R_2)} \cdots \xrightarrow{(I,R_m)} (I-A)R.$$

Next, let D_k be a product of elementary matrices, $E = E_n \cdots E_1$, where E_t acts from the left to add ϵ_t ($\epsilon_t = 1$ or $\epsilon_t = -1$) times row 1 to row j_t, and $2 \leq j_t \leq m$. Consider the equivalence $(E_1, I)\colon (I-A)R \to E_1(I-A)R$. Notice $(E_1(I-A)R)_{ii} = ((I-A)R)_{ii}$. So, this equivalence (E_1, I) is positive unless $(E_1(I-A)R)(j,k) > 0$ for some columns p to the right of the ii block. Let F_1 be the product of basic elementary matrices $F_{1,t}$, $1 \leq t \leq T$ say, which act from the right to subtract column j_1 from such columns p enough times to guarantee (with $F_1 = F_{1,1} \cdots F_{1,T}$) that $(E_1(I-A)RF_1)(j_1, p) < 0$. Then

$$(I-A)R \xrightarrow{(I,F_{1,1})} \cdots \xrightarrow{(I,F_{1,T})} \cdot \xrightarrow{(E_1,I)} E_1(I-A)RF_1$$

gives a positive equivalence $(E_1, F_1)\colon (I-A)R \to E_1(I-A)RF_1$. Recursively, for $1 \leq t < m$, apply this procedure, to produce F_{t+1} giving a positive equivalence

$$E_t \cdots E_1(I-A)RF_1 \cdots F_t \xrightarrow{(I,F_{t+1})} \cdot \xrightarrow{(E_{t+1},I)} E_{t+1} \cdots E_1(I-A)RF_1 \cdots F_{t+1}.$$

Let $Q_k = F_1 \cdots F_m$: then we have a positive equivalence

$$(I-A) \xrightarrow{(D_k, RQ_k)} D_k(I-A)RQ_k \xrightarrow{(I,R^{-1})} D_k(I-A)RQ_kR^{-1}.$$

Because $RQ_k = Q_kR$, altogether we get

$$(I-A) \xrightarrow{(D_k,Q_k)} D_k(I-A)RQ_k.$$

Notice, $Q_k \in \mathcal{U}_{\mathcal{P}}(\mathbf{n}, \mathbb{Z})$. Moreover, if $j \in \mathcal{P}$ and $j < i_k$, then for any t the tj blocks of $(I-A)$ and $(D_k(I-A)_kQ_k)$ are equal.

Next, for the cycle components $i_{k-1}, \ldots, i_1$ (in that order) we repeat the procedure used above for (D_k, Q_k) to produce pairs $(D_{k-1}, Q_{k-1}), \ldots, (D_1, Q_1)$ with $D = D_kD_{k-1} \cdots D_1$ and $Q_{(-)} := Q_kQ_{k-1} \cdots Q_1$ giving a positive equivalence

$$(I-A) \xrightarrow{(D_k,Q_k)} \cdot \xrightarrow{(D_{k-1},Q_{k-1})} \cdots \xrightarrow{(D_1,Q_1)} D(I-A)Q_{(-)}.$$

To see that the (D_i, Q_i) define positive equivalences, note that for $i_s \neq i_k$, the column-subtracting moves we use to prepare the entries in a block i_sj to the right of the i_si_s block do not change the sign of entries outside the

$i_s j$ block (because we are subtracting columns through the $i_s i_s$ block with diagonal entry 1, and these columns have no other nonzero entry at this stage, because the earlier subtractions of columns through block $i_r i_r$ with $r \geq s$ do not affect the i_s block column).

For every cycle component i, the ii block of the matrix $D(I-A)Q_{(-)}$ equals Id. Suppose there exists (r,s) such that $r \in \mathcal{C}^{\text{sec}}$ and $(D(I-A)Q_{(-)})(r,s) < 0$; then choose such an (r,s) with r minimal, and add column r to column s. Because the (r,s) entry cannot lie in a diagonal block, this elementary positive equivalence is implemented by multiplication from the right by a matrix in $\mathcal{U}_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$. Repeat this move until a matrix is produced in which the (r,s) entry is zero whenever $r \in \mathcal{C}^{\text{sec}}$. Let the corresponding positive equivalence be denoted $(I, Q_{(+)})\colon D(I-A)Q_{(-)} \to D(I-A)Q_{(-)}Q_{(+)}$. The proof is finished by setting $Q = Q_{(-)}Q_{(+)}$. □

The next lemma gives the last ingredient, Stage III, for the proof of Proposition 4.5.

Lemma A.7. *Suppose U, V, A, A' satisfy the assumptions of Lemma* A.3 *and in addition assume that $U_{ii} = V_{ii} = \mathrm{Id}$ for every $i \in \mathcal{C}$. Then there is a commuting diagram of $SL_{\mathcal{P}}(\mathbf{n},\mathbb{Z})$ equivalences*

$$\begin{array}{ccc} (I-A) & \longrightarrow & (I-\overline{A}) \\ {\scriptstyle (U,V)}\downarrow & & \downarrow{\scriptstyle (\overline{U},\overline{V})} \\ (I-A') & \longrightarrow & (I-\overline{A'}) \end{array}$$

satisfying the conclusion of Proposition 4.5. (*Moreover, $\overline{A} = A$ and $\overline{A'} = A'$.*)

Proof. We will build a suitable commuting diagram

$$\begin{array}{ccccc} (I-A) & \xrightarrow{(E^{-1},H)} & (I-A) & \xrightarrow{(\overline{H},\overline{E}^{-1})} & (I-A) \\ {\scriptstyle (U,V)}\downarrow & & {\scriptstyle (U_3,V_3)}\downarrow & & \downarrow{\scriptstyle (\overline{U},\overline{V})} \\ (I-A') & \xrightarrow[(I,I)]{} & (I-A') & \xrightarrow[(I,I)]{} & (I-A') \end{array}$$

and then use $(\overline{H}E^{-1}, H\overline{E}^{-1})$ and (I,I) for the upper and lower horizontal arrows in the diagram required for the lemma. First we work on the left half of the diagram. We will choose E, H, U_3 satisfying:

(i) $U_3(i,j) = \delta_{ij}, \quad \forall i \in \mathcal{C}^{\text{sec}}$,
(ii) $(E^{-1}, H)\colon (I-A) \to (I-A)$ is a positive equivalence, and
(iii) $H^{-1}(i,j) = \delta_{ij}, \quad \forall i \notin \mathcal{C}^{\text{sec}}$.

Recall, $\mathcal{I}_s$ denotes the set of indices for rows/columns through A_{ss}. To choose E, let the entries (i,j) for which $i \in \mathcal{C}^{\text{sec}}$ and $U(i,j) \neq \delta_{ij}$ be listed as $(i_1,j_1),\ldots,(i_n,j_n)$, where $i_k \in \mathcal{I}_{s(k)}$ and $s(1) \preceq s(2) \preceq \cdots \preceq s(n)$.

(So, $j_k \in \mathcal{I}_{t(k)}$ with $s(k) \prec t(k)$ since by assumption $U_{s(k)s(k)} = \mathrm{Id}$.) Let $\mu_k = U(i_k, j_k)$. Define matrices E_k, $1 \le k \le n$, by $E_k(i_k, j_k) = -\mu_k$ and otherwise $E_k(i,j) = \delta_{ij}$. Then $(UE_k)(i_k, j_k) = 0$. Define $E = E_1 E_2 \cdots E_n$. Then by our ordering $s(1) \preceq s(2) \preceq \cdots \preceq s(n)$, we have $(UE)(i,j) = \delta_{ij}$ for $i \in \mathcal{C}^{\mathrm{sec}}$. Let $U_3 = UE$, now (i) holds, and $U = (UE)E^{-1} = U_3 E^{-1}$ as required for the diagram to commute. Also $E(i,j) = \delta_{ij}$ if $i \notin \mathcal{C}^{\mathrm{sec}}$, so $E^{-1}(i,j) = \delta_{ij}$ if $i \notin \mathcal{C}^{\mathrm{sec}}$.

Next for $1 \le k \le n$, we will define H_k such that $(E_k^{-1}, H_k)\colon (I-A) \to (I-A)$ is a positive equivalence and $H_k(i,j) = \delta_{ij}$ when $i \notin \mathcal{C}^{\mathrm{sec}}$. Then we will set $(E^{-1}, H) = (E_n^{-1} \cdots E_1^{-1}, H_1 \cdots H_n)$, so that $(E^{-1}, H)\colon (I-A) \to (I-A)$ is the composition of positive equivalences and satisfies (ii). To prepare for the definition of H_k, given k pick M a positive integer greater than the absolute value of any entry in row i_k of $E_k^{-1}(I-A)$, and define a matrix F_k as follows: $F_k(i_k, j) = -M$ if $i_k \neq j$ and $(E_k^{-1}(I-A))(i_k, j) \neq 0$, and $F_k(i,j) = \delta_{ij}$ otherwise. Define a matrix G_k by setting $G_k(i_k, j) = -(E_k^{-1}(I-A)F_k)(i_k, j)$ and $G_k(i,j) = \delta_{ij}$ otherwise. Then we have the positive equivalence

$$(I-A) \xrightarrow{(I,F_k)} \cdot \xrightarrow{(E_k^{-1},I)} \cdot \xrightarrow{(I,G_k)} (I-A).$$

Let $H_k = F_k G_k$. Note $H_k(i,j) = \delta_{ij}$ if $i \notin \mathcal{C}^{\mathrm{sec}}$, so $H(i,j) = \delta_{ij}$ if $i \notin \mathcal{C}^{\mathrm{sec}}$, and therefore also $H^{-1}(i,j) = \delta_{ij}$ if $i \notin \mathcal{C}^{\mathrm{sec}}$. We now have E, H, U_3 satisfying (i)-(iii).

To get the right half of the commuting diagram, we apply to the equivalence (U_3, V_3) the transpose of the procedure above to get matrices $\overline{E}, \overline{H}, \overline{V}$, $\overline{U}$ satisfying:

$\overline{\text{(i)}}$ $\overline{V}(i,j) = \delta_{ij}$, $\forall j \in \mathcal{C}^{\mathrm{sec}}$,
$\overline{\text{(ii)}}$ $(\overline{H}, \overline{E}^{-1})\colon (I-A) \to (I-A)$ is a positive equivalence, and
$\overline{\text{(iii)}}$ $\overline{H}^{-1}(i,j) = \delta_{ij}$, $\forall j \notin \mathcal{C}^{\mathrm{sec}}$,

where $\overline{U}$ and $\overline{V}$ are defined by $\overline{U} = U_3 \overline{H}^{-1}$ and $\overline{V} = \overline{E} V_3$. Using $\overline{\text{(i)}}$ and the forms of $(I-A)$ and $(I-A')$, we get for every $j \in \mathcal{C}^{\mathrm{sec}}$ and every i that

$$\overline{U}(i,j) = (\overline{U}(I-A))(i,j) = (\overline{U}(I-A)\overline{V})(i,j) = (I-A')(i,j) = \delta_{ij} = \overline{V}(i,j).$$

Now suppose $i \in \mathcal{C}^{\mathrm{sec}}$. We claim that $\overline{U}(i,j) = \delta_{ij}$. Suppose not. Pick $j \neq i$ such that $\overline{U}(i,j) \neq 0$. Because $\overline{U} = U_3 \overline{H}^{-1}$, it follows from (i) that $\overline{U}(i,j) = \overline{H}^{-1}(i,j)$, and then from $\overline{\text{(iii)}}$ that $j \in \mathcal{C}^{\mathrm{sec}}$. This is a contradiction.

Finally, for $i \in \mathcal{C}^{\mathrm{sec}}$ we obtain

$$\overline{V}(i,j) = ((I-A)\overline{V})(i,j) = (\overline{U}(I-A)\overline{V})(i,j) = (I-A')(i,j) = \delta_{ij}.$$

This finishes the proof. □

References

[Ba1] L. Badoian, *Flow equivalences of shifts of finite type and flow K-theory*, Ph.D. thesis, University of California, Berkeley, 1998.

[Ba2] ———, *personal communication*, 2000.

[Bow] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Springer Lecture Notes in Math., **470** (1975), Springer-Verlag, MR 56 #1364, Zbl 0308.28010.

[BowF] R. Bowen and J. Franks, *Homology for zero-dimensional basic sets*, Annals of Math., **106** (1977), 73-92, MR 56 #16692, Zbl 0375.58018.

[B1] M. Boyle, *Positive K-theory and symbolic dynamics*, in 'Dynamics and Randomness' (editors A. Maass, S. Martinez and J. San Martin), Kluwer, 2002, 31-52.

[B2] ———, *Talk at Mt. Holyoke conference* 'Classification problems in C^*-algebras and dynamics,' 1996.

[BFF] M. Boyle, D. Fiebig and U. Fiebig, *A dimension group for local homeomorphisms and endomorphisms of onesided shifts of finite type*, J. Reine Angew. Math., **487** (1997), 27-59, MR 98i:54020, Zbl 0877.58038.

[BHa] M. Boyle and D. Handelman, *Orbit equivalence, flow equivalence and ordered cohomology*, Israel J. Math., **95** (1996), 169-210, MR 98a:46082, Zbl 0871.58071.

[BH] M. Boyle and D. Huang, *Poset block equivalence of integral matrices*, Trans. Amer. Math. Soc., to appear.

[BW] M. Boyle and J. B. Wagoner, in preparation.

[C] J. Cuntz, *A class of C^*-algebras and topological Markov chains* II: *Reducible chains and the* Ext*-functor for C^*-algebras*, Inventiones Math., **63** (1981), 25-40, MR 82f:46073b, Zbl 0461.46047.

[CK] J. Cuntz and W. Krieger, *A class of C^*-algebras and topological Markov chains*, Inventiones Math., **56** (1980), 251-268, MR 82f:46073a, Zbl 0434.46045.

[DGS] M. Denker, C. Grillenberger and K. Sigmund, *Ergodic theory on compact spaces*, Springer Lecture Notes in Math., **527** (1976), MR 56 #15879, Zbl 0328.28008.

[F] J. Franks, *Flow equivalence of subshifts of finite type*, Ergod. Th. & Dynam. Sys., **4** (1984), 53-66, MR 86j:58078, Zbl 0555.54026.

[G] R. Gomez, *Finitary isomorphisms of Markov chains via positive K-theory*, Ph.D. thesis, University of Maryland, College Park, 2000.

[H1] D. Huang, *Flow equivalence of reducible shifts of finite type*, Ergod. Th. & Dynam. Sys., **14** (1994), 695-720, MR 95k:46110, Zbl 0819.46051.

[H2] ———, *The classification of two-component Cuntz-Krieger algebras*, Proc. Amer. Math. Soc., **124**(**2**) (1996), 505-512, MR 96d:46078, Zbl 0846.46040.

[H3] ———, *Flow equivalence of reducible shifts of finite type and Cuntz-Krieger algebra*, J. Reine. Angew. Math., **462** (1995), 185-217, MR 96m:46123, Zbl 0820.46065.

[H4] ———, *Automorphisms of Bowen-Franks groups for shifts of finite type*, Ergod. Th. & Dynam. Sys., **21**(**4**) (2001), 1113-1137, CMP 1 849 604.

[H5] ———, *The K-web invariant and flow equivalence of reducible shifts of finite type*, in preparation.

[KR1] K.H. Kim and F.W. Roush, *Free Z_p actions on subshifts*, Pure Math. Appl., **8**(**2**-**4**) (1997), 293-322, MR 99f:58066, Zbl 0910.58012.

[KR2] ———, *The Williams conjecture is false for irreducible subshifts*, Annals of Math., **149**(**2**) (1999), 545-558, MR 2001b:37012.

[KRW1] K.H. Kim, F.W. Roush and J.B. Wagoner, *Automorphisms of the dimension group and gyration numbers of automorphisms of a shift*, J. Amer. Math. Soc., **5** (1992), 191-212, MR 93h:54026, Zbl 0749.54012.

[KRW2] ———, *Characterization of inert actions on periodic points*, Part I, Forum Math., **12** (2000), 565-602, MR 2001g:37009.

[KRW3] ———, *Characterization of inert actions on periodic points*, Part II, Forum Math., **12**(**6**) (2000), 671-712, MR 2001j:37024.

[Ki] B. Kitchens, *Symbolic Dynamics. One-Sided, Two-Sided and Countable State Markov Shifts*, Springer-Verlag, 1998, MR 98k:58079, Zbl 0892.58020.

[Kr1] W. Krieger, *On a dimension for a class of homeomorphism groups*, Math. Ann., **252** (1980), 239-250, MR 82b:46083, Zbl 0472.54028.

[Kr2] ———, *On dimension functions and topological Markov chains*, Inventiones Math., **56** (1980), 239-250, MR 81m:28018, Zbl 0431.54024.

[LM] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, 1995, MR 97a:58050.

[Ma] K. Matsumoto, *Presentations of subshifts and their topological conjugacy invariants*, Documenta Math., **4** (1999), 285-340, MR 2000h:37013, Zbl 0926.37002.

[Ne] M. Newman, *Integral Matrices*, Academic Press, New York, 1972, MR 49 #5038, Zbl 0254.15009.

[PS] W. Parry and D. Sullivan, *A topological invariant for flows on one-dimensional spaces*, Topology, **14** (1975), 297-299, MR 53 #9179, Zbl 0314.54045.

[PT] W. Parry and S. Tuncel, *Classification problems in ergodic theory*, LMS Lecture Note Series, **67**, Cambridge Press, Cambridge, 1982, MR 84g:28024, Zbl 0487.28014.

[Rob] C. Robinson, *Dynamical systems. Stability, symbolic dynamics, and chaos*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1995, MR 97e:58064, Zbl 0914.58021.

[R] M. Rørdam, *Classification of Cuntz-Krieger algebras*, K-theory, **9** (1995), 31-58, MR 96k:46103, Zbl 0826.46064.

[Ros] J. Rosenberg, *Algebraic K-theory and its applications*, Graduate Texts in Mathematics, **147**, Springer-Verlag, 1994, MR 95e:19001, Zbl 0801.19001.

[S] S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc., **73** (1967), 747-817, MR 37 #3598, Zbl 0202.55202.

[W1] J.B. Wagoner, *Strong shift equivalence theory and the shift equivalence problem*, Bull. Amer. Math. Soc. (N.S.), **36**(**3**) (1999), 271-296, MR 2001e:54061, Zbl 0927.19001.

[W2] ———, *Strong shift equivalence and K_2 of the dual numbers*, J. Reine Angew. Math., **521** (2000), 119-160, MR 2001i:19001, Zbl 0969.37003.

[Wi] R.F. Williams, *Classification of subshifts of finite type*, Annals of Math., **98** (1973), 120-153; *Erratum*, Annals of Math., **99** (1974), 380-381, MR 48 #9769, Zbl 0282.58008.

Received September 20, 2000 and revised December 1, 2000. The research of the author was supported by NSF Grant DMS9706852.

Department of Mathematics
University of Maryland
College Park, MD 20742-4015
E-mail address: mmb@math.umd.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

MANIFOLDS WITH 2-NONNEGATIVE RICCI OPERATOR

Martha P. Dussan and Maria Helena Noronha

In this paper we study compact manifolds with 2-nonnegative Ricci operator, assuming that their Weyl operator satisfies certain conditions which generalize conformal flatness. As a consequence, we obtain that such manifolds are either locally symmetric or their Betti numbers between 2 and $n-2$ vanish. We also study the topology of compact hypersurfaces with 2-nonnegative Ricci operator.

1. Introduction.

One of the most powerful methods for studying the Betti numbers of compact manifolds is the Bochner technique. This technique is used in the context of manifolds with some type of curvature condition which will imply that harmonic forms are parallel. For 2-forms, this fact is implied by the nonnegativity of the Weitzenböck operator. In dimension four the nonnegativity of the Weitzenböck operator is equivalent (see for instance [**14**]) to the nonnegativity of the isotropic curvature, a notion introduced by Micallef-Moore ([**11**]) to study stability of harmonic 2-spheres. In that paper, the authors also point out that conformally flat manifolds with nonnegative scalar curvature have nonnegative isotropic curvature. Actually, denoting the Weitzenböck operator by $\mathcal{Q}_2$, they show that a necessary and sufficient condition for the nonnegativity of $\mathcal{Q}_2$ is $-\mathcal{W}+S/6 \geq 0$, where S is the scalar curvature and $\mathcal{W}$ is the operator induced by the Weyl tensor on the space of 2-forms $\Lambda^2(T_xM)$. The condition above follows from the fact that, in dimension 4, the isotropic curvature (and hence the Weitzenböck operator) does not depend on the traceless Ricci tensor.

In dimensions greater than 4, conformal flatness and the nonnegativity of the scalar curvature do not imply nonnegative isotropic curvature, as can be seen through the conformally flat hypersurfaces constructed in [**10**] which have $S \geq 0$ but some isotropic curvatures are negative. The same examples of conformally flat manifolds show that for $n > 4$, $S \geq 0$ does not imply $\mathcal{Q}_2 \geq 0$.

The role of the Ricci tensor in the study of the isotropic curvature and the Weitzenböck operator for dimensions $n > 4$ is not yet clear. It turns out (see below) that the condition $-\mathcal{W} + S/6 \geq 0$ used for 4-manifolds generalizes to $-\mathcal{W} + S/[(n-2)(n-1)] \geq 0$ and this paper searches for hypotheses on

the Ricci tensor that together with $-\mathcal{W} + S/[(n-2)(n-1)] \geq 0$ imply that $\mathcal{Q}_2 \geq 0$. Our first result in this paper is the following:

Theorem 1. *Let M^n, $n > 4$, be a compact, locally irreducible manifold with nonnegative Ricci curvature. If $-\mathcal{W} + S/[(n-2)(n-1)] \geq 0$ then one of the following occurs:*

(a) *M is covered by a compact symmetric space.*
(b) *The Betti numbers $\beta_p(M) = 0$, for $1 \leq p \leq n-1$.*

The key point of the proof of this result is to conclude that if $n > 4$ then the restricted holonomy group of metrics with $\mathcal{Q}_2 \geq 0$ and $-\mathcal{W} + S/[(n-2)(n-1)] \geq 0$ is essentially $SO(n)$.

In the next result we assume a weaker condition for the Ricci curvature, namely, that the manifold M has 2-nonnegative Ricci operator, that is to say *the sum of the smallest 2 eigenvalues of the Ricci operator is nonnegative.* We will consider such a condition on manifolds whose Weyl operator $\mathcal{W}$ commutes with $\mathrm{Ric} \wedge I$, where Ric and I denote the Ricci and the identity operators respectively. For such manifolds we have the following result:

Theorem 2. *Let M^n, $n > 4$, be a compact, locally irreducible manifold with 2-nonnegative Ricci operator. Let us suppose that $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$ and $-\mathcal{W} + S/[(n-2)(n-1)] \geq 0$. Then one of following occurs:*

(a) *M is covered by a compact symmetric space.*
(b) *The Betti numbers $\beta_p(M) = 0$, for $2 \leq p \leq n-2$.*

Observe that the Weyl operator of three important classes of manifolds commutes with $\mathrm{Ric} \wedge I$: Conformally flat, Einstein and manifolds with pure curvature tensor (see definition on page 439 of [**4**]). We also show that two other types of metrics satisfy the condition $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$. They are G-manifolds of cohomogeneity one and Riemannian manifolds with harmonic curvature, non-parallel Ricci tensor and such that the operator Ric has less than three distinct eigenvalues. The last class of manifolds was studied by Derdzisnki in [**7**] and [**8**]. Such manifolds were the first examples of compact manifolds with harmonic curvature and non-parallel Ricci tensor and hence not Einstein. Among them we find, for $n > 4$, examples that are not conformally flat either.

We also prove that locally reducible conformally flat manifolds with 2-nonnegative Ricci operator in fact have nonnegative Ricci curvature. Using this fact, Theorem 1 above implies the corollary below, which generalizes Theorem 1 of [**13**].

Corollary 1. *Let M^n, $n \geq 4$, be a compact conformally flat manifold with 2-nonnegative Ricci operator. Then either M is flat or $\beta_p(M) = 0$ for $2 \leq p \leq n-2$. Moreover if $\beta_1(M) \neq 0$ then M is a quotient of $S^{n-1} \times \mathbf{R}$ or $\mathbf{R}^n$ by a group of fixed point free isometries in the standard metrics.*

We point out that for locally irreducible conformally flat manifolds, Theorem 2 above and Theorem 2 of [**10**] have the same conclusion, namely, that $\beta_p(M) = 0$, for $2 \leq p \leq n-2$. This gives rise to a corollary with the same proof as Corollary 1 in [**10**]:

Corollary 2. *Let $f : M^n \to \mathbf{R}^{n+p}$, $2 \leq p \leq n/2-1$, be an isometric immersion of a compact, orientable, locally irreducible conformally flat manifold M with 2-nonnegative Ricci operator. Then $H_i(M; \mathbf{Z}) = 0$ for $p \leq i \leq n-p$.*

Another similarity between the topology of manifolds of 2-nonnegative Ricci operator and nonnegative isotropic curvature appears in the context of hypersurfaces of Euclidean spaces. For these, we prove the result below (compare with Theorem 1 of [**10**]):

Theorem 3. *Let $f : M^n \to \mathbf{R}^{n+1}$, $n \geq 4$, be an isometric immersion of a compact manifold M with 2-nonnegative Ricci operator. Then the homology groups*

$$H_i(M; \mathbf{Z}) = 0 \quad \textit{for} \quad 2 \leq i \leq n-2$$

and the fundamental group $\pi_1(M)$ is a free group on β_1 elements.

2. Manifolds with nonnegative Weitzenböck operator.

Let M be a Riemannian manifold and $\mathrm{Ric} : T_xM \to T_xM$ denote the Ricci operator given by

$$\langle \mathrm{Ric}\,(X), Y \rangle = \mathrm{Ric}\,(X, Y).$$

In this paper we will use the same notation for a tangent vector X and its dual form. With this in mind, we define the *Weitzenböck operator* $\mathcal{Q}_2 : \Lambda^2(T_xM) \to \Lambda^2(T_xM)$ as

$$\begin{aligned} \mathcal{Q}_2(X \wedge Y) &= (\mathrm{Ric} \wedge I)(X \wedge Y) - 2\mathcal{R}\,(X \wedge Y) \\ &= \mathrm{Ric}\,(X) \wedge Y + X \wedge \mathrm{Ric}\,(Y) - 2\mathcal{R}\,(X \wedge Y), \end{aligned}$$

where $\mathcal{R}$ is the curvature operator and $\Lambda^2(T_xM)$ denotes the space of 2-forms. This operator satisfies the well-known *Weitzenböck formula*, e.g., $\Delta\omega = -\mathrm{div}\nabla\omega + \mathcal{Q}_2(\omega)$, where Δ is the Laplace-Beltrami operator and $\nabla\omega$ the covariant derivative of ω.

It is easy to see that $\mathcal{Q}_2$ is a self-adjoint operator, and therefore it makes sense to study it when it is nonnegative. The nonnegativity of the Weitzenböck operator has been used to study the second Betti number of compact manifolds. In this section we collect some results along this line.

Lemma 2.1. *Let M be a Riemannian manifold with nonnegative Weitzenböck operator. Then:*

(a) *If e_1, e_2 are orthonormal vectors, we have $\mathrm{Ric}\,(e_1, e_1) + \mathrm{Ric}\,(e_2, e_2) - 2K_{12} \geq 0$, where K_{12} is the sectional curvature of the plane spanned by e_1 and e_2.*

(b) *The scalar curvature S is nonnegative.*
(c) *If $S \equiv 0$ and $n > 4$ then M is flat.*
(d) *If $S \equiv 0$ and $n = 4$ then M is conformally flat.*

Proof. Since $\mathcal{Q}_2 \geq 0$, (a) comes straightforward from the definition of $\mathcal{Q}_2$.

Now we have that $\mathrm{Ric}\,(e_1, e_1) + \mathrm{Ric}\,(e_j, e_j) - 2K_{1j} \geq 0$ for all unit vectors e_j that are orthogonal to e_1. We obtain

(2.2)
$$(n-1)\mathrm{Ric}\,(e_1, e_1) + \sum_{j \neq 1} [\mathrm{Ric}\,(e_j, e_j) - 2K_{1j}] = (n-4)\mathrm{Ric}\,(e_1, e_1) + S \geq 0.$$

Therefore, if $n = 4$ we have $S \geq 0$. If $n > 4$ and the Ricci curvature is nonnegative then $S \geq 0$. If $\mathrm{Ric}\,(e_1, e_1) < 0$ then $S > 0$.

To prove (c), observe that Equation (2.2) also implies that if $S \equiv 0$ and $n > 4$ then the Ricci curvature is nonnegative. If $S \equiv 0$, we conclude that M is Ricci flat. This substituted in (a) implies that the sectional curvature $K \leq 0$ which gives $K = 0$, again because $S \equiv 0$. The result in (d) is well-known (see for instance [**12**], Proposition 2.5 or [**15**], Proposition 2.5).

Proposition 2.3. *Let M^n, $n \geq 4$, be a locally irreducible compact manifold with nonnegative Weitzenböck operator. Then:*

(a) *If M is even dimensional and $\beta_2(M) \neq 0$ then $\beta_2(M) = 1$ and M is a simply connected Kähler manifold with positive first Chern class. Further, if $n = 4$, then M is biholomorphic to the complex projective space $\mathbf{CP}^2$.*
(b) *If M is odd dimensional and $\beta_2(M) \neq 0$ then M is covered by symmetric space of [compact type] and* rank > 1.

Proof. Since M is compact, it follows from the nonnegativity of $\mathcal{Q}_2$ and the Weitzenböck formula that a harmonic 2-form ω is parallel.

If M is even dimensional, the proof of Theorem 2.1(b) of [**12**] applies here, since it depends only on the fact that harmonic 2-forms are parallel and $S \geq 0$ but not zero. Since $S = 0$ implies that M is flat and this contradicts the irreducibility of M we conclude the first assertion of (a). The second part follows from Theorem 1 of [**14**].

If M is odd dimensional, since we are supposing that M is locally irreducible, then so is the restricted holonomy group G. Recall that in [**2**], Berger proved that if for some $x \in M$, G acts irreducibly on T_xM, then either M is locally symmetric or G is one of the standard subgroups of $SO(n)$:

$$SO(n), U(m)(n = 2m), Sp(m) \times Sp(1)(n = 4m > 4), \mathrm{Spin}\,(9)(n = 16)$$
$$SU(m)(n = 2m > 2), Sp(m)(n = 4m > 4), G_2(n = 7), \mathrm{Spin}\,(7)(n = 8).$$

In the case that M is locally symmetric, the universal cover $\widetilde{M}$ is an irreducible symmetric space. Since $\widetilde{M}$ is Einstein, if $S = 0$, M would be

Ricci flat and then flat by Lemma 2.1. Therefore $S > 0$ and $\widetilde{M}$ is compact. If $\operatorname{rank}(\widetilde{M}) = 1$, being odd-dimensional, it would be isometric to a sphere contradicting that $\beta_2(M) \neq 0$. Berger also proved that if G is one of the possibilities listed on the second line above, M is Ricci flat, which in this case implies that M is flat. Note that in the other possibilities for G, M is even dimensional, except if $G = SO(n)$. In this case, the existence of a parallel 2-form ω would give rise to a parallel and hence harmonic 2-form on S^n by the holonomy principle, and this is a contradiction.

3. A special condition on the Weyl tensor.

We start this section proving a result for manifolds with nonnegative Weitzenböck operator and Weyl tensor satisfying a condition which generalizes conformal flatness. Before we state the result, we recall that the Weyl tensor induces an operator

$$\mathcal{W} : \Lambda^2(T_xM) \to \Lambda^2(T_xM)$$

given by

$$\mathcal{W}(X \wedge Y) = \mathcal{R}(X \wedge Y) - \Gamma(X) \wedge Y - X \wedge \Gamma(Y)$$

where $\Gamma : T_xM \to T_xM$ is defined by

$$\Gamma(X) = \frac{1}{n-2}\left(\operatorname{Ric}(X) - \frac{S}{2(n-1)}X\right).$$

It is well-known that conformal flatness for manifolds of dimension $n \geq 4$ is equivalent to $\mathcal{W} \equiv 0$.

Lemma 3.1. *Let $\mathcal{Q}_2, \mathcal{R}, \mathcal{W}$ denote the Weitzenböck, curvature and Weyl operator respectively. We have:*

(a)

$$\mathcal{Q}_2 - (n-4)\mathcal{R} = \frac{S}{n-1} - (n-2)\mathcal{W}.$$

(b) *If $-\mathcal{W} + S/[(n-2)(n-1)] \geq 0$ and $\mathcal{Q}_2$ is a nonnegative operator ($\mathcal{Q}_2 \geq 0$) then*

$$\mathcal{Q}_2 - 2(p-2)\mathcal{R} \geq 0 \quad \textit{whenever} \quad p \leq [n/2].$$

Proof. Using the definition of $\mathcal{W}$ we obtain (a).

For (b), observe first that the assumptions imply that $\mathcal{Q}_2 - (n-4)\mathcal{R}$ is a nonnegative operator. Now let μ be an eigenvalue of $\mathcal{Q}_2 - 2(p-2)\mathcal{R}$ with corresponding eigenvector ϕ. If $\langle \mathcal{R}(\phi), \phi\rangle \leq 0$, then $\langle (\mathcal{Q}_2 - 2(p-2)\mathcal{R})(\phi), \phi\rangle \geq 0$, since we are supposing that $\mathcal{Q}_2 \geq 0$. If $\langle \mathcal{R}(\phi), \phi\rangle \geq 0$, we have for $p \leq [n/2]$

$$\langle (\mathcal{Q}_2 - 2(p-2)\mathcal{R})(\phi), \phi\rangle \geq \langle (\mathcal{Q}_2 - (n-4)\mathcal{R})(\phi), \phi\rangle \geq 0.$$

Theorem 3.2. *Let $M^n, n \geq 4$ be a locally irreducible compact manifold with nonnegative Weitzenböck operator and such that $-\mathcal{W}+S/[(n-2)(n-1)] \geq 0$. Then one of the following occurs:*

(a) *M is locally symmetric and covered by a compact symmetric space.*
(b) *The Betti numbers $\beta_p(M) = 0$ for $2 \leq p \leq n-2$.*
(c) *M is 4-dimensional manifold biholomorphic to the complex projective space $\mathbf{CP}^2$.*

Proof. Without loss of generality, we asume that M is orientable. Let ω be a harmonic p-form. We use the Weitzenböck formula for p-forms (see [**9**])

$$(\Delta\omega, \omega) = (\nabla\omega, \nabla\omega) + \int_M F(\omega) dV$$

where $(\,,)$ denotes the L^2-product with respect to the Riemannian volume density dV and

$$F(\omega) = \frac{1}{(p-1)!}\left[A - \frac{p-1}{2}B\right],$$

with

$$\begin{aligned} A = \sum_{i_3,\dots,i_p} \sum_{r,s,u,t} & \omega(X_s, X_r, X_{i_3}, \dots, X_{i_p}) \\ & \cdot \omega(X_t, X_r, X_{i_3}, \dots, X_{i_p})\langle R(X_s, X_u)X_u, X_t\rangle \\ B = \sum_{i_3,\dots,i_p} \sum_{r,s,u,t} & \omega(X_r, X_s, X_{i_3}, \dots, X_{i_p}) \\ & \cdot \omega(X_t, X_u, X_{i_3}, \dots, X_{i_p})\langle R(X_r, X_s)X_u, X_t\rangle. \end{aligned}$$

Notice that $F(\omega)$ can be written as

$$F(\omega) = \frac{1}{(p-1)!} \sum_{i_3,\dots,i_p} \langle (\mathcal{Q}_2 - 2(p-2)\mathcal{R})(\phi_{i_3,\dots,i_p}), \phi_{i_3,\dots,i_p}\rangle$$

where $\phi_{i_3,\dots,i_p}$ is a 2-form obtained by fixing $X_{i_3}, \dots, X_{i_p}$ and defining

$$\phi_{i_3,\dots,i_p}(u, v) = \omega(u, v, X_{i_3}, \dots, X_{i_p}).$$

Therefore, $\mathcal{Q}_2 - 2(p-2)\mathcal{R} \geq 0$ implies $F(\omega) \geq 0$.

Proceeding as the proof of Lemma 3.1(b), we conclude that $(\mathcal{Q}_2 - 2(p-2)\mathcal{R}) \geq 0$ for $2 \leq p \leq [n/2]$ and hence for p in this range, a harmonic p-form is parallel. Again, we study each possibility for the restricted holonomy group G and use the holonomy principle.

If M is locally symmetric, being locally irreducible, $\widetilde{M}$ is an irreducible symmetric space and therefore an Einstein space. Since it cannot be Ricci flat, it has positive Ricci curvature and hence compact.

The fact that M cannot be Ricci flat leaves us with the following possibilities:

$$SO(n), U(m)(n = 2m), Sp(m) \times Sp(1)(n = 4m > 4), \mathrm{Spin}\,(9)(n = 16).$$

In the last case, a result in [**5**] implies that M is locally symmetric, and we repeat the previous argument.

Recall that if $\beta_2(M) \neq 0$, and M is even dimensional then M is a Kähler manifold and then there exists a parallel form ω for which $\mathcal{Q}_2(\omega) = 0$. Moreover we can find orthonomal vectors $e_1, \ldots, e_m$, $n = 2m$, such that

$$\omega = J(e_1) \wedge e_1 + \cdots + J(e_m) \wedge e_m,$$

where J denotes the complex structure on M. Using the fact that $\mathcal{Q}_2(\omega) = 0$, from the definition of $\mathcal{Q}_2$ we get

$$\begin{aligned} 0 = {} & \operatorname{Ric}(e_1, e_1) + \operatorname{Ric}(J(e_1), J(e_1)) + \cdots + \operatorname{Ric}(e_m, e_m) \\ & + \operatorname{Ric}(J(e_m), J(e_m)) - 2\langle \mathcal{R}(\omega), \omega \rangle, \end{aligned}$$

which gives

$$\langle \mathcal{R}(\omega), \omega \rangle = \frac{S}{2}$$

yielding

$$\langle \mathcal{W}(\omega), \omega \rangle = \frac{(n-2)S}{2(n-1)}.$$

On the other hand, let $\{\phi_i\}$ denote an orthonormal basis which diagonalizes $\mathcal{W}$ with corresponding eigenvalues ν_i. We then write $\omega = \sum_i a_i \phi_i$, and then

$$\langle \mathcal{W}(\omega), \omega \rangle = \frac{(n-2)S}{2(n-1)} = \sum_i a_i^2 \nu_i.$$

Let us suppose that the eigenvalues ν_i's are increasing and let i_0 denote the index such that $\nu_i \geq 0$, for $i \geq i_0$. Therefore, from our assumption on the eigenvalues of the Weyl operator, we get

$$\frac{(n-2)S}{2(n-1)} \leq \sum_{i \geq i_0} a_i^2 \frac{S}{(n-1)(n-2)} \leq \frac{S}{(n-1)(n-2)} \frac{n}{2}.$$

But the above implies either $(n-2)^2 \leq n$, which is clearly a contradiction for $n > 4$, or $S = 0$. But $S = 0$ contradicts the irreducibility of M, since it implies that M is flat. Therefore, if $n > 4$, $\beta_2(M) = 0$ and hence the holonomy G cannot be $U(m)$. If $n = 4$, we obtain that (c) follows from Proposition 2.3.

If $G = Sp(m) \times Sp(1)$, M is Einstein (see [**3**]) and hence has positive Ricci curvature. Furthermore, M is a quaternionic Kähler manifold which implies the existence of a parallel 4-form (V.Y. Kraines, see [**4**] p. 419), which we denote by ω, and then $F(\omega) = 0$. From the equation

$$F(\omega) = \frac{1}{(3)!} \sum_{i_3, i_4} \langle (\mathcal{Q}_2 - 4\mathcal{R})(\phi_{i_3, i_4}), \phi_{i_3, i_4} \rangle,$$

and the fact that $(\mathcal{Q}_2 - 4\mathcal{R})$ is a nonnegative operator, we obtain

$$\langle(\mathcal{Q}_2 - 4\mathcal{R})(\phi_{i_3,i_4}), (\phi_{i_3,i_4})\rangle = 0, \tag{3.3}$$

for all 2-forms of type ϕ_{i_3,i_4}. On the other hand, we have $\langle(\mathcal{Q}_2 - (n-4)\mathcal{R})(\phi_{i_3,i_4}), (\phi_{i_3,i_4})\rangle \geq 0$, and using (3.3) we obtain

$$(8-n)\langle\mathcal{R}(\phi_{i_3,i_4}), \phi_{i_3,i_4}\rangle \geq 0.$$

If $n > 8$, the above implies that $\langle\mathcal{Q}_2(\phi_{i_3,i_4}), \phi_{i_3,i_4}\rangle = 4\langle\mathcal{R}(\phi_{i_3,i_4}), \phi_{i_3,i_4}\rangle \leq 0$. Since $\mathcal{Q}_2 \geq 0$, we then have that

$$\langle\mathcal{Q}_2(\phi_{i_3,i_4}), \phi_{i_3,i_4}\rangle = \langle\mathcal{R}(\phi_{i_3,i_4}), \phi_{i_3,i_4}\rangle = 0.$$

But M is Einstein and hence $\mathcal{Q}_2 = 2(S/n) - 2\mathcal{R}$. Therefore, the equation above gives

$$\frac{2S||\phi_{i_3,i_4}||^2}{n} = 0,$$

implying $S = 0$, which is the desired contradiction.

If $n = 8$, (3.3) substituted in Lemma 3.1(a) immediately implies

$$\langle\mathcal{W}(\phi_{i_3,i_4}), \phi_{i_3,i_4}\rangle = \frac{S||\phi_{i_3,i_4}||^2}{42},$$

and we claim that ϕ_{i_3,i_4} is an eigenvector of $\mathcal{W}$. In fact, if not, we consider again an orthonormal basis $\{\phi_i\}$ which diagonalizes $\mathcal{W}$, and let i_0 denote the index such that $\nu_i \geq 0$, for $i \geq i_0$. We would have

$$\langle\mathcal{W}(\phi_{i_3,i_4}), \phi_{i_3,i_4}\rangle < \sum_{i\geq i_0} a_i^2 \frac{S}{42} < \frac{S||\phi_{i_3,i_4}||^2}{42}.$$

Since M is Einstein, an eigenvector of $\mathcal{W}$ is also an eigenvector $\mathcal{R}$, and then we obtain that

$$\mathcal{R}(\phi_{i_3,i_4}) = \frac{S}{24}\phi_{i_3,i_4}.$$

We will show now that one can obtain a basis of $\Lambda^2(T_xM)$ whose elements are 2-forms of type ϕ_{i_3,i_4}. This implies $\mathcal{R} = (S/24)I$, that is, M is a manifold of constant curvature contradicting that its restricted holonomy group G is $Sp(2)\times Sp(1)$. For that, let I, J, K denote the almost complex structures of M which satisfy the relations $IJ = -JI$, and $K = IJ$. Let $\{e_1, \ldots, e_8\}$ be an orthonormal basis with the property

$$\begin{aligned} e_2 &= I(e_1), \qquad e_3 = J(e_1), \qquad e_4 = K(e_1) \\ e_6 &= I(e_5), \qquad e_7 = J(e_5), \qquad e_8 = K(e_5). \end{aligned}$$

The form ω is given by

$$\omega = \alpha\wedge\alpha + \beta\wedge\beta + \gamma\wedge\gamma,$$

where

$$\alpha(X,Y) = \langle I(X), Y\rangle \quad \beta(X,Y) = \langle J(X), Y\rangle \quad \gamma(X,Y) = \langle K(X), Y\rangle.$$

Therefore α, β and γ are written as

$$\alpha = e_1 \wedge e_2 + e_3 \wedge e_4 + e_5 \wedge e_6 + e_7 \wedge e_8$$

$$\beta = e_1 \wedge e_3 - e_2 \wedge e_4 + e_5 \wedge e_7 - e_6 \wedge e_8$$

$$\gamma = e_1 \wedge e_4 + e_2 \wedge e_3 + e_5 \wedge e_8 + e_6 \wedge e_7,$$

and then

$$\begin{aligned}\omega = {} & 6\, e_1 \wedge e_2 \wedge e_3 \wedge e_4 + 2\, e_1 \wedge e_2 \wedge e_5 \wedge e_6 \\ & + 2\, e_1 \wedge e_2 \wedge e_7 \wedge e_8 + 2\, e_1 \wedge e_3 \wedge e_5 \wedge e_7 - 2\, e_1 \wedge e_3 \wedge e_6 \wedge e_8 \\ & + 2\, e_1 \wedge e_4 \wedge e_5 \wedge e_8 + 2\, e_1 \wedge e_4 \wedge e_6 \wedge e_7 + 2\, e_2 \wedge e_3 \wedge e_5 \wedge e_8 \\ & + 2\, e_2 \wedge e_3 \wedge e_6 \wedge e_7 - 2\, e_2 \wedge e_4 \wedge e_5 \wedge e_7 + 2\, e_2 \wedge e_4 \wedge e_6 \wedge e_8 \\ & + 2\, e_3 \wedge e_4 \wedge e_5 \wedge e_6 + 2\, e_3 \wedge e_4 \wedge e_7 \wedge e_8 + 6\, e_5 \wedge e_6 \wedge e_7 \wedge e_8.\end{aligned}$$

As before, let us consider the 2-form $\phi_{i,j}(u, v) = \omega(u, v, e_i, e_j)$. From the expression of ω it is straightforward to conclude that $\{\phi_{i,j}, i < j\}$ is a basis of $\wedge^2(T_xM)$. Since $G \neq U(m), Sp(m) \times Sp(1)$, if $n > 4$ then the only possibility for G is $SO(n)$. The holonomy principle implies that $\beta_p(M) = 0$ for $2 \leq p \leq [n/2]$ and we conclude $\beta_p(M) = 0$ for $2 \leq p \leq n - 2$ by duality.

Now we use Theorem 3.2 to prove Theorem 1 stated in the introduction.

Proof of Theorem 1. We show that the hypotheses imply $\mathcal{Q}_2 \geq 0$. In fact, let ω be a unit eigenvector of $\mathcal{Q}_2$. There exist an orthonormal set $\{e_1, \ldots, e_{2m}\}$ of T_xM and numbers $a_1, \ldots, a_m$ such that

$$\omega = a_1\, e_1 \wedge e_2 + \cdots + a_m\, e_{2m-1} \wedge e_{2m}.$$

From the definition of $\mathcal{Q}_2$ we obtain

$$\begin{aligned}\langle \mathcal{Q}_2(\omega), \omega \rangle &= \sum_{i=1}^{m} a_i^2 \big(\mathrm{Ric}\,(e_{2i-1}, e_{2i-1}) + \mathrm{Ric}\,(e_{2i}, e_{2i})\big) - 2\langle \mathcal{R}(\omega), \omega \rangle \\ &= \frac{n-4}{n-2} \sum_{i=1}^{m} a_i^2 \big(\mathrm{Ric}\,(e_{2i-1}, e_{2i-1}) + \mathrm{Ric}\,(e_{2i}, e_{2i})\big) \\ &\quad - 2\left[\langle \mathcal{W}(\omega), \omega \rangle - \frac{S}{(n-2)(n-1)}\right] \geq 0.\end{aligned}$$

Now, Theorem 3.2 implies (a) or $\beta_p(M) = 0$ for $2 \leq p \leq n - 2$. Since we are also assuming that all Ricci curvatures are nonnegative, we apply the well-known generalization of Bochner's theorem, namely, that either M is covered by a compact symmetric space or it is Ricci flat or $\beta_1(M) = 0$. Since our hypotheses imply $\mathcal{Q}_2 \geq 0$, M cannot be Ricci flat and this finishes the proof of the theorem.

Next we want to examine another condition on the Weyl operator that also generalizes $\mathcal{W} = 0$. Such a condition is

$$[\text{Ric} \wedge I, \mathcal{W}] = 0.$$

This condition is satisfied by several important classes of Riemannian manifolds. Among them, we easily find the Einstein manifolds. In this section we show other classes of manifolds whose Weyl operator commutes with $\text{Ric} \wedge I$.

Recall that the curvature operator $\mathcal{R}$ is said to be *pure* if there exists an orthonormal basis $\{e_1, \dots, e_n\}$ of the tangent space such that the basis of 2-forms $\{e_i \wedge e_j\}_{i<j}$ diagonalizes $\mathcal{R}$. We call the basis $\{e_1, \dots, e_n\}$ an $\mathcal{R}$-basis.

Notice that the Weyl tensor of a manifold with pure curvature operator satisfies $[\text{Ric} \wedge I, \mathcal{W}] = 0$. This class of manifolds also includes hypersurfaces of Euclidean spaces, and more generally, manifolds which admit isometric immersions into a space of constant curvature with flat normal bundle. To see this, just use the Ricci equation which implies that there is an orthonormal basis that diagonalizes simultaneously all the Weingarten operators; then from the Gauss equation one obtains the $\mathcal{R}$-basis. The technical condition of the next lemma will appear naturally in two other classes of manifolds.

Lemma 3.4. *Let M be a Riemannian manifold such that for every point $x \in M$, the Ricci operator Ric_x has an eigenvalue $\lambda(x)$ of constant multipicity $n-1$. Suppose that the eigenspaces E_λ corresponding to λ form an integrable distribution. If their leaves are totally umbilic submanifolds and have constant mean curvature then $[\text{Ric} \wedge I, \mathcal{W}] = 0$.*

Proof. Let $\{e_1, \dots, e_n\}$ be an orthonormal basis such that $\text{Ric}(e_1) = \mu e_1$ and $\text{Ric}(e_i) = \lambda e_i$, for $i \geq 2$. We show first that

$$\langle \mathcal{R}(e_1 \wedge e_k), e_i \wedge e_j \rangle = 0, \quad i, j, k \geq 2.$$

For that, let Σ denote a maximal leaf of E_λ. Let A denote the shape operator of the inclusion $i : \Sigma \to M$ with only an eigenvalue of multiplicity $n-1$ denoted by a. Since a is constant, it is straightforward to verify that A satifies the Codazzi equation

$$\langle (\overline{\nabla}_{e_i} A)(e_j), e_k \rangle = \langle (\overline{\nabla}_{e_j} A)(e_i), e_k \rangle, \quad \forall\, i, j, k \geq 2,$$

where $\overline{\nabla}$ is the induced connection on Σ. This fact implies $\langle \mathcal{R}(e_i \wedge e_j), e_1 \wedge e_k \rangle = 0$. Then, we have that

$$\mathcal{W}(e_1 \wedge e_k) = \mathcal{R}(e_1 \wedge e_k) - \frac{\mu}{n-1} e_1 \wedge e_k$$

lies in the space $V = \text{span}\{e_1 \wedge e_k, k \geq 2\}$. Since $\text{Ric} \wedge I$ restricted to V is a multiple of the identity, we have that $\mathcal{W}$ and $\text{Ric} \wedge I$ commute on V. We

also have, for $i, j \geq 2$,

$$\mathcal{W}(e_i \wedge e_j) = \mathcal{R}(e_i \wedge e_j) - \frac{(n-1)\lambda - \mu}{(n-1)(n-2)} e_i \wedge e_j,$$

and then $\mathcal{W}(e_i \wedge e_j) \in U = \operatorname{span}\{e_i \wedge e_j, i, j \geq 2, i < j\}$ implying that $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$ on U.

Definition 3.5. A Riemannian G-manifold is said to be of *cohomogeneity one* if the group G acts effectively and isometrically with principal orbits of codimension one.

Proposition 3.6. *Let M be a cohomogeneity one G-manifold such that its principal orbits are isotropy-irreducible homogeneous spaces* (*see* [**4**, p. 187]). *Then the set of regular points M_{reg} of M satisfies the conditions of Lemma* 3.4. *It follows that* $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$ *for all points of M.*

Proof. Since Σ is an isotropy-irreducible homogeneous space, the immersion $i : \Sigma \to M$ is totally umbilic. Further, a G invariant metric defined on an isotropy-irreducible homogeneous space is Einstein. From this and the fact that the immersion of the orbit Σ into M is totally umbilic, we obtain that the operator Ric_x is almost umbilic for all $x \in M_{\mathrm{reg}}$. Note that such an immersion has constant mean curvature, by the homogeneity of Σ . Therefore, from the Lemma above we get that $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$ on 2-forms defined on $\Lambda^2(T_xM)$ for $x \in M_{\mathrm{reg}}$. Since M_{reg} is dense in M, we have the result.

Proposition 3.7. *Let M be a Riemannian manifold with harmonic curvature and non-parallel Ricci tensor. If* Ric *has less than three distinct eigenvalues at any point of M then M satisfies the conditions of Lemma* 3.4 *and hence* $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$ *for all points of M.*

Proof. The proof that M satisfies the conditions of Lemma 3.4 is Lemma 3 of [**7**].

Lemma 3.8. *Let M be Riemannian manifold with the property that* $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$. *Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of eigenvectors of* Ric *with corresponding eigenvalues μ_i. Then:*

(a) $\mathcal{R}(e_i \wedge e_j)$ *and* $F(e_i \wedge e_j)$ *are eigenvectors of* $\mathrm{Ric} \wedge I$ *with corresponding eigenvalue $\mu_i + \mu_j$.*

(b) *Let E_{μ_i} denote the eigenspace of μ_i. If $\{e_1, \ldots, e_k\}$ is a basis of E_{μ_i} and $\{e_{k+1}, \ldots, e_m\}$ a basis of E_{μ_j} then the space* $\operatorname{span}\{e_r \wedge e_s,\ r = 1, \ldots, k,\ s = k+1, \ldots, m\}$ *is invariant by $\mathcal{R}$ and $\mathcal{Q}_2$.*

Proof. Since the condition $[\mathrm{Ric} \wedge I, \mathcal{W}] = 0$ implies that $\mathrm{Ric} \wedge I$ commutes with $\mathcal{R}$ and $\mathcal{Q}_2$ and $\mu_i + \mu_j$ is an eigenvalue of $\mathrm{Ric} \wedge I$ with corresponding eigenvector $e_i \wedge e_j$, we have Part (a), which immediately implies (b).

4. Manifolds with 2-nonnegative Ricci operator.

It has been shown by H. Chen ([**6**]) that, from the topological point of view, compact manifolds with 2-nonnegative curvature operator are the same as the ones with nonnegative curvature operator. In this section we want to investigate to what extent the topology of manifolds with nonnegative Ricci curvature and 2-nonnegative Ricci operator can be compared.

Definition 4.1. The Ric operator is said to be 2-nonnegative (respectively, positive) if the sum of the first 2 eigenvalues is nonnegative (respectively, positive).

Proposition 4.2. *Let M^n be a locally reducible Riemannian manifold with 2-nonnegative Ricci operator. If M does not have nonnegative Ricci curvature then the universal cover $\widetilde{M}$ is isometric $N_1 \times \cdots \times N_m$ where each N_i is irreducible and non-flat and one the N_i's is at least 3-dimensional and has 2-nonnegative Ricci operator and all other factors have nonnegative Ricci curvature.*

Proof. The universal covering $\widetilde{M}$ is isometric to $\mathbf{R}^k \times N_1 \times \cdots \times N_m$ by the decomposition theorem of de Rham. If $k \geq 1$ then $\mathrm{Ric}\,(X) = 0$ for all X that is tangent to $\mathbf{R}^k$ and then M has nonnegative Ricci curvature. If $k = 0$, and one of the N_i's is 2-dimensional, its curvature must be nonnegative, otherwise Ric would have 2 negative eigenvalues. Therefore, if Ric has a negative eigenvalue, one of the N_i's has dimension at least 3. The remanining statements now are obvious.

Corollary 4.3. *Let M^n be a locally reducible conformally flat manifold with 2-nonnegative Ricci operator. Then M has nonnegative Ricci curvature.*

Proof. It follows from Proposition 4.2 that the only case to be studied here is $\widetilde{M} = N_1 \times \cdots \times N_m$ where a factor N_{i_0} has dimension $k \geq 3$, 2-nonnegative Ricci operator, and all other factors are at least 2-dimensional. Let $\{e_1, \ldots, e_k\}$ be a basis of vectors tangent to N_{i_0} and e_r, e_s orthonormal vectors tangent to $N_j, j \neq i_0$. Since M is conformally flat we have that

$$K_{12} + K_{34} = K_{13} + K_{24},$$

whenever e_1, e_2, e_3, e_4 are orthonormal vectors. Using this relation we get

$$K_{ij} + K_{rs} = K_{ir} + K_{js} = 0, \quad \forall\, i, j = 1, \ldots k, i \neq j.$$

This implies that N_{i_0} has constant curvature and hence positive Ricci curvature.

Now we are ready to prove Theorem 2.

Proof of Theorem 2. Again we show that the hypotheses of the theorem imply that the Weitzenböck operator $\mathcal{Q}_2$ is nonnegative and the result will follow from Theorem 3.2.

Let ω be a unit eigenvector of $\mathcal{Q}_2$. From Part (b) of Lemma 3.8 (and with same notation) we get that $\omega \in \operatorname{span}\{e_r \wedge e_s,\ r = 1, \dots, k,\ s = k+1, \dots, m\}$, for some i, j. Since we have

$$\mathcal{R}(\omega) = \mathcal{W}(\omega) + \Big(\frac{\mu_i + \mu_j}{n-2}\Big)\,\omega - \frac{S}{(n-1)(n-2)}\,\omega$$

$$\mathcal{Q}_2(\omega) = (\mu_i + \mu_j)\,\omega - 2\mathcal{R}(\omega),$$

we obtain

$$\langle \mathcal{Q}_2(\omega), \omega\rangle = \frac{(n-4)(\mu_i+\mu_j)}{n-2} - 2\left[\langle \mathcal{W}(\omega), \omega\rangle - \frac{S}{(n-1)(n-2)}\right] \geq 0.$$

Combining the results of Corollary 4.3, Theorem 2 and Theorem of [**13**] we obtain Corollary 1 stated in the introduction.

Now we use Theorem 2 to conclude the following results for the manifolds studied in the last section.

Theorem 4.4. *Let $M^n, n \geq 5$, be a compact, locally irreducible cohomogeneity one G-manifold such that its principal orbits are isotropy-irreducible homogeneous spaces. If M has 2-nonnegative Ricci operator and $-\mathcal{W} + S/\big((n-1)(n-2)\big) \geq 0$ then $\beta_i(M) = 0$, for $2 \leq i \leq n-2$.*

Proof. Observe first that combining Proposition 3.6 and Theorem 2 we get either $\beta_i(M) = 0$, for $2 \leq i \leq n-2$ or $\widetilde{M}$ is a compact symmetric space, and in particular a homogeneous space. A theorem of Podestà states (see [**16**]) that a compact homogeneous space that is also a cohomogeneity one manifold with isotropy-irreducible principal orbits is isometric to the sphere or to the real projective space which implies again that $\beta_i(M) = 0$, for $2 \leq i \leq n-2$.

Theorem 4.4 above generalizes in some sense a result of Podestà in [**17**], which states that a compact G-cohomogeneity one manifold of positive Ricci curvature and isotropy-irreducible principal orbits is covered by a manifold conformally difeomorphic to a sphere.

For manifolds with harmonic curvature we obtain the following result.

Theorem 4.5. *Let $M^n, n \geq 5$, be a compact locally irreducible Riemannian manifold with harmonic curvature and non-parallel Ricci tensor. Let us suppose that* Ric *has less than three distinct eigenvalues at any point of M and the eigenvalues of the Weyl operator satisfy $-\mathcal{W}+S/\big((n-1)(n-2)\big) \geq 0$. If M has 2-nonnegative Ricci operator then $\beta_i(M) = 0$, for $2 \leq i \leq n-2$.*

Proof. It is immediate from Proposition 3.7 and Theorem 2, since we are assuming that the Ricci tensor is non-parallel and hence M cannot be locally symmetric.

5. Hypersurfaces of 2-nonnegative Ricci operator.

It is well-known that on hypersurfaces of Euclidean spaces, nonnegative Ricci curvature implies the nonnegativity of the sectional curvatures. The next result shows that compact hypersurfaces with 2-nonnegative Ricci operator and compact hypersurfaces with nonnegative isotropic curvatures are topologically the same. This is the content of Theorem 3 stated in the introduction that we now prove.

Proof of Theorem 3. Let ξ be a unit vector such that $\pm\xi$ are regular values of the Gauss map $\Phi : M^n \to S^n \subseteq \mathbf{R}^{n+1}$. Then the height function $h_\xi : M \to \mathbf{R}$ given by $h_\xi(x) = \langle f(x), \xi\rangle$ is a Morse fuction with critical points $\Phi^{-1}(\pm\xi)$. At such points the Hessian of h_ξ is given, up to a sign, by the Weingarten operator A_ξ. Let $\{e_1, \dots, e_n\}$ be an orthonormal basis of T_xM that diagonalizes A_ξ, say, $A_\xi e_i = \lambda_i e_i$. By the Gauss equation $K_{ij} = \lambda_i\lambda_j$ and since the critical points are nondegenerate, we have that $\lambda_i \neq 0$ for $i = 1, \dots, n$. As before, we denote the eigenvalues of the Ricci operator by μ_i. If all eigenvalues of the Ricci operator are nonnegative, then the sectional curvatures $K_{ij} \geq 0$ and all eigenvalues of A_ξ have the same sign.

Suppose $\mu_1 < 0$. We claim that in this case $n-1$ eigenvalues of A_ξ have the same sign.

If $\lambda_1 < 0$, we reorder the λ_i's for $i \geq 2$, such that $\lambda_2 \leq \cdots \leq \lambda_n$. Thus

$$\mu_1 = \lambda_1\,(\lambda_2 + \lambda_3 + \cdots + \lambda_n) < 0 \quad \Rightarrow \quad \lambda_2 + \lambda_3 + \cdots + \lambda_n > 0.$$

Therefore $\lambda_n > 0$. Now we suppose that $\lambda_2 < 0$ and this will give a contradiction. Indeed, since our hypothesis implies $\mu_2 \geq 0$ we have

$$\mu_2 = \lambda_2\,(\lambda_1 + \lambda_3 + \cdots + \lambda_n) > 0 \quad \Rightarrow \quad \lambda_1 + \lambda_3 + \cdots + \lambda_n < 0,$$

which in turn implies

$$\lambda_3 + \cdots + \lambda_n < -\lambda_1. \tag{5.1}$$

Since $\mu_n \geq 0$ and $\lambda_n > 0$ we also have

$$\mu_n = \lambda_2\lambda_n + \lambda_n\,(\lambda_1 + \lambda_3 + \cdots + \lambda_{n-1}) > 0 \quad \Rightarrow \quad \lambda_1 + \lambda_3 + \cdots + \lambda_{n-1} > 0,$$

yielding

$$\lambda_3 + \cdots + \lambda_{n-1} > -\lambda_1. \tag{5.2}$$

From (5.1) and (5.2) we get $-\lambda_1 + \lambda_n < -\lambda_1$ implying that $\lambda_n < 0$ and this is a contradiction.

If $\lambda_1 > 0$ we then have that $\lambda_2 + \lambda_3 + \cdots + \lambda_n < 0$. Since $\mu_1 < 0$ not all λ_i's have the same sign, otherwise all sectional curvatures would be positive. Let us the suppose that $\lambda_2 < 0$, after we have reordered such that $\lambda_2 \leq \cdots \leq \lambda_n$. If $\lambda_i < 0$ for $i \geq 3$ we have the claim. If not, then $\lambda_n > 0$. Again we obtain (5.1) and (5.2) which gives the desired contradiction.

Therefore we conclude that for each regular point, all but at most one of the $\lambda_i's$ have the same sign and hence the index of a critical point of h_ξ has to be $0, 1, n-1$ or n. By the standard Morse Theory, M has the homotopy type of a CW-complex, with no cells of dimension i for $2 \leq i \leq n-2$. Therefore the homology group $H_i(M; \mathbf{Z}) = 0$ for $2 \leq i \leq n-2$. Moreover, since there are no 2-cells ($n \geq 4$), we conclude by the cellular aproximation theorem that the inclusion of the 1-skeleton $M^{(1)} \hookrightarrow M$ induces an isomorphism between the fundamental groups. Therefore the fundamental group $\pi_1(M)$ is a free group on β_1 elements and $H_1(M; \mathbf{Z})$ is a free abelian group with the same number of generators.

Acknowledgements. The results of this article are extensions of some results in the first author's Ph.D. dissertation, which was directed by the second author. The second author wants to thank the Department of Mathematics of UNICAMP in Campinas, Brasil, for their hospitality and both authors would like to thank Professor Francesco Mercuri for helpful and valuable discussions.

References

[1] A.C. Asperti, F. Mercuri and M.H. Noronha, *Cohomogeneity one manifolds and hypersurfaces of revolution*, Bolletino U.M.I, **11** (1997), 199-215, MR 98c:53061, Zbl 0882.53006.

[2] M. Berger, *Sur les groupes d'holonomie des variétés a connexion affine et des variétés riemannienes*, Bull. Soc. Math. France, **83** (1955), 279-310, MR 18,149a, Zbl 0068.36002.

[3] ———, *Sur les groupes d'holonomie homogenes des variétés riemannienes*, C.R. Acad. Sci. Paris serie A, **262** (1966), 1316.

[4] A. Besse, *Einstein Manifolds*, Ergeb. Math. Grenzgeb (3), **10**, Springer-Verlag, Berlin, 1987, MR 88f:53087, Zbl 0613.53001.

[5] R.B. Brown and A. Gray, *Riemannian manifolds with holonomy group immersed* Spin (9), Differential Geometry (in honor of K. Yano), Kinokuniya, Tokyo, (1972), 41-59, MR 48 #7159, Zbl 0245.53020.

[6] H. Chen, *Manifolds with* 2*-nonnegative curvature operator*, Proceed. of Symposia in Pure Math., **54** (1993), 129-133, MR 94f:53063, Zbl 0796.53040.

[7] A. Derdzinski, *Classification of certain compact Riemannian manifolds with harmonic curvature and non-parallel Ricci tensor*, Math. Z., **172** (1980), 273-280, MR 82e:53053, Zbl 0453.53037.

[8] ———, *On compact Riemannian manifolds with harmonic curvature*, Math. Ann., **259** (1982), 145-152, MR 83i:53070, Zbl 0489.53042.

[9] S. Gallot and D. Meyer, *Opérateur de courbure et Laplacien des formes différentelles d'une variété Riemanniene*, J. Math. Pures et Appl., **54** (1975), 285-304, MR 56 #13128, Zbl 0316.53036.

[10] F. Mercuri and M.H. Noronha, *Low codimensional submanifolds of euclidean space with nonnegative isotropic curvature*, Trans. A.M.S., **348** (1996), 2711-2724, MR 96j:53049, Zbl 0862.53003.

[11] M. Micallef and J.D. Moore, *Minimal two-spheres and the topology of manifolds with positive curvature on totally isotropic two-planes*, Ann. of Math., **127** (1988), 199-227, MR 89e:53088, Zbl 0661.53027.

[12] M. Micallef and M.Y. Wang, *Metrics with nonnegative isotropic curvature*, Duke Math. J., **72** (1993), 649-672, MR 94k:53052, Zbl 0804.53058.

[13] M.H. Noronha, *Some compact conformally flat manifolds with nonnegative scalar curvature*, Geometriae Dedicata, **47** (1993), 255-268, MR 94f:53068, Zbl 0792.53035.

[14] ———, *Self-duality and four-manifolds with nonnegative curvature on totally isotropic two-planes*, Michigan Math. J., **41** (1994), 3-12, MR 95e:53069, Zbl 0816.53023.

[15] ———, *Positively curved* 4*-manifolds and the nonnegativity of isotropic curvatures*, Michigan Math. J., **44** (1997), 211-229, MR 98d:53055, Zbl 0888.53029.

[16] F. Podestà, *Cohomogeneity one Riemannian manifolds and Killing fields*, Diff. Geom. Appl., **5** (1995), 311-320, MR 96k:53072, Zbl 0846.53029.

[17] ———, *Immersions of cohomogeneity one Riemannian manifolds*, Monatsh. Math., **122** (1996), 215-225, MR 98a:53067, Zbl 0880.53040.

Received September 29, 2000 and revised March 26, 2001. The first author was partially supported by CNPq, Brasil. The second author was partially supported by FAPESP, Brasil.

IMECC-UNICAMP
Universidade Estadual de Campinas
13081-970, Campinas
SP, Brasil
E-mail address: dussan@ime.unicamp.br

Department of Mathematics
California State University Northridge
Northridge, CA, 91330-8313
E-mail address: maria.noronha@csun.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

DISCRETE PRODUCT SYSTEMS OF HILBERT BIMODULES

Neal J. Fowler

A Hilbert bimodule is a right Hilbert module X over a C^*-algebra A together with a left action of A as adjointable operators on X. We consider families $X = \{X_s : s \in P\}$ of Hilbert bimodules, indexed by a semigroup P, which are endowed with a multiplication which implements isomorphisms $X_s \otimes_A X_t \to X_{st}$; such a family is a called a product system. We define a generalized Cuntz-Pimsner algebra $\mathcal{O}_X$, and we show that every twisted crossed product of A by P can be realized as $\mathcal{O}_X$ for a suitable product system X. Assuming P is quasi-lattice ordered in the sense of Nica, we analyze a certain Toeplitz extension $\mathcal{T}_{\mathrm{cv}}(X)$ of $\mathcal{O}_X$ by embedding it in a crossed product $B_P \rtimes_{\tau,X} P$ which has been "twisted" by X; our main Theorem is a characterization of the faithful representations of $B_P \rtimes_{\tau,X} P$.

Introduction.

Suppose X is a right Hilbert module over a C^*-algebra A. If X also carries a left action of A as adjointable operators on X_A, we call X a *Hilbert bimodule* over A. In [**22**], Pimsner associated with every such bimodule X a C^*-algebra $\mathcal{O}_X$, which we shall call the *Cuntz-Pimsner algebra* of X, and showed that every crossed product by $\mathbb{Z}$ and every Cuntz-Krieger algebra can be realized as $\mathcal{O}_X$ for suitable X. He also commented that the algebras $\mathcal{O}_X$ include the crossed products by $\mathbb{N}$; that is, for each endomorphism α of a C^*-algebra A there is a bimodule $X = X(\alpha)$ such that $\mathcal{O}_X$ is canonically isomorphic to the semigroup crossed product $A \rtimes_\alpha \mathbb{N}$ of [**6**, **24**].

The work in this paper is motivated by the following observation, which also serves as our primary example. Suppose β is an action of a discrete semigroup P as endomorphisms of a C^*-algebra A. For each $s \in P$ let $X_s := X(\beta_s)$ be the bimodule canonically associated with the endomorphism β_s. Then the family $X = \{X_s : s \in P\}$ admits an associative multiplication $(x, y) \in X_s \times X_t \mapsto xy \in X_{ts}$ which implements isomorphisms $X_s \otimes_A X_t \to X_{ts}$; we call a family with this structure a *product system of Hilbert bimodules*. (In this example X is a product system over the opposite semigroup P^o.) Such families generalize the product systems

of [**7**, **8**, **12**, **10**], where the fibers X_s are complex Hilbert spaces (bimodules over $\mathbb{C}$).

To each product system X we associate a generalized Cuntz-Pimsner algebra $\mathcal{O}_X$. When X is the product system associated with the semigroup dynamical system (A, P, β), $\mathcal{O}_X$ is canonically isomorphic to the semigroup crossed product $A \rtimes_\beta P$. Moreover, if we "twist" X by a multiplier $\omega : P \times P \to \mathbb{T}$, then the corresponding Cuntz-Pimsner algebra is isomorphic to the twisted semigroup crossed product $A \rtimes_{\beta,\omega} P$. Our construction applies even when A is nonunital provided each endomorphism β_s extends to the multiplier algebra $M(A)$.

The aim of this paper is to take a first step towards analyzing the Cuntz-Pimsner algebra of a product system X. Following Pimsner [**22**], we begin by studying the structure of its Toeplitz extension $\mathcal{T}_X$. This algebra is universal for *Toeplitz representations* of X; these are multiplicative maps whose restriction to each fiber X_s is a Toeplitz representation in the sense of [**13**]. Our results generalize those of [**12**] for product systems of Hilbert spaces; indeed, much of the paper is devoted to adapting the methods of [**12**] to the bimodule setting. Thus our basic assumptions about the underlying semigroup P are as in [**12**]: To allow our analysis to extend beyond the totally-ordered case, we assume that P is the positive cone of a group G such that (G, P) is quasi-lattice ordered in the sense of Nica [**20**]. The class of such (G, P) includes all direct sums and free products of totally ordered groups. We also impose a covariance condition, called *Nica covariance*, on Toeplitz representations of X. This means that the universal C^*-algebra $\mathcal{T}_{\mathrm{cov}}(X)$ which we analyze is in general a quotient of $\mathcal{T}_X$. However, if (G, P) is totally-ordered, then Nica-covariance is automatic, and hence $\mathcal{T}_{\mathrm{cov}}(X)$ is the same as $\mathcal{T}_X$.

Our main goal is to characterize the faithful representations of $\mathcal{T}_{\mathrm{cov}}(X)$. We accomplish this by embedding $\mathcal{T}_{\mathrm{cov}}(X)$ in a certain twisted semigroup crossed product $B_P \rtimes_{\tau,X} P$ (Theorem 6.3), and then characterizing its faithful representations (Theorem 7.2). When $P = \mathbb{N}$, $\mathcal{T}_{\mathrm{cov}}(X)$ is precisely the Toeplitz algebra of the Hilbert bimodule X_1 (the fiber over $1 \in \mathbb{N}$), and our Theorem 7.2 reduces to [**13**, Theorem 2.1]. In fact, the analysis in [**13**] was motivated by our preliminary work on this paper. We would like to point out in particular how the stronger result [**13**, Theorem 3.1] arose from our investigations into product systems, for it serves as a good illustration of the usefulness of Nica covariance. Suppose Z is an orthogonal direct sum $\bigoplus_{\lambda\in\Lambda} Z^\lambda$ of Hilbert bimodules. Let G be the free group on Λ, let P be the subsemigroup of G generated by Λ, and let X be the unique product system over P whose fiber over λ is Z^λ. Then $\mathcal{T}_{\mathrm{cov}}(X)$ is canonically isomorphic to the Toeplitz algebra of the bimodule Z, and [**13**, Theorem 3.1] follows from our Theorem 7.2.

The main application of [**13**, Theorem 3.1] was to establish the simplicity of the graph algebras associated with certain infinite directed graphs [**13**, Corollary 4.3]. Although here we confine our applications to twisted semigroup crossed products, we anticipate that our results will also give interesting information about $\mathcal{O}_X$ when each of the fibers of X arise from infinite directed graphs.

We begin in Section 1 by giving a brief review of Hilbert bimodules, their representations, and their C^*-algebras. In Section 2 we introduce product systems of Hilbert bimodules, discuss their representations, and define the algebras $\mathcal{T}_X$ and $\mathcal{O}_X$. In Section 3 we associate with each twisted semigroup dynamical system (A, P, β, ω) a product system $X = X(A, P, \beta, \omega)$ whose Cuntz-Pimsner algebra $\mathcal{O}_X$ is the twisted semigroup crossed product $A \rtimes_{\beta,\omega} P$. We show that the Toeplitz algebra of $X(A, P, \beta, \omega)$ also has a crossed product structure, and this motivates the definition of a "Toeplitz" crossed product $\mathcal{T}(A \rtimes_{\beta,\omega} P)$ in which the endomorphisms are implemented not by isometries, but rather by partial isometries.

In Section 4 we generalize the notion of twisted crossed product by replacing the multiplier ω by a product system X of Hilbert bimodules. This extends the philosophy developed in [**12**] that one should regard product systems as noncommutative cocycles. Hence given an action β of P as endomorphisms of a C^*-algebra C, we consider (C, P, β, X) as a twisted semigroup dynamical system, and we define a twisted crossed product $C \rtimes_{\beta,X} P$.

In Section 5 we assume that (G, P) is quasi-lattice ordered, and we discuss the notion of Nica covariance for a Toeplitz representation. As illustrated in [**10**, Example 1.3] using product systems of Hilbert spaces, when (G, P) is not a total order it is possible that the C^*-algebra $\mathcal{T}_{\mathrm{cov}}(X)$ which is "universal" for such representations may admit representations which are not the integrated form of a Nica-covariant Toeplitz representation. To avoid this pathology we adapt the methods of [**10**] to our setting: We define the notion of a product system being *compactly aligned*, and show that $\mathcal{T}_{\mathrm{cov}}(X)$ is truly universal when X is compactly aligned (Proposition 5.9). We show that X is compactly aligned if the left action of A on each fiber X_s is by compact operators (Proposition 5.8); it follows that the product systems $X(A, P, \beta, \omega)$ associated with twisted semigroup dynamical systems are compactly aligned.

In Section 6 we consider a certain C^*-subalgebra B_P of $\ell^\infty(P)$ which is invariant under left translation $\tau : P \to \mathrm{End}(\ell^\infty(P))$. As in [**17**, **12**], covariant representations of the twisted system (B_P, P, τ, X) are in one-one correspondence with Toeplitz representations of X which are Nica-covariant (Proposition 6.1), and hence $\mathcal{T}_{\mathrm{cov}}(X)$ embeds naturally as a subalgebra of $B_P \rtimes_{\tau,X} P$ (Theorem 6.3). When the left action of A on each fiber X_s is by compact operators, $\mathcal{T}_{\mathrm{cov}}(X)$ is all of $B_P \rtimes_{\tau,X} P$.

In Section 7 we prove our main result, Theorem 7.2, which characterizes the faithful representations of $B_P \rtimes_{\tau,X} P$ under the assumption that

X is compactly aligned and (B_P, P, τ, X) satisfies a certain amenability hypothesis. In Section 8 we give conditions on (G, P) which ensure that (B_P, P, τ, X) is amenable. In particular, (B_P, P, τ, X) is amenable if X is compactly aligned and (G, P) is a free product $*(G^\lambda, P^\lambda)$ with each G^λ amenable (Corollary 8.2).

Finally, in Section 9 we apply our Theorem 7.2 to the product system $X(A, P, \beta, \omega)$ of Section 3. When (G, P) is a total order, $B_P \rtimes_{\tau,X} P$ is isomorphic to the Toeplitz crossed product $\mathcal{T}(A \rtimes_{\beta,\omega} P)$; in general $B_P \rtimes_{\tau,X} P$ is a certain quotient $\mathcal{T}_{\rm cov}(A \rtimes_{\beta,\omega} P)$ which also has a crossed product structure, and Theorem 9.3 characterizes its faithful representations. Applying this to the twisted system (B_P, P, τ, ω), we show that $\mathcal{T}_{\rm cov}(B_P \rtimes_{\tau,\omega} P)$ is universal for partial isometric representations of P which are *bicovariant* (Proposition 9.6), and we obtain a characterization of its faithful representations (Theorem 9.7) which is particularly nice when P is the free semigroup on infinitely many generators (Theorem 9.9).

The author thanks Iain Raeburn for the many helpful discussions while this research was being conducted.

1. Preliminaries.

Let A be a separable C^*-algebra. A *Hilbert bimodule over A* is a right Hilbert A-module X together with a *-homomorphism $\phi : A \to \mathcal{L}(X)$ which is used to define a left action of A on X via $a \cdot x := \phi(a)x$ for $a \in A$ and $x \in X$. A *Toeplitz representation* of X in a C^*-algebra B is a pair (ψ, π) consisting of a linear map $\psi : X \to B$ and a homomorphism $\pi : A \to B$ such that

$$\begin{aligned} \psi(x \cdot a) &= \psi(x)\pi(a), \\ \psi(x)^*\psi(y) &= \pi(\langle x, y\rangle_A), \quad \text{and} \\ \psi(a \cdot x) &= \pi(a)\psi(x) \end{aligned}$$

for $x, y \in X$ and $a \in A$. Given such a representation, there is homomorphism $\pi^{(1)} : \mathcal{K}(X) \to B$ which satisfies

$$\pi^{(1)}(\Theta_{x,y}) = \psi(x)\psi(y)^* \qquad \text{for all } x, y \in X, \tag{1.1}$$

where $\Theta_{x,y}(z) := x \cdot \langle y, z\rangle_A$ for $z \in X$; see [**22**, p. 202], [**16**, Lemma 2.2], and [**13**, Remark 1.7] for details. We say that (ψ, π) is *Cuntz-Pimsner covariant* if

$$\pi^{(1)}(\phi(a)) = \pi(a) \qquad \text{for all } a \in \phi^{-1}(\mathcal{K}(X)).$$

The *Toeplitz algebra* of X is the C^*-algebra $\mathcal{T}_X$ which is universal for Toeplitz representations of X [**22**, **13**], and the *Cuntz-Pimsner algebra* of X is the C^*-algebra $\mathcal{O}_X$ which is universal for Toeplitz representations which are Cuntz-Pimsner covariant [**22**, **9**, **16**, **18**, **19**, **11**].

Every right Hilbert A-module X is essential, in the sense that X is the closed linear span of elements $x \cdot a$. We say that a Hilbert bimodule X is *essential* if it is also essential as a left A-module; that is, if

$$X = \overline{\text{span}}\{\phi(a)x : a \in A, x \in X\}.$$

When X is essential, two applications of the Hewitt-Cohen Factorization Theorem allow us to write any $x \in X$ as $\phi(a)y \cdot b$ for some $y \in X$ and $a, b \in A$. Hence if (a_i) is an approximate identity for A, then

$$\|x - x \cdot a_i\| \to 0 \quad \text{and} \quad \|x - \phi(a_i)x\| \to 0 \qquad \text{for all } x \in X. \tag{1.2}$$

2. Product systems of Hilbert bimodules.

For each $n \geq 1$ the n-fold internal tensor product $X^{\otimes n} := X \otimes_A \cdots \otimes_A X$ has a natural structure as a Hilbert bimodule over A; see [**19**, Section 2.2] for details. The following definition, based on Arveson's continuous tensor product systems over $(0, \infty)$ [**3**], generalizes the collection $\{X^{\otimes n} : n \in \mathbb{N}\}$ to semigroups other than $\mathbb{N}$.

Definition 2.1. Suppose P is a countable semigroup with identity e and $p : X \to P$ is a family of Hilbert bimodules over A. Write X_s for the fibre $p^{-1}(s)$ over $s \in P$, and write $\phi_s : A \to \mathcal{L}(X_s)$ for the homomorphism which defines the left action of A on X_s. We say that X is a (*discrete*) *product system over* P if X is a semigroup, p is a semigroup homomorphism, and for each $s, t \in P \setminus \{e\}$ the map $(x, y) \in X_s \times X_t \mapsto xy \in X_{st}$ extends to an isomorphism of the Hilbert bimodules $X_s \otimes_A X_t$ and X_{st}. We also require that $X_e = A$ (with its usual right Hilbert module structure and $\phi_e(a)b = ab$ for $a, b \in A$), and that the multiplications $X_e \times X_s \to X_s$ and $X_s \times X_e \to X_s$ satisfy

$$ax = \phi_s(a)x, \qquad xa = x \cdot a \qquad \text{for } a \in X_e \text{ and } x \in X_s. \tag{2.1}$$

Remark 2.2. Multiplication $X_e \times X_s \to X_s$ will not induce an isomorphism $X_e \otimes_A X_s \to X_s$ unless X_s is essential as a left A-module.

Remark 2.3. The associativity of multiplication in X implies that $\phi_{st}(a) = \phi_s(a) \otimes_A 1^t$ for all $a \in A$; that is, $\phi_{st}(a)(xy) = (\phi_s(a)x)y$ for all $x \in X_s$ and $y \in X_t$.

Remark 2.4. It is possible that some of the X_s may be zero.

Definition 2.5. Suppose B is a C^*-algebra and $\psi : X \to B$; write ψ_s for the restriction of ψ to X_s. We call ψ a *Toeplitz representation* of X if:

(1) For each $s \in P$, (ψ_s, ψ_e) is a Toeplitz representation of X_s; and
(2) $\psi(xy) = \psi(x)\psi(y)$ for $x, y \in X$.

If in addition each (ψ_s, ψ_e) is Cuntz-Pimsner covariant, we say that ψ is *Cuntz-Pimsner covariant*.

Remark 2.6. By [**13**, Remark 1.1], every Toeplitz representation ψ is contractive; moreover, if the homomorphism $\psi_e : A \to B$ is isometric, then so is ψ. Also, since we are assuming (2.1), a map $\psi : X \to B$ is a Toeplitz representation if it satisfies both (2) and

(1′) $\psi_s(x)^*\psi_s(y) = \psi_e(\langle x, y\rangle_A)$ whenever $s \in P$ and $x, y \in X_s$.

Notation 2.7. We write $\psi^{(s)}$ for the homomorphism of $\mathcal{K}(X_s)$ into B which corresponds to the pair (ψ_s, ψ_e), as in (1.1); that is,

$$\psi^{(s)}(\Theta_{x,y}) = \psi_s(x)\psi_s(y)^* \qquad \text{for all } x, y \in X_s.$$

The Fock representation. Let $F(X)$ be the right Hilbert A-module

$$F(X) := \bigoplus_{s\in P} X_s.$$

By this we mean the following: As a set, $F(X)$ is the subset of $\prod_{s\in P} X_s$ consisting of all elements (x_s) for which $\sum_{s\in P}\langle x_s, x_s\rangle_A$ is summable in A; that is, for which $\sum_{s\in F}\langle x_s, x_s\rangle_A$ converges in norm as F increases over the finite subsets of P. We write $\oplus x_s$ for (x_s) to indicate that the above series is summable. The right action of A is given by $(\oplus x_s)\cdot a := \oplus(x_s \cdot a)$, and the inner product by $\langle \oplus x_s, \oplus y_s\rangle_A := \sum_{s\in P}\langle x_s, y_s\rangle_A$. The algebraic direct sum $\bigodot_{s\in P} X_s$ is dense in $F(X)$.

Suppose P is left-cancellative. Then for any $x \in X$ and $\oplus x_t \in F(X)$ we have $p(xx_s) = p(xx_t)$ if and only if $s = t$, so there is an element $(y_s) \in \prod X_s$ such that

$$y_s = \begin{cases} xx_t & \text{if } s = p(x)t \\ 0 & \text{if } s \notin p(x)P; \end{cases}$$

we write (xx_t) for (y_s). Since $\langle xx_s, xx_s\rangle_A \le \|x\|^2\langle x_s, x_s\rangle_A$ for each $s \in P$, the series $\sum\langle xx_s, xx_s\rangle_A$ is summable. It is routine to check that

$$l(x)(\oplus x_s) := \oplus xx_s \qquad \text{for } \oplus x_s \in F(X)$$

determines an adjointable operator $l(x)$ on $F(X)$; indeed, the adjoint $l(x)^*$ is zero on any summand X_s for which $s \notin p(x)P$, and on $X_{p(x)t} = \overline{\operatorname{span}}\, X_{p(x)}X_t$ it is determined by the formula $l(x)^*(yz) = \langle x, y\rangle_A \cdot z$ for $y \in X_{p(x)}$ and $z \in X_t$. It follows that $l : X \to \mathcal{L}(F(X))$ is a Toeplitz representation, called the *Fock representation of* X. The homomorphism $l_e : A \to \mathcal{L}(F(X))$ is simply the diagonal left action of A; that is, $l_e(a) = \oplus\phi_s(a)$. Since ϕ_e is just left multiplication on $X_e = A$, it is isometric, and hence so is l_e; by Remark 2.6, l is isometric.

Proposition 2.8. *Let X be a product system over P of Hilbert A–A bimodules. Then there is a C^*-algebra $\mathcal{T}_X$, called the* Toeplitz algebra *of X, and a Toeplitz representation $i_X : X \to \mathcal{T}_X$, such that*

(a) *for every Toeplitz representation ψ of X, there is a homomorphism ψ_* of $\mathcal{T}_X$ such that $\psi_* \circ i_X = \psi$; and*
(b) *$\mathcal{T}_X$ is generated as a C^*-algebra by $i_X(X)$.*

The pair $(\mathcal{T}_X, i_X)$ is unique up to canonical isomorphism, and i_X is isometric.

Proof. It is straightforward to translate the proof of [**13**, Proposition 1.3] to this setting. □

Proposition 2.9. *Let X be a product system over P of Hilbert A–A bimodules. Then there is a C^*-algebra $\mathcal{O}_X$, called the* Cuntz-Pimsner algebra *of X, and a Toeplitz representation $j_X : X \to \mathcal{O}_X$ which is Cuntz-Pimsner covariant, such that*

(a) *for every Cuntz-Pimsner covariant Toeplitz representation ψ of X, there is a homomorphism ψ_* of $\mathcal{O}_X$ such that $\psi_* \circ j_X = \psi$; and*
(b) *$\mathcal{O}_X$ is generated as a C^*-algebra by $j_X(X)$.*

The pair $(\mathcal{O}_X, j_X)$ is unique up to canonical isomorphism.

Remark 2.10. Although the universal map $i_X : X \to \mathcal{T}_X$ is always isometric, it is quite possible that X might not admit any nontrivial Toeplitz representations which are Cuntz-Pimsner covariant, in which case $\mathcal{O}_X$ is trivial.

Proof of Proposition 2.9. With $(\mathcal{T}_X, i_X)$ as in Proposition 2.8, let $\mathcal{I}$ be the ideal in $\mathcal{T}_X$ generated by

$$\{i_X(a) - i_X^{(s)}(\phi_s(a)) : s \in P, a \in \phi_s^{-1}(\mathcal{K}(X))\}.$$

Define $\mathcal{O}_X := \mathcal{T}_X/\mathcal{I}$ and $j_X := q \circ i_X$, where $q : \mathcal{T}_X \to \mathcal{O}_X$ is the canonical projection. Obviously j_X is a Toeplitz representation which generates $\mathcal{O}_X$, and it is Cuntz-Pimsner covariant because $j_X^{(s)} = q \circ i_X^{(s)}$. If ψ is another Cuntz-Pimsner covariant Toeplitz representation, then the homomorphism ψ_* of $\mathcal{T}_X$ satisfies

$$\psi_*(i_X(a) - i_X^{(s)}(\phi_s(a))) = \psi(a) - \psi^{(s)}(\phi_s(a)) = 0$$

whenever $\phi_s(a) \in \mathcal{K}(X_s)$, and hence ψ_* descends to the required homomorphism of $\mathcal{O}_X$ (also denoted ψ_*). □

Proposition 2.11. *Let X be a product system over $\mathbb{N}$ of Hilbert A–A bimodules. Then $\mathcal{T}_X$ is canonically isomorphic to the Toeplitz algebra $\mathcal{T}_{X_1}$ of the Hilbert bimodule X_1. If the left action on each fiber is isometric, or if the left action on each fiber is by compact operators, then $\mathcal{O}_X$ is canonically isomorphic to $\mathcal{O}_{X_1}$.*

Proof. Let $i_X : X \to \mathcal{T}_X$ be universal for Toeplitz representations of X, and define $\mu := (i_X)_1 : X_1 \to \mathcal{T}_X$ and $\pi := (i_X)_0 : A = X_0 \to \mathcal{T}_X$. Since (μ, π) is a Toeplitz representation of X_1, we get a homomorphism $\mu \times \pi : \mathcal{T}_{X_1} \to \mathcal{T}_X$.

To construct the inverse of $\mu \times \pi$, let (i_{X_1}, i_A) be the universal Toeplitz representation of X_1 in $\mathcal{T}_{X_1}$, and fix $n \geq 1$. By [**13**, Proposition 1.8(1)], there is a linear map $\psi_n : X_n \to \mathcal{T}_{X_1}$ which satisfies

$$\psi_n(x_1 \cdots x_n) := i_{X_1}(x_1) \cdots i_{X_1}(x_n) \qquad \text{for all } x_1, \ldots, x_n \in X_1,$$

and then (ψ_n, i_A) is a Toeplitz representation of X_n. Defining $\psi_0 := i_A$ thus gives a Toeplitz representation $\psi : X \to \mathcal{T}_{X_1}$, and it is routine to check that $\psi_* : \mathcal{T}_X \to \mathcal{T}_{X_1}$ is the inverse of $\mu \times \pi$.

Now let $i_X : X \to \mathcal{O}_X$ be universal for Cuntz-Pimsner covariant Toeplitz representations of X. As above, we get a homomorphism $\mu \times \pi : \mathcal{O}_{X_1} \to \mathcal{O}_X$. To construct the inverse, we let $(i_{X_1}, i_A) : (X_1, A) \to \mathcal{O}_{X_1}$ be universal and define a Toeplitz representation $\psi : X \to \mathcal{O}_{X_1}$ as before; we need to check that ψ is Cuntz-Pimsner covariant under each of the hypotheses on the left action. By definition (ψ_1, ψ_0) is Cuntz-Pimsner covariant, so we use induction. Assume that (ψ_n, ψ_0) is Cuntz-Pimsner covariant for some $n \geq 1$, and suppose $a \in A$ acts compactly on the left of X_{n+1}; that is, $\phi(a) \otimes_A 1^n \in \mathcal{K}(X_{n+1})$. Since the left action is isometric on each fiber, by [**11**, Lemma 4.2] we have $\phi(a) \otimes_A 1^{n-1} \in \mathcal{K}(X_n)$; hence $\psi^{(n)}(\phi(a) \otimes_A 1^{n-1}) = \psi_0(a)$. But [**11**, Lemma 4.5] gives $\psi^{(n+1)}(\phi(a) \otimes_A 1^n) = \psi^{(n)}(\phi(a) \otimes_A 1^{n-1})$, so (ψ_{n+1}, ψ_0) is Cuntz-Pimsner covariant.

Now suppose that A acts by compact operators on each fiber. By representing $\mathcal{O}_{X_1}$ faithfully on a Hilbert space $\mathcal{H}$ we can assume that ψ is a Toeplitz representation of X on $\mathcal{H}$. Assuming again that (ψ_n, ψ_0) is Cuntz-Pimsner covariant for some $n \geq 1$, [**11**, Lemma 1.9] gives $\overline{\psi_0(A)\mathcal{H}} \subseteq \overline{\operatorname{span}}(\psi_n(X_n)\mathcal{H})$. Let $x \in X_n$, and express $x = y \cdot a$ with $y \in X_n$ and $a \in A$. Since (ψ_1, ψ_0) is Cuntz-Pimsner covariant and $\phi(a) \in \mathcal{K}(X_1)$, we have

$$\psi_n(x) = \psi_n(y)\psi_0(a) = \psi_n(y)\psi^{(1)}(\phi(a)).$$

Now $\phi(a)$ can be approximated by a finite sum $\sum \Theta_{x_i, y_i}$, hence $\psi_n(x)$ can be approximated by a finite sum $\psi_n(y)\psi(x_i)\psi(y_i)^* = \psi_{n+1}(yx_i)\psi(y_i)^*$. Thus

$$\overline{\psi_0(A)\mathcal{H}} \subseteq \overline{\operatorname{span}}(\psi_n(X_n)\mathcal{H}) \subseteq \overline{\operatorname{span}}(\psi_{n+1}(X_{n+1})\mathcal{H}),$$

and (ψ_{n+1}, ψ_0) is Cuntz-Pimsner covariant by [**11**, Lemma 1.9]. □

Definition 2.12. Let X be a product system over P of Hilbert A–A bimodules. A Toeplitz representation $\psi : X \to B$ is *nondegenerate* if the induced homomorphism $\psi_* : \mathcal{T}_X \to B$ is nondegenerate.

Lemma 2.13. *Suppose each fiber X_s is essential as a left A-module. Then a Toeplitz representation $\psi : X \to B$ is nondegenerate if and only if the homomorphism $\psi_e : A \to B$ is nondegenerate.*

Proof. Let (a_i) be an approximate identity for $A = X_e$. By (1.2), $i_X(a_i)$ is an approximate identity for $\mathcal{T}_X$, and the result follows. □

3. Crossed products twisted by multipliers.

Our main examples of product systems come from C^*-dynamical systems. Suppose β is an action of P as endomorphisms of A such that β_e is the identity endomorphism. We will assume that each β_s is *extendible*; that is, that each β_s extends to a strictly continuous endomorphism $\overline{\beta_s}$ of $M(A)$. For P the positive cone of a totally ordered abelian group, Adji has shown that extendibility is necessary to define a reasonable crossed product $A \rtimes_\beta P$ [**1**].

In this section we will consider crossed products which are twisted by a multiplier ω of P; that is, by a function $\omega : P \times P \to \mathbb{T}$ which satisfies $\omega(e, e) = 1$ and

$$\omega(r, s)\omega(rs, t) = \omega(r, st)\omega(s, t) \qquad \text{for all } r, s, t \in P.$$

We call (A, P, β, ω) a *twisted semigroup dynamical system*.

Definition 3.1. Let B be a C^*-algebra. A function $V : P \to B$ is called an *ω-representation* of P if

$$V_s V_t = \omega(s, t)V_{st} \qquad \text{for all } s, t \in P. \tag{3.1}$$

If in addition each V_s is an isometry (resp. partial isometry), V is called *isometric* (resp. *partial isometric*) ω-representation. A *covariant representation* of (A, P, β, ω) on a Hilbert space $\mathcal{H}$ is a pair (π, V) consisting of a nondegenerate representation $\pi : A \to B(\mathcal{H})$ and an isometric ω-representation $V : P \to B(\mathcal{H})$ such that

$$\pi(\beta_s(a)) = V_s \pi(a) V_s^* \qquad \text{for all } s \in P \text{ and } a \in A. \tag{3.2}$$

A *crossed product* for (A, P, β, ω) is a triple (B, i_A, i_P) consisting of a C^*-algebra B, a nondegenerate homomorphism $i_A : A \to B$, and a map $i_P : P \to M(B)$ such that

(a) if σ is a nondegenerate representation of B, then $(\sigma \circ i_A, \overline{\sigma} \circ i_P)$ is a covariant representation of (A, P, β, ω);
(b) for every covariant representation (π, V), there is a representation $\pi \times V$ such that $(\pi \times V) \circ i_A = \pi$ and $\overline{\pi \times V} \circ i_P = V$; and
(c) B is generated as a C^*-algebra by $\{i_A(a)i_P(s) : a \in A, s \in P\}$.

After establishing the existence of a crossed product, it is easily seen to be unique up to canonical isomorphism; we denote the crossed product algebra $A \rtimes_{\beta,\omega} P$.

We will construct a product system $X = X(A, P, \beta, \omega)$ over the opposite semigroup P^o, and show that its Cuntz-Pimsner algebra $\mathcal{O}_X$ is a crossed product for (A, P, β, ω). Moreover, we will show that the Toeplitz algebra of this product system also has a crossed product structure: It will be universal

for pairs (π, V) satisfying (3.2) in which π is a nondegenerate representation of A and V is a partial isometric ω-representation such that

$$V_s^* V_s \pi(a) = \pi(a) V_s^* V_s \text{ for all } s \in P \text{ and } a \in A. \tag{3.3}$$

We call such a pair (π, V) a *Toeplitz covariant representation* of (A, P, β, ω), and write $\mathcal{T}(A \rtimes_{\beta,\omega} P)$ for the corresponding universal C^*-algebra, called the *Toeplitz crossed product* of (A, P, β, ω).

For each $s \in P$ let

$$X_s := \{s\} \times \overline{\beta_s}(1)A,$$

and give X_s the structure of a Hilbert bimodule over A via

$$(s, x) \cdot a := (s, xa), \quad \langle (s, x), (s, y) \rangle_A := x^* y,$$

and

$$\phi_s(a)(s, x) := (s, \beta_s(a)x).$$

Let $X = \bigsqcup_{s \in P} X_s$, let $p(s, x) := s$, and define multiplication in X by

$$(s, x)(t, y) := (ts, \overline{\omega(t, s)}\beta_t(x)y) \qquad \text{for } x \in \overline{\beta_s}(1)A \text{ and } y \in \overline{\beta_t}(1)A.$$

Lemma 3.2. *$X = X(A, P, \beta, \omega)$ is a product system over the opposite semigroup P^o. For each $s \in P$, the fiber X_s is essential as a left A-module, and the left action of A on X_s is by compact operators.*

Proof. Let $(s, x) \in X_s$ and $(t, y) \in X_t$. If $x = \overline{\beta_s}(1)a$ and $y = \overline{\beta_t}(1)b$, then

$$\beta_t(x)y = \beta_t(\overline{\beta_s}(1)a)\overline{\beta_t}(1)b = \overline{\beta_t}(\overline{\beta_s}(1))\beta_t(a)b = \overline{\beta_{ts}}(1)\beta_t(a)b,$$

and hence the product $(s, x)(t, y)$ belongs to X_{ts}. Letting a vary over an approximate identity for A, this product converges in norm to $\overline{\beta_{ts}}(1)b$, so the set of products $(s, x)(t, y)$ has dense linear span in X_{ts}. Hence to see that multiplication induces an isomorphism $X_s \otimes_A X_t \to X_{ts}$, it suffices to check that it preserves the inner product of any pair of elementary tensors:

$$\begin{aligned}
\langle (s, x) \otimes_A (t, y), (s, x') \otimes_A (t, y') \rangle_A &= \langle (t, y), \phi_t(\langle (s, x), (s, x') \rangle_A)(t, y') \rangle_A \\
&= \langle (t, y), \phi_t(x^* x')(t, y') \rangle_A \\
&= \langle (t, y), (t, \beta_t(x^* x')y') \rangle_A \\
&= y^* \beta_t(x^* x') y' \\
&= \langle (ts, \overline{\omega(t, s)}\beta_t(x)y), (ts, \overline{\omega(t, s)}\beta_t(x')y') \rangle_A \\
&= \langle (s, x)(t, y), (s, x')(t, y') \rangle_A.
\end{aligned}$$

Multiplication is associative since

$$\begin{aligned}((s,x)(t,y))(r,z) &= (ts, \overline{\omega(t,s)}\beta_t(x)y)(r,z)\\ &= (rts, \overline{\omega(r,ts)}\beta_r(\overline{\omega(t,s)}\beta_t(x)y)z)\\ &= (rts, \overline{\omega(rt,s)}\beta_{rt}(x)\overline{\omega(r,t)}\beta_r(y)z)\\ &= (s,x)(rt, \overline{\omega(r,t)}\beta_r(y)z)\\ &= (s,x)((t,y)(r,z)).\end{aligned}$$

If (a_i) is an approximate identity for A, then for each $a \in A$ and $s \in P$ we have $\lim \phi(a_i)(s, \overline{\beta_s}(1)a) = \lim(s, \beta_s(a_i)a) = (s, \overline{\beta_s}(1)a)$, so each X_s is essential. If $a \in A$, then by writing $a = bc^*$ with $b, c \in A$, we see that $\phi_s(a) = \Theta_{(s,\beta_s(b)),(s,\beta_s(c))} \in \mathcal{K}(X_s)$ is compact. □

Lemma 3.3. *Let $i_X : X \to \mathcal{T}_X$ be universal for Toeplitz representations of X, and let (a_i) be an approximate identity for A. Then for each $s \in P$, $i_X(s, \beta_s(a_i))$ converges strictly in $M\mathcal{T}_X$.*

Proof. Since each fiber X_t is essential, any vector $\xi \in X_t$ can be written in the form $\xi = \phi_t(a)\eta \cdot b$ with $a, b \in A$ and $\eta \in X_t$. But then $i_X(\xi) = i_X(e,a)i_X(\eta)i_X(e,b)$, and since elements of the form $i_X(\xi)$ generate $\mathcal{T}_X$ as a C^*-algebra, the result follows from the calculations

$$i_X(s, \beta_s(a_i))i_X(e,a) = i_X(s, \beta_s(a_i)a) \to i_X(s, \overline{\beta_s}(1)a) \tag{3.4}$$

and

$$i_X(e,a)i_X(s, \beta_s(a_i)) = i_X(s, \beta_s(aa_i)) \to i_X(s, \beta_s(a)). \tag{3.5}$$

□

Define $i_A : A \to \mathcal{T}_X$ by $i_A(a) := i_X(e,a)$, and define $i_P : P \to M\mathcal{T}_X$ by $i_P(s) := \lim i_X(s, \beta_s(a_i))^*$.

Proposition 3.4. *$\mathcal{T}_X$ and $\mathcal{O}_X$ are canonically isomorphic to $A \rtimes_{\beta,\omega} P$ and $\mathcal{T}(A \rtimes_{\beta,\omega} P)$, respectively. More precisely, $(\mathcal{T}_X, i_A, i_P)$ is a Toeplitz crossed product for (A, P, β, ω), and, with $q : \mathcal{T}_X \to \mathcal{O}_X$ the canonical projection, $(\mathcal{O}_X, q \circ i_A, \overline{q} \circ i_P)$ is a crossed product for (A, P, β, ω).*

Proof. Taking $s = e$ in (3.4) and (3.5), shows that $i_A(a_i)$ converges strictly to the identity in $M(\mathcal{T}_X)$; that is, i_A is nondegenerate. For Condition (a) of a Toeplitz crossed product, suppose σ is a nondegenerate representation of $\mathcal{T}_X$; we must show that $(\pi, V) := (\sigma \circ i_A, \overline{\sigma} \circ i_P)$ is a Toeplitz covariant representation of (A, P, β, ω). First note that π is nondegenerate since σ and i_A are. Equation (3.5) gives

$$i_A(a)i_P(s)^* = i_X(s, \beta_s(a)) \qquad \text{for all } a \in A \text{ and } s \in P, \tag{3.6}$$

so

$$\begin{aligned} i_P(s)i_A(a)i_P(s)^* &= \lim i_X(s,\beta_s(a_i))^* i_X(s,\beta_s(a)) \\ &= \lim i_X(e,\beta_s(a_i^*a)) = i_X(e,\beta_s(a)) = i_A(\beta_s(a)), \end{aligned}$$

and applying $\overline{\sigma}$ gives $V_s\pi(a)V_s^* = \pi(\beta_s(a))$. In particular

$$i_P(s)i_P(s)^* = \lim i_P(s)i_A(a_i)i_P(s)^* = \lim i_A(\beta_s(a_i)) = \overline{i_A}(\overline{\beta_s}(1))$$

is a projection, so $i_P(s)$, and hence V_s, is a partial isometry.

To establish (3.3), take any $a \in A$, write $a = bc^*$ with $b, c \in A$, and compute:

$$\begin{aligned} i_A(bc^*)i_P(s)^*i_P(s) &= \lim i_X(s,\beta_s(bc^*))i_X(s,\beta_s(a_i))^* \qquad \text{(by (3.6))} \\ &= \lim i_X(s,\beta_s(b))i_X(e,\beta_s(c^*))i_X(s,\beta_s(a_i))^* \\ &= \lim i_X(s,\beta_s(b))\big(i_X(s,\beta_s(a_i))i_X(e,\beta_s(c))\big)^* \\ &= \lim i_X(s,\beta_s(b))i_X(s,\beta_s(a_ic))^* \\ &= i_X(s,\beta_s(b))i_X(s,\beta_s(c))^*. \end{aligned} \tag{3.7}$$

Taking adjoints, interchanging b and c, and applying $\overline{\sigma}$ gives $V_s^*V_s\pi(a) = \pi(a)V_s^*V_s$.

For every $s, t \in P$ we have

$$\begin{aligned} i_P(t)^*i_P(s)^* &= (\lim_i j_X(t,\beta_t(a_i)))(\lim_j j_X(s,\beta_s(a_j))) \\ &= \lim_i \lim_j j_X(st, \overline{\omega(s,t)}\beta_s(\beta_t(a_i))\beta_s(a_j)) \\ &= \lim_i j_X(st, \overline{\omega(s,t)}\beta_{st}(a_i)) \\ &= \overline{\omega(s,t)}i_P(st)^*; \end{aligned}$$

taking adjoints and applying $\overline{\sigma}$ gives $V_sV_t = \omega(s,t)V_{st}$. This completes the proof of Condition (a).

For Condition (b), suppose (π, V) is a Toeplitz covariant representation of (A, P, β, ω) on a Hilbert space $\mathcal{H}$. Define $\psi : X \to B(\mathcal{H})$ by

$$\psi(s,x) := V_s^*\pi(x).$$

Since π is nondegenerate and $\pi(a) = \pi(\beta_e(a)) = V_e\pi(a)V_e^*$ for all $a \in A$, V_e is a coisometry. Since $V_e^2 = \omega(e,e)V_e = V_e$, we deduce that $V_e = 1$. Thus

$$\begin{aligned} \psi(s,x)^*\psi(s,y) &= \pi(x)^*V_sV_s^*\pi(y) = \pi(x^*\overline{\beta_s}(1)y) \\ &= V_e^*\pi(x^*y) \qquad \text{(since } y \in \overline{\beta_s}(1)A \text{ and } V_e = 1) \\ &= \psi(e, x^*y) = \psi(e, \langle (s,x),(s,y)\rangle_A), \end{aligned}$$

and since we also have

$$\begin{aligned}
\psi(s,x)\psi(t,y) &= V_s^*\pi(x)V_t^*\pi(y) \\
&= V_s^*\pi(x)V_t^*V_tV_t^*\pi(y) \qquad (V_t \text{ is a partial isometry}) \\
&= V_s^*V_t^*V_t\pi(x)V_t^*\pi(y) \qquad \text{(by (3.3))} \\
&= (V_tV_s)^*\pi(\beta_t(x))\pi(y) \\
&= \overline{\omega(t,s)}V_{ts}^*\pi(\beta_t(x)y) \\
&= \overline{\omega(t,s)}\psi(ts,\beta_t(x)y) \\
&= \psi((s,x)(t,y)),
\end{aligned}$$

ψ is a Toeplitz representation of X. Let $\pi \times V$ be the representation $\psi_* : \mathcal{T}_X \to B(\mathcal{H})$. Then

$$(\pi \times V) \circ i_A(a) = \psi_* \circ i_X(e,a) = \psi(e,a) = V_e^*\pi(a) = \pi(a),$$

and

$$\begin{aligned}
\overline{\pi \times V} \circ i_P(s)\pi(a) &= \psi_*(i_P(s)i_A(a)) \\
&= \psi_*(i_X(s,\beta_s(a^*))^*) \qquad \text{(by (3.6))} \\
&= \psi(s,\beta_s(a^*))^* = \pi(\beta_s(a))V_s \\
&= V_s\pi(a)V_s^*V_s = V_sV_s^*V_s\pi(a) = V_s\pi(a);
\end{aligned}$$

since π is nondegenerate, this implies that $\overline{\pi \times V} \circ i_P = V$, as required. For Condition (c), simply note that $i_A(a)i_P(s) = i_X(s,\overline{\beta_s}(1)a^*)^*$, and elements of this form generate $\mathcal{T}_X$.

We now show that $(\mathcal{O}_X, q \circ i_A, \overline{q} \circ i_P)$ is a crossed product for (A,P,β,ω). Since i_A and q are nondegenerate, so is $q \circ i_A$. If ρ is a nondegenerate representation of $\mathcal{O}_X$, then $\sigma := \rho \circ q$ is a nondegenerate representation of $\mathcal{T}_X$. Hence $(\pi, V) := (\rho \circ q \circ i_A, \overline{\rho} \circ \overline{q} \circ i_P) = (\sigma \circ i_A, \overline{\sigma} \circ i_P)$ is a Toeplitz covariant representation of (A,P,β,ω). To see that each V_s is an isometry, let $b,c \in A$. Since $q \circ i_X$ is Cuntz-Pimsner covariant, (3.7) gives

$$\begin{aligned}
q \circ i_A(bc^*)\overline{q} \circ i_P(s)^*\overline{q} \circ i_P(s) &= q \circ i_X(s,\beta_s(b))q \circ i_X(s,\beta_s(c))^* \\
&= (q \circ i_X)^{(s)}(\Theta_{(s,\beta_s(b)),(s,\beta_s(c))}) \\
&= (q \circ i_X)^{(s)}(\phi_s(bc^*)) \\
&= q \circ i_X(e,bc^*) = q \circ i_A(bc^*).
\end{aligned}$$

Since $q \circ i_A$ is nondegenerate, this implies that $\overline{q} \circ i_P(s)$, and hence V_s, is an isometry. This gives Condition (a) for a crossed product. Condition (c) is obvious, so it remains only to verify (b). Suppose (π, V) is a covariant representation of (A,P,β,ω) on a Hilbert space $\mathcal{H}$, and define $\psi(s,x) := V_s^*\pi(x)$ as before. We have already seen that ψ is a Toeplitz representation

of X, and it is Cuntz-Pimsner covariant since, for any $b, c \in A$,

$$\begin{aligned}\psi^{(s)}(\phi_s(bc^*)) &= \psi^{(s)}(\Theta_{(s,\beta_s(b)),(s,\beta_s(c))}) = \psi(s,\beta_s(b))\psi(s,\beta_s(c))^* \\ &= V_s^*\pi(\beta_s(b))\pi(\beta_s(c^*))V_s = V_s^*V_s\pi(bc^*)V_s^*V_s = \psi(e,bc^*).\end{aligned}$$

Defining $\pi \times V := \psi_* : \mathcal{O}_X \to B(\mathcal{H})$ gives Condition (c). □

4. Crossed products twisted by product systems.

Multipliers of P correspond to product systems over P of one-dimensional Hilbert spaces: Given a multiplier ω, one defines multiplication on $P \times \mathbb{C}$ by $(s,z)(t,w) := (st, \omega(s,t)zw)$. In this section we consider twisted semigroup dynamical systems in which the multiplier ω is replaced by a product system X of Hilbert bimodules, and we construct a crossed product which is "twisted by X". For this, we first need to see how semigroups of endomorphism arise from Toeplitz representations of product systems.

Proposition 4.1. (1) *Let X be a Hilbert bimodule over A, and suppose (ψ, π) is a Toeplitz representation of X on a Hilbert space $\mathcal{H}$. Then there is a unique endomorphism $\alpha = \alpha^{\psi,\pi}$ of $\pi(A)'$ such that*

$$\alpha(S)\psi(x) = \psi(x)S \qquad \text{for all } S \in \pi(A)' \text{ and } x \in X, \tag{4.1}$$

and such that $\alpha(1)$ vanishes on $(\psi(X)\mathcal{H})^\perp$.

(2) *Let X be a product system over P of Hilbert A–A bimodules in which each fiber X_s is essential as a left A-module. Let ψ be a nondegenerate Toeplitz representation of X on a Hilbert space $\mathcal{H}$, and let α_s^ψ be the endomorphism α^{ψ_s,ψ_e} above. Then $\alpha^\psi : P \to \mathrm{End}(\psi_e(A)')$ is a semigroup homomorphism, and α_e^ψ is the identity endomorphism.*

Proof. (1) The uniqueness of α is obvious. By [**23**, Proposition 2.69], there is a unital homomorphism $S \in \pi(A)' \mapsto 1 \otimes_A S \in \mathrm{Ind}\,\pi(\mathcal{L}(X))' \subseteq B(X \otimes_A \mathcal{H})$ determined by

$$1 \otimes_A S(x \otimes_A h) = x \otimes_A Sh \qquad \text{for } x \in X \text{ and } h \in \mathcal{H}.$$

Let $U : X \otimes_A \mathcal{H} \to \mathcal{H}$ be the isometry which satisfies $U(x \otimes_A h) = \psi(x)h$ (see the proof of [**13**, Proposition 1.6(1)]), and define

$$\alpha(S) := U(1 \otimes_A S)U^* \qquad \text{for all } S \in \pi(A)'.$$

Then α is a homomorphism, and $\alpha(1) = UU^*$ vanishes on $(\psi(X)\mathcal{H})^\perp$. If $S \in \pi(A)'$ and $x \in X$, then for any $h \in \mathcal{H}$ we have

$$\alpha(S)\psi(x)h = U(1 \otimes_A S)(x \otimes_A h) = U(x \otimes_A Sh) = \psi(x)Sh,$$

giving (4.1).

Since $\pi(a)\psi(x)h = \psi(\phi(a)x)h$, the space $\overline{\mathrm{span}}\{\psi(x)h : x \in X, h \in \mathcal{H}\}$ reduces π; hence for any $S \in \pi(A)'$ and $a \in A$, both $\alpha(S)\pi(a)$ and $\pi(a)\alpha(S)$

vanish on $(\psi(X)\mathcal{H})^\perp$. This and

$$\begin{aligned}\alpha(S)\pi(a)\psi(x)h &= \alpha(S)\psi(\phi(a)x)h = \psi(\phi(a)x)Sh\\ &= \pi(a)\psi(x)Sh = \pi(a)\alpha(S)\psi(x)h\end{aligned}$$

show that $\alpha(\pi(A)') \subseteq \pi(A)'$.

(2) Let $s, t \in P$, and suppose $x \in X_s$ and $y \in X_t$. Vectors of the form xy have dense linear span in X_{st}; since X_t is essential, this holds even when $s = e$ (see Remark 2.2). Since

$$\begin{aligned}\alpha^\psi_s(\alpha^\psi_t(S))\psi_{st}(xy) &= \alpha^\psi_s(\alpha^\psi_t(S))\psi_s(x)\psi_t(y)\\ &= \psi_s(x)\alpha^\psi_t(S)\psi_t(y) = \psi_s(x)\psi_t(y)S = \psi_{st}(xy)S,\end{aligned}$$

we deduce that

$$\alpha^\psi_s \circ \alpha^\psi_t(S)\psi_{st}(z) = \psi_{st}(z)S \qquad \text{for all } S \in \psi_e(A)' \text{ and } z \in X_{st}. \tag{4.2}$$

Once we show that $\alpha^\psi_s \circ \alpha^\psi_t(1) = \alpha^\psi_{st}(1)$, it follows from the uniqueness of α^ψ_{st} that $\alpha^\psi_s \circ \alpha^\psi_t = \alpha^\psi_{st}$.

From (4.2) we see that $\alpha^\psi_s \circ \alpha^\psi_t(1) \geq \alpha^\psi_{st}(1)$. Suppose that $\alpha^\psi_s \circ \alpha^\psi_t(1)f = f$; we will show that $\alpha^\psi_{st}(1)f = f$, which will complete the proof. Since f is in the range of $\alpha^\psi_s(1)$, it can be approximated by a finite sum $\sum_i \psi_s(x_i)g_i$. Then

$$f \doteq \alpha^\psi_s \circ \alpha^\psi_t(1)\sum_i \psi_s(x_i)g_i = \sum_i \psi_s(x_i)\alpha^\psi_t(1)g_i.$$

Now each $\alpha^\psi_t(1)g_i$ can be approximated by a finite sum $\sum_j \psi_t(y_{ij})h_{ij}$, and then

$$f \doteq \sum_{i,j} \psi_s(x_i)\psi_t(y_{ij})h_{ij} = \sum_{i,j} \psi_{st}(x_i y_{ij})h_{ij}.$$

Thus f can be approximated arbitrarily closely by a vector in the range of $\alpha^\psi_{st}(1)$, and hence $\alpha^\psi_{st}(1)f = f$.

Since each X_t is essential, the assumption that ψ is nondegenerate implies that ψ_e is a nondegenerate representation of A. Since $\alpha^\psi_e(S)\psi_e(a)h = \psi_e(a)Sh = S\psi_e(a)h$ for all $a \in A$ and $h \in \mathcal{H}$, we have $\alpha^\psi_e(S) = S$ for all $S \in \psi_e(A)'$. □

Consider a twisted semigroup dynamical system (C, P, β, X) in which C is a separable C^*-algebra, $\beta : P \to \operatorname{End} C$ is an action of the semigroup P as extendible endomorphisms of C, and X is a product system over P of Hilbert A–A bimodules. We assume that β_e is the identity endomorphism, and that each fiber X_s is essential as a left A-module.

Definition 4.2. A *covariant representation* of (C, P, β, X) on a Hilbert space $\mathcal{H}$ is a pair (L, ψ) consisting of a nondegenerate representation L :

$C \to B(\mathcal{H})$ and a nondegenerate Toeplitz representation $\psi : X \to B(\mathcal{H})$ such that

(i) $L(C) \subseteq \psi_e(A)'$, and
(ii) $L \circ \beta_s = \alpha_s^{\psi} \circ L$ for $s \in P$.

Definition 4.3. A *crossed product* for (C, P, β, X) is a triple (B, i_C, i_X) consisting of a C^*-algebra B, a nondegenerate homomorphism $i_C : C \to M(B)$, and a nondegenerate Toeplitz representation $i_X : X \to M(B)$ such that

(a) there is a faithful nondegenerate representation σ of B such that $(\overline{\sigma} \circ i_C, \overline{\sigma} \circ i_X)$ is a covariant representation of (C, P, β, X);
(b) for every covariant representation (L, ψ) of (C, P, β, X), there is a representation $L \times \psi$ of B such that $\overline{(L \times \psi)} \circ i_C = L$ and $\overline{(L \times \psi)} \circ i_X = \psi$; and
(c) the C^*-algebra B is generated by $\{i_C(c) i_X(x) : c \in C,\ x \in X\}$.

Remark 4.4. If each fiber X_s has a finite basis $\{u_{s,1}, \dots, u_{s,n(s)}\}$ (in the sense that $x = \sum_k u_{s,k} \cdot \langle u_{s,k}, x \rangle_A$ for every $x \in X_s$), it is not hard to show that (a) is equivalent to asking that $i_C(c) i_X(a) = i_X(a) i_C(c)$ for all $c \in C$ and $a \in A = X_e$, and that

$$i_C(\beta_s(c)) = \sum_k i_X(u_{s,k}) i_C(c) i_X(u_{s,k})^* \qquad \text{for all } s \in P \text{ and } c \in C.$$

In this case, $(\overline{\sigma} \circ i_C, \overline{\sigma} \circ i_X)$ will be a covariant representation of (C, P, β, X) for *every* nondegenerate representation σ of B; however, as demonstrated in [**12**, Example 2.5] for product systems of Hilbert spaces, in general one cannot expect this to be the case.

Proposition 4.5. *If (C, P, β, X) has a covariant representation, then it has a crossed product $(C \rtimes_{\beta,X} P, i_C, i_X)$ which is unique in the following sense: If (B, i'_C, i'_X) is another crossed product for (C, P, β, X), then there is an isomorphism $\theta : C \rtimes_{\beta,X} P \to B$ such that $\overline{\theta} \circ i_C = i'_C$ and $\theta \circ i_X = i'_X$.*

Remark 4.6. When X is the product system $P \times \mathbb{C}$ with multiplication given by a multiplier ω, it is not hard to see that $C \rtimes_{\beta,X} P$ is precisely the crossed product $C \rtimes_{\beta,\omega} P$ defined in the previous section. If C is unital and $A = \mathbb{C}$, then $C \rtimes_{\beta,X} P$ is the crossed product defined in [**12**, Section 2].

Proof of Proposition 4.5. If S is a set of pairs (L, ψ) consisting of maps $L : C \to B(\mathcal{H}_{L,\psi})$ and $\psi : X \to B(\mathcal{H}_{L,\psi})$, then $(\oplus L, \oplus \psi)$ is a covariant representation of (C, P, β, X) if and only if each (L, ψ) is. The main point here is that the value of $\alpha_s^{\oplus \psi}$ on an element of $(\oplus \psi)_e(A)'$ of the form $\oplus L(c)$ is $\oplus \alpha_s^{\psi}(L(c))$.

Suppose (L, ψ) is a nondegenerate covariant representation on a separable Hilbert space $\mathcal{H}$; that is, the C^*-algebra

$$\mathcal{U} := C^*(\{L(c)\psi(x) : c \in C, x \in X\})$$

acts nondegenerately on $\mathcal{H}$. We will identify the multiplier algebra of $\mathcal{U}$ with the concrete C^*-algebra

$$M(\mathcal{U}) = \{S \in B(\mathcal{H}) : ST, TS \in \mathcal{U} \text{ for every } T \in \mathcal{U}\}.$$

We claim that $L(C) \cup \psi(X) \subseteq M(\mathcal{U})$. For this, it suffices to check that multiplying a generator $L(c)\psi(x)$ of $\mathcal{U}$ on either the left or the right by an operator of the form $L(d)$, $\psi(y)$, or $\psi(y)^*$ yields another element of $\mathcal{U}$. Certainly $L(d)L(c)\psi(x) = L(dc)\psi(x) \in \mathcal{U}$ and $L(c)\psi(x)\psi(y) = L(c)\psi(xy) \in \mathcal{U}$, and since

$$\psi(y)L(c) = \alpha^{\psi}_{p(y)}(L(c))\psi(y) = L(\beta_{p(y)}(c))\psi(y), \tag{4.3}$$

we also have $\psi(y)L(c)\psi(x) = L(\beta_{p(y)}(c))\psi(yx) \in \mathcal{U}$ and $L(c)\psi(x)L(d) = L(c\beta_{p(x)}(d))\psi(x) \in \mathcal{U}$. Writing $c = c_1^* c_2$ with $c_1, c_2 \in C$ gives

$$\psi(y)^* L(c)\psi(x) = (L(c_1)\psi(y))^* L(c_2)\psi(x) \in \mathcal{U}.$$

Finally, to see that $L(c)\psi(x)\psi(y)^* \in \mathcal{U}$, we use a trick from [**1**]. Let (c_i) be an approximate identity for C; we claim that

$$L(c)\psi(x)L(c_i) \xrightarrow{\|\ \|} L(c)\psi(x). \tag{4.4}$$

Since L is nondegenerate, $L(c)\psi(x)L(c_i)$ converges strongly to $L(c)\psi(x)$. On the other hand, using (4.3) we see that $L(c)\psi(x)L(c_i) = L(c\beta_{p(x)}(c_i))\psi(x)$ converges in *norm* (to $L(c\overline{\beta_{p(x)}}(1))\psi(x)$), and (4.4) follows. Hence

$$L(c)\psi(x)L(c_i)\psi(y)^* \xrightarrow{\|\ \|} L(c)\psi(x)\psi(y)^*,$$

and since

$$L(c)\psi(x)L(c_i)\psi(y)^* = L(c)\psi(x)\psi(y)^* L(\beta_{p(y)}(c_i)) \in \mathcal{U},$$

we deduce that $L(c)\psi(x)\psi(y)^* \in \mathcal{U}$.

Since $M(\mathcal{U}) \subseteq \mathcal{U}''$, we have shown in particular that the ranges of L and ψ are contained in $\mathcal{U}''$. Consequently, any decomposition $1 = \sum Q_\lambda$ of the identity as a sum of mutually orthogonal projections $Q_\lambda \in \mathcal{U}'$ gives corresponding decompositions $L = \oplus Q_\lambda L$ and $\psi = \oplus Q_\lambda \psi$, and by the first paragraph each pair $(Q_\lambda L, Q_\lambda \psi)$ is a covariant representation of (C, P, β, X). By the usual Zorn's Lemma argument we can choose these projections such that $\mathcal{U}$ acts cyclically on $Q_\lambda \mathcal{H}$; since $C^*(\{Q_\lambda L(c) Q_\lambda \psi(x) : c \in C,\ x \in X\}) = Q_\lambda \mathcal{U}$ acts cyclically on $Q_\lambda \mathcal{H}$, this shows that every covariant representation of (C, P, β, X) decomposes as a direct sum of cyclic representations.

Let S be a set of cyclic covariant representations with the property that every cyclic covariant representation of (C, P, β, X) is unitarily equivalent

to an element in S. It can be shown that such a set S exists by fixing a Hilbert space $\mathcal{H}$ of sufficiently large cardinality (depending on the cardinalities of C and X) and considering only representations on $\mathcal{H}$. Note that S is nonempty because the system has a covariant representation, which has a cyclic summand.

Define $i_C := \bigoplus_{(L,\psi)\in S} L$ and $i_X := \bigoplus_{(L,\psi)\in S} \psi$, and let $C\rtimes_{\beta,X}P$ be the C^*-algebra generated by $\{i_C(c)i_X(x) : c \in C,\ x \in X\}$. By the first paragraph, (i_C, i_X) is a covariant representation of (C, P, β, X), and it is nondegenerate since each (L, ψ) is. We deduce that both i_C and i_X map into $M(C\rtimes_{\beta,X}P)$, and that Condition (a) for a crossed product is satisfied by taking σ to be the identity representation. Condition (c) is trivial, and (b) holds because every covariant representation decomposes as a direct sum of cyclic representations. We need to show that $i_C : C \to M(C\rtimes_{\beta,X}P)$ and $i_X : X \to M(C\rtimes_{\beta,X}P)$ are nondegenerate. For this, let $c \in C$ and $x \in X$. If (a_i) is an approximate identity for $A = X_e$, then by (1.2) we have $i_C(c)i_X(x)i_X(a_i) = i_C(c)i_X(x \cdot a_i) \to i_C(c)i_X(x)$ and $i_X(a_i)i_C(c)i_X(x) = i_C(c)i_X(a_i)i_X(x) = i_C(c)i_X(\phi(a_i)x) \to i_C(c)i_X(x)$, so i_X is nondegenerate (Lemma 2.13). If (c_i) is an approximate identity for C, then $i_C(c_i)i_C(c)i_X(x) = i_C(c_ic)i_X(x) \to i_C(c)i_X(x)$, and since i_C is nondegenerate as a representation on Hilbert space, (4.4) gives $i_C(c)i_X(x)i_C(c_i) \to i_C(c)i_X(x)$. Thus i_C is nondegenerate.

For the uniqueness assertion, suppose (B, i'_C, i'_X) is another crossed product. Condition (a) allows us to assume that (i_C, i_X) and (i'_C, i'_X) are covariant representations of (C, P, β, X) on Hilbert spaces $\mathcal{H}$ and $\mathcal{H}'$. Condition (b) then gives a representation $i'_C \times i'_X : C\rtimes_{\beta,X}P \to B(\mathcal{H}')$ whose image is contained in B since $i'_C \times i'_X(i_C(c)i_X(x)) = i'_C(c)i'_X(x)$. Similarly one obtains a map $i_C \times i_X : B \to C\rtimes_{\beta,X}P$ which is obviously an inverse for $i'_C \times i'_X : C\rtimes_{\beta,X}P \to B$. □

If P is a subsemigroup of a group G, then there is a *dual coaction* of G on $C\rtimes_{\beta,X}P$:

Proposition 4.7. *Suppose (C, P, β, X) is a twisted system which has a covariant representation. If P is a subsemigroup of a group G, then there is an injective coaction*

$$\delta : C\rtimes_{\beta,X}P \to (C\rtimes_{\beta,X}P) \otimes_{\min} C^*(G)$$

such that

$$\delta(i_C(c)i_X(x)) = i_C(c)i_X(x) \otimes i_G(p(x)).$$

If G is abelian, there is a strongly continuous action $\widehat{\beta}$ of $\widehat{G}$ on $C\rtimes_{\beta,X}P$ such that

$$\widehat{\beta}_\gamma(i_C(c)i_X(x)) = \gamma(p(x))i_C(c)i_X(x).$$

Proof. We follow [**12**, Proposition 2.7]. Let σ be a faithful nondegenerate representation σ of $C \rtimes_{\beta,X} P$ such that $(L, \psi) := (\overline{\sigma} \circ i_C, \overline{\sigma} \circ i_X)$ is a covariant representation of (C, P, β, X), and let U be a unitary representation of G whose integrated form π_U is faithful on $C^*(G)$. We claim that $(L \otimes 1, \psi \otimes (U \circ p))$ is a covariant representation of (C, P, β, X). Most of the verifications are routine, so we check only that

$$L(\beta_s(c)) \otimes 1 = \alpha_s^{\psi \otimes (U \circ p)}(L(c) \otimes 1) \qquad \text{for all } s \in P \text{ and } c \in C. \tag{4.5}$$

For this, we show that $L(\beta_s(c)) \otimes 1$ satisfies the properties which characterize $\alpha_s^{\psi \otimes (U \circ p)}(L(c) \otimes 1)$ (Proposition 4.1). First, let $x \in X_s$; we show that (4.5) holds on any vector in the range of $(\psi \otimes (U \circ p))(x) = \psi_s(x) \otimes U_s$:

$$\begin{aligned}(L(\beta_s(c)) \otimes 1)(\psi_s(x) \otimes U_s) &= \alpha_s^{\psi}(L(c))\psi_s(x) \otimes U_s = \psi_s(x)L(c) \otimes U_s \\ &= (\psi_s(x) \otimes U_s)(L(c) \otimes 1) = \alpha_s^{\psi \otimes (U \circ p)}(L(c) \otimes 1)(\psi_s(x) \otimes U_s).\end{aligned}$$

Next, note that $\alpha_s^{\psi \otimes (U \circ p)}(1)$ is the projection onto

$$\begin{aligned}&\overline{\operatorname{span}}\{(\psi \otimes (U \circ p))(x)\xi : x \in X_s, \xi \in \mathcal{H}_\sigma \otimes \mathcal{H}_U\} \\ &\qquad = \overline{\operatorname{span}}\{\psi_s(x)h \otimes U_s k : x \in X_s, h \in \mathcal{H}_\sigma, k \in \mathcal{H}_U\} \\ &\qquad = \overline{\operatorname{span}}\{\psi_s(x)h : x \in X_s, h \in \mathcal{H}_\sigma\} \otimes \mathcal{H}_U,\end{aligned}$$

which is precisely the range of $\alpha_s^{\psi}(1) \otimes 1$. Since $L(\beta_s(c)) \otimes 1 = \alpha_s^{\psi}(L(c)) \otimes 1$ vanishes on the range of $1 - \alpha_s^{\psi}(1) \otimes 1$, (4.5) follows from the uniqueness assertion of Proposition 4.1.

Since $(L \otimes 1, \psi \otimes (U \circ p))$ is covariant, there is a representation ρ of $C \rtimes_{\beta,X} P$ such that

$$\begin{aligned}\rho(i_C(c)i_X(x)) &= (L(c) \otimes 1)(\psi(x) \otimes U_{p(x)}) \\ &= (\sigma \otimes \pi_U)(i_C(c)i_X(x) \otimes i_G(p(x))).\end{aligned}$$

Since σ and π_U are faithful, $\sigma \otimes \pi_U$ is faithful on $(C \rtimes_{\beta,X} P) \otimes_{\min} C^*(G)$, and we can define $\delta := (\sigma \otimes \pi_U)^{-1} \circ \rho$.

By checking on generators it is easy to see that δ satisfies the coaction identity $(\mathrm{id} \otimes \delta_G) \circ \delta = (\delta \otimes \mathrm{id}) \circ \delta$, and δ is injective since $\sigma = (\sigma \otimes \epsilon) \circ \delta$, with ϵ the augmentation representation of $C^*(G)$ (i.e., $\epsilon(i_G(s)) = 1$ for all $s \in G$). When G is abelian, $\widehat{\beta}$ is the action canonically associated with δ. □

5. Nica covariance.

Now suppose P is a subsemigroup of a group G such that $P \cap P^{-1} = \{e\}$. Then $s \leq t$ iff $s^{-1}t \in P$ defines a partial order on G which is left-invariant: For any $r, s, t \in P$ we have $s \leq t$ iff $rs \leq rt$. Following Nica [**20**], we say that (G, P) is a *quasi-lattice ordered group* if every finite subset of G which has an upper bound in P has a least upper bound in P. When $s, t \in P$ have a common upper bound, we denote their least upper bound by $s \vee t$; when s

and t have no common upper bound we write $s \vee t = \infty$. For a finite subset $C = \{t_1, \dots, t_n\}$ of P, we write σC for $t_1 \vee \cdots \vee t_n$.

Definition 5.1. Suppose (G, P) is a quasi-lattice ordered group and X is a product system over P of essential Hilbert A–A bimodules. We call a Toeplitz representation $\psi : X \to B(\mathcal{H})$ *Nica covariant* if

$$\alpha_s^\psi(1)\alpha_t^\psi(1) = \begin{cases} \alpha_{s\vee t}^\psi(1) & \text{if } s \vee t < \infty \\ 0 & \text{otherwise.} \end{cases}$$

Remark 5.2. If (G, P) is totally ordered, then every Toeplitz representation of X is Nica covariant.

Lemma 5.3. *Let $l : X \to \mathcal{L}(F(X))$ be the Fock representation, and suppose π is a representation of A on a Hilbert space $\mathcal{H}$. Then*

$$\Psi := F(X)\text{-Ind}_A^{\mathcal{L}(F(X))}\, \pi \circ l$$

is a Nica-covariant Toeplitz representation of X. If π is faithful, then Ψ is isometric.

Proof. Since l is a Toeplitz representation, so is Ψ. Let $s \in P$. The range of $\alpha_s^\Psi(1)$ is

$$\begin{aligned} &\overline{\operatorname{span}}\{\Psi(x)\xi : x \in X_s, \xi \in F(X) \otimes_A \mathcal{H}\} \\ &\qquad = \overline{\operatorname{span}}\{l(x)y \otimes_A h : x \in X_s, y \in F(X), h \in \mathcal{H}\} = \bigoplus_{s \le r} X_r \otimes_A \mathcal{H}. \end{aligned}$$

Hence for any $s, t \in P$, the range of $\alpha_s^\Psi(1)\alpha_t^\Psi(1)$ is

$$\left(\bigoplus_{s\le r} X_r \otimes_A \mathcal{H}\right) \cap \left(\bigoplus_{t \le r} X_r \otimes_A \mathcal{H}\right),$$

which is $\bigoplus_{s\vee t \le r} X_r \otimes_A \mathcal{H} = \operatorname{ran} \alpha_{s\vee t}^\Psi(1)$ if $s \vee t < \infty$, and is zero otherwise. If π is faithful then so is $F(X)\text{-Ind}_A^{\mathcal{L}(F(X))}\, \pi$; since l is isometric, this implies that Ψ is isometric. □

Proposition 5.4. *Let (G, P) be a quasi-lattice ordered group such that every $s, t \in P$ have a common upper bound. Let X be a product system over P of essential Hilbert A–A bimodules such that the left action of A on each fiber X_s is by compact operators. Then every Toeplitz representation $\psi : X \to B(\mathcal{H})$ which is Cuntz-Pimsner covariant is also Nica covariant.*

Proof. Fix $s \in P$. Since (ψ_s, ψ_e) is Cuntz-Pimsner covariant and $\phi_s(A) \subseteq \mathcal{K}(X_s)$, [**11**, Lemma 1.9] gives $\psi_e(A)\mathcal{H} \subseteq \overline{\operatorname{span}}\, \psi_s(X_s)\mathcal{H}$. But X_s is essential, so the reverse inclusion holds as well, and since $\overline{\operatorname{span}}\, \psi_s(X_s)\mathcal{H}$ is precisely the range of $\alpha_s^\psi(1)$, we deduce that $\alpha_s^\psi(1)$ is constant in s. Since $s \vee t < \infty$ for all $s, t \in P$, this implies that ψ is Nica covariant. □

There are product systems for which Nica covariance is not a C^*-algebraic condition; that is, if $\psi : X \to B(\mathcal{H})$ is Nica covariant and $\sigma : C^*(\psi(X)) \to B(\mathcal{K})$ is a homomorphism, the composition $\sigma \circ \psi$ need not be Nica covariant [**10**, Example 1.3]. We pause a moment to show how to adapt the methods of [**10**] to avoid this pathology. The following Lemma collects some results we shall need for both this and the sequel.

Lemma 5.5. *Suppose X is a product system over P of essential Hilbert A–A bimodules, $\psi : X \to B(\mathcal{H})$ is a Toeplitz representation, and $s \in P$.*

(1) *There is a strict–strong continuous representation $\rho^\psi_s : \mathcal{L}(X_s) \to B(\mathcal{H})$ such that*

$$\rho^\psi_s(S)\psi_s(x)h = \psi_s(Sx)h \qquad \text{for all } S \in \mathcal{L}(X_s),\ x \in X_s, \text{ and } h \in \mathcal{H},$$

and such that $\rho^\psi_s(S)$ vanishes on $(\psi_s(X_s)\mathcal{H})^\perp$. Moreover, $\rho^\psi_s(S) = \psi^{(s)}(S)$ for every $S \in \mathcal{K}(X_s)$.

(2) $\rho^\psi_s(1) = \alpha^\psi_s(1)$.

(3) *If $a \in A$ satisfies $\phi_s(a) \in \mathcal{K}(X_s)$, then*

$$\psi_e(a)\rho^\psi_s(1) = \psi^{(s)}(\phi_s(a)) = \rho^\psi_s(1)\psi_e(a). \tag{5.1}$$

(4) *If $Q \in \psi_e(A)'$, then $\alpha^\psi_s(Q) \in \rho^\psi_s(\mathcal{L}(X_s))'$. Further, if Q is a projection such that ψ_e acts faithfully on $Q\mathcal{H}$, then ρ^ψ_s acts faithfully on $\alpha^\psi_s(Q)\mathcal{H}$.*

(5) *For all $S \in \mathcal{L}(X_s)$ and $t \in P$ we have*

$$\rho^\psi_{st}(S \otimes_A 1) = \rho^\psi_s(S)\rho^\psi_{st}(1) = \rho^\psi_{st}(1)\rho^\psi_s(S),$$

where $S \otimes_A 1(xy) := (Sx)y$ for all $x \in X_s$ and $y \in X_t$.

(6) *If $t \in P$ and $z, w \in X_s$, then $\rho^\psi_{st}(\Theta_{z,w} \otimes_A 1) = \psi(z)\alpha^\psi_t(1)\psi(w)^*$.*

Proof. (1) See [**13**, Proposition 1.6(1)]. For the continuity assertion, suppose $S_\lambda \to S$ strictly in $\mathcal{L}(X_s) = M\mathcal{K}(X_s)$, $x \in X_s$, and $h \in \mathcal{H}$. There exists $K \in \mathcal{K}(X_s)$ and $y \in X_s$ such that $x = Ky$, and then

$$\rho^\psi_s(S_\lambda)\psi_s(x)h = \rho^\psi_s(S_\lambda K)\psi_s(y)h \to \rho^\psi_s(SK)\psi_s(y)h = \rho^\psi_s(S)\psi_s(x)h.$$

(2) Both $\rho^\psi_s(1)$ and $\alpha^\psi_s(1)$ are the projection onto $\overline{\operatorname{span}}\{\psi_s(X_s)\mathcal{H}\}$.

(3) If $x \in X_s$ and $h \in \mathcal{H}$, then

$$\psi_e(a)\rho^\psi_s(1)\psi_s(z)h = \psi_e(a)\psi_s(z)h = \psi_s(\phi_s(a)z)h = \psi^{(s)}(\phi_s(a))\psi_s(z)h;$$

since both sides of (5.1) are supported on $\overline{\operatorname{span}}\,\psi_s(X_s)\mathcal{H}$, this implies that $\psi_e(a)\rho^\psi_s(1) = \psi^{(s)}(\phi_s(a))$. By (2), $\rho^\psi_s(1)$ commutes with $\psi_e(a)$, giving the other half of (5.1).

(4) When Q is a projection, $\alpha^\psi_s(Q)$ is the projection onto $\overline{\operatorname{span}}\,\psi_s(X_s)Q\mathcal{H}$, and the result follows from [**13**, Proposition 1.6(2)].

(5) See [**13**, Proposition 1.8(2)].

(6) $\rho^\psi_{st}(\Theta_{z,w} \otimes_A 1) = \rho^\psi_{st}(1)\rho^\psi_s(\Theta_{z,w}) = \alpha^\psi_{st}(1)\psi(z)\psi(w)^* = \psi(z)\alpha^\psi_t(1)\psi(w)^*$. □

Proposition 5.6. *Suppose (G, P) is a quasi-lattice ordered group and X is a product system over P of essential Hilbert A–A bimodules. A Toeplitz representation $\psi : X \to B(\mathcal{H})$ is Nica covariant if and only if*

$$\rho^\psi_s(S)\rho^\psi_t(T) = \begin{cases} \rho^\psi_{s\vee t}((S \otimes_A 1)(T \otimes_A 1)) & \text{if } s \vee t < \infty \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

holds whenever $S \in \mathcal{K}(X_s)$ and $T \in \mathcal{K}(X_t)$.

Proof [**10**, Proposition 1.4]. If ψ is Nica covariant, then

$$\begin{aligned} \rho^\psi_s(S)\rho^\psi_t(T) &= \rho^\psi_s(S)\rho^\psi_s(1)\rho^\psi_t(1)\rho^\psi_t(T) \\ &= \rho^\psi_s(S)\rho^\psi_{s\vee t}(1)\rho^\psi_t(T) = \rho^\psi_{s\vee t}((S \otimes_A 1)(T \otimes_A 1)), \end{aligned}$$

where the last equality uses Lemma 5.5(5). Conversely, suppose (5.2) holds for all compact S and T. If $S \to 1$ strictly, then

$$\rho^\psi_{s\vee t}((S \otimes_A 1)) = \rho^\psi_s(S)\rho^\psi_{s\vee t}(1) \to \rho^\psi_s(1)\rho^\psi_{s\vee t}(1) = \rho^\psi_{s\vee t}(1),$$

where the convergence is in the strong operator topology. Hence

$$\rho^\psi_s(1)\rho^\psi_t(T) = \begin{cases} \rho^\psi_{s\vee t}(T \otimes_A 1) & \text{if } s \vee t < \infty \\ 0 & \text{otherwise} \end{cases}$$

for every $T \in \mathcal{K}(X_t)$. Letting $T \to 1$ strictly shows that ψ is Nica covariant. □

When each product $(S\otimes_A 1)(T\otimes_A 1)$ is compact, the previous Proposition allows us to give a C^*-algebraic characterization of Nica covariance:

Definition 5.7. Suppose (G, P) is a quasi-lattice ordered group and X is a product system over P of essential Hilbert A–A bimodules. We say that X is *compactly aligned* if whenever $s, t \in P$ have a common upper bound and S and T are compact operators on X_s and X_t, respectively, $(S\otimes_A 1)(T\otimes_A 1)$ is a compact operator on $X_{s\vee t}$. If X is compactly aligned and ψ is a Toeplitz representation of X in a C^*-algebra B, we say that ψ is *Nica covariant* if

$$\psi^{(s)}(S)\psi^{(t)}(T) = \begin{cases} \psi^{(s\vee t)}((S \otimes_A 1)(T \otimes_A 1)) & \text{if } s \vee t < \infty \\ 0 & \text{otherwise} \end{cases}$$

whenever $s, t \in P$, $S \in \mathcal{K}(X_s)$ and $T \in \mathcal{K}(X_t)$.

Proposition 5.8. *If (G, P) is a total order, or if the left action of A on each fiber X_s is by compact operators, then X is compactly aligned.*

Proof. Suppose $s, t \in P$, $s\vee t < \infty$, $S \in \mathcal{K}(X_s)$, and $T \in \mathcal{K}(X_t)$. If (G, P) is a total order then either $S\otimes_A 1 = S$ or $T\otimes_A 1 = T$; either way $(S\otimes_A 1)(T\otimes_A 1)$ is compact. If the left action of A on each fiber X_s is by compact operators, then by [**22**, Corollary 3.7], both $S \otimes_A 1$ and $T \otimes_A 1$ are compact. □

Proposition 5.9. *Suppose X is compactly aligned. Let B and C be C^*-algebras, let $\psi : X \to B$ be a Nica-covariant Toeplitz representation, and let $\sigma : B \to C$ be a homomorphism. Then $\sigma \circ \psi$ is Nica covariant.*

Proof. By checking on an operator $\Theta_{x,y} \in \mathcal{K}(X_s)$, one verifies that $(\sigma \circ \psi)^{(s)} = \sigma \circ \psi^{(s)}$, and the result follows easily from this. □

Proposition 5.10. *Suppose X is a compactly-aligned product system, ψ is a Nica-covariant Toeplitz representation of X, $s, t \in P$, $y \in X_s$, and $z \in X_t$. If $s \vee t = \infty$, then $\psi(y)^*\psi(z) = 0$; otherwise*

$$\psi(y)^*\psi(z) \in \overline{\operatorname{span}}\{\psi(f)\psi(g)^* : f \in X_{s^{-1}(s\vee t)}, g \in X_{t^{-1}(s\vee t)}\}.$$

Proof. Express $y = Sy'$ with $S \in \mathcal{K}(X_s)$ and $y' \in X_s$; similarly, express $z = Tz'$ with $T \in \mathcal{K}(X_t)$ and $z' \in X_t$. Since ψ is Nica covariant,

$$\psi(y)^*\psi(z) = \psi(y')^*\rho^\psi_s(S^*)\rho^\psi_t(T)\psi(z')$$

is zero if $s \vee t = \infty$, and otherwise

$$\psi(y)^*\psi(z) = \psi(y')^*\rho^\psi_{s\vee t}(K)\psi(z'),$$

where $K = (S^* \otimes_A 1)(T \otimes_A 1) \in \mathcal{K}(X_{s\vee t})$. Since K is compact it can be approximated in norm by a finite sum of operators $\Theta_{u,v}$ with $u, v \in X_{s\vee t}$, and hence $\rho^\psi_{s\vee t}(K)$ can be approximated by finite sums of the form $\psi(u)\psi(v)^*$. But any such u can be approximated by finite sums of products $u_1 f'$ with $u_1 \in X_s$ and $f' \in X_{s^{-1}(s\vee t)}$; similarly, any such v can be approximated by finite sums of products $v_1 g'$ with $v_1 \in X_t$ and $g' \in X_{t^{-1}(s\vee t)}$. Hence $\psi(y')^*\rho^\psi_{s\vee t}(K)\psi(z')$ can be approximated in norm by finite sums of operators of the form

$$\psi(y')^*\psi(u_1)\psi(f')\psi(g')^*\psi(v_1)^*\psi(z') = \psi(\langle y', u_1\rangle_A f')\psi(\langle z', v_1\rangle_A g')^*.$$

□

The following Lemma is useful when working with Nica-covariant Toeplitz representations.

Lemma 5.11. *Suppose (G, P) is a quasi-lattice ordered group, X is a product system over P of essential Hilbert A–A bimodules, ψ is a Toeplitz representation of X on $\mathcal{H}$, $x \in X$, and $s \in P$.*

(1) *If $p(x) \le s$, then $\alpha^\psi_s(S)\psi(x) = \psi(x)\alpha^\psi_{p(x)^{-1}s}(S)$ for all $S \in \psi_e(A)'$.*
(2) *If ψ is Nica covariant, then*

$$\alpha^\psi_s(1)\psi(x) = \begin{cases} \psi(x)\alpha^\psi_{p(x)^{-1}(p(x)\vee s)}(1) & \text{if } p(x) \vee s < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. The proof is formally identical to that of [**12**, Lemma 3.6]. □

6. The system (B_P, P, τ, X).

For each $t \in P$, let $1_t \in \ell^\infty(P)$ be the characteristic function of tP. Since the product $1_s 1_t$ is either $1_{s\vee t}$ or 0, $B_P := \overline{\text{span}}\{1_t : t \in P\}$ is a C^*-subalgebra of $\ell^\infty(P)$. Left translation on $\ell^\infty(P)$ restricts to an action τ of P on B_P, determined by $\tau_s(1_t) = 1_{st}$ for $s, t \in P$.

Proposition 6.1. *Suppose (G, P) is a quasi-lattice ordered group and X is a product system over P of essential Hilbert A–A bimodules.*

(1) *If (L, ψ) is a covariant representation of (B_P, P, τ, X), then ψ is a nondegenerate Nica-covariant Toeplitz representation of X and $L(1_s) = \alpha^\psi_s(1)$.*
(2) *If ψ is a nondegenerate Nica-covariant Toeplitz representation of X on a Hilbert space $\mathcal{H}$, then there is a representation $L^\psi : B_P \to B(\mathcal{H})$ such that $L^\psi(1_s) = \alpha^\psi_s(1)$; moreover, (L^ψ, ψ) is then a covariant representation of (B_P, P, τ, X).*

Proof. The proof is formally identical to that of [**12**, Proposition 4.1], except that in (2) one must also note that $L^\psi(B_P) \subseteq \psi_e(A)'$ since $L^\psi(1_s) = \alpha^\psi_s(1) \in \psi_e(A)'$ and $\{1_s : s \in P\}$ generates B_P. □

Corollary 6.2. *The system (B_P, P, τ, X) has a covariant representation.*

Proof. Let π be a nondegenerate representation of A on a Hilbert space $\mathcal{H}$, and let $l : X \to \mathcal{L}(F(X))$ be the Fock representation of X. By Lemma 5.3, $\Psi := F(X)\text{-Ind}_A^{\mathcal{L}(F(X))}\, \pi \circ l$ is a Nica-covariant Toeplitz representation of X. Since π is nondegenerate, so is $F(X)\text{-Ind}_A^{\mathcal{L}(F(X))}\, \pi$; since l is nondegenerate, Ψ is as well. The previous Proposition thus gives a covariant representation (L^Ψ, Ψ) of (B_P, P, τ, X). □

Let i_X and i_{B_P} be the canonical maps of X and B_P into $M(B_P \rtimes_{\tau,X} P)$. Since B_P is unital, $i_X(x) = i_{B_P}(1) i_X(x) \in B_P \rtimes_{\tau,X} P$ for each $x \in X$. We write $\mathcal{T}_{\text{cov}}(X)$ for the C^*-subalgebra of $B_P \rtimes_{\tau,X} P$ generated by $i_X(X)$; the following Theorem justifies this notation.

Theorem 6.3. *$(\mathcal{T}_{\text{cov}}(X), i_X)$ is universal for Nica-covariant Toeplitz representations of X, in the sense that:*

(a) *There is a faithful representation θ of $\mathcal{T}_{\text{cov}}(X)$ on Hilbert space such that $\theta \circ i_X$ is a Nica-covariant Toeplitz representation of X; and*
(b) *for every Nica-covariant Toeplitz representation ψ of X, there is a representation ψ_* of $\mathcal{T}_{\text{cov}}(X)$ such that $\psi = \psi_* \circ i_X$.*

Up to canonical isomorphism, $(\mathcal{T}_{\text{cov}}(X), i_X)$ is the unique pair with this property. If X is compactly aligned, then i_X is Nica covariant,

$$\mathcal{T}_{\text{cov}}(X) = \overline{\text{span}}\{i_X(x) i_X(y)^* : x, y \in X\}, \tag{6.1}$$

and

$$B_P \rtimes_{\tau,X} P = \overline{\operatorname{span}}\{i_X(x) i_{B_P}(1_s) i_X(y)^* : x, y \in X, s \in P\}. \tag{6.2}$$

If the left action of A on each fiber X_s is by compact operators, then $\mathcal{T}_{\mathrm{cov}}(X)$ is all of $B_P \rtimes_{\tau,X} P$; if in addition every $s, t \in P$ have a common upper bound, then the Cuntz-Pimsner algebra $\mathcal{O}_X$ is a quotient of $\mathcal{T}_{\mathrm{cov}}(X)$.

Proof of Theorem 6.3. Let σ be a faithful representation of $B_P \rtimes_{\tau,X} P$ on a Hilbert space $\mathcal{H}$ such that $(\overline{\sigma} \circ i_{B_P}, \sigma \circ i_X)$ is a covariant representation of (B_P, P, τ, X). By Proposition 6.1(1), $\sigma \circ i_X$ is a Nica-covariant Toeplitz representation of X, so we can take θ to be the restriction of σ to $\mathcal{T}_{\mathrm{cov}}(X)$. Suppose ψ is a (nondegenerate) Nica-covariant Toeplitz representation of X. Proposition 6.1(2) gives us a covariant representation (L^ψ, ψ) of (B_P, P, τ, X), and hence a representation $L^\psi \times \psi$ of $B_P \rtimes_{\tau,X} P$ such that $(L^\psi \times \psi) \circ i_X = \psi$. Restricting $L^\psi \times \psi$ to $\mathcal{T}_{\mathrm{cov}}(X)$ gives the required representation ψ_*. Uniqueness of $(\mathcal{T}_{\mathrm{cov}}(X), i_X)$ follows by the usual argument.

Suppose X is compactly aligned. Since i_X is the composition of the Nica-covariant Toeplitz representation $\sigma \circ i_X$ and the homomorphism σ^{-1} (restricted to $\sigma(\mathcal{T}_{\mathrm{cov}}(X))$), i_X is Nica covariant by Proposition 5.9. Let $w \in X$, and express $w = z \cdot a$ for some $z \in X$ and $a \in A$. Then $i_X(w) = i_X(z) i_X(a^*)^*$, so $\mathcal{A} := \overline{\operatorname{span}}\{i_X(x) i_X(y)^* : x, y \in X\}$ contains $i_X(X)$. Obviously $\mathcal{A}$ is a closed self-adjoint subspace of $\mathcal{T}_{\mathrm{cov}}(X)$, and since X is compactly aligned, Proposition 5.10 shows that $\mathcal{A}$ is closed under multiplication. This gives (6.1).

Now let $\mathcal{B} := \overline{\operatorname{span}}\{i_X(x) i_{B_P}(1_s) i_X(y)^* : x, y \in X, s \in P\}$. Using Lemma 5.11 with $\psi := \sigma \circ i_X$, and then applying σ^{-1}, gives

$$i_X(x) i_{B_P}(1_s) = i_{B_P}(1_{p(x)s}) i_X(x) \tag{6.3}$$

and

$$i_{B_P}(1_s) i_X(x) = \begin{cases} i_X(x) i_{B_P}(1_{p(x)^{-1}(p(x) \vee s)}) & \text{if } p(x) \vee s < \infty \\ 0 & \text{otherwise.} \end{cases} \tag{6.4}$$

Equation (6.3) shows that

$$i_X(x) i_{B_P}(1_s) i_X(y)^* = i_{B_P}(1_{p(x)s}) i_X(x) (i_{B_P}(1_{p(y)s}) i_X(y))^* \in B_P \rtimes_{\tau,X} P,$$

so $\mathcal{B} \subseteq B_P \rtimes_{\tau,X} P$. Since B_P is generated by $\{1_s : s \in P\}$, elements of the form $i_{B_P}(1_s) i_X(w)$ generate $B_P \rtimes_{\tau,X} P$ as a C^*-algebra; with $w = z \cdot a$ as

above, (6.4) shows that

$$\begin{aligned} i_{B_P}(1_s)i_X(w) &= i_{B_P}(1_s)i_X(z)i_X(a^*)^* \\ &= \begin{cases} i_X(z)i_{B_P}(1_{p(z)^{-1}(p(z)\vee s)})i_X(a^*)^* & \text{if } p(z)\vee s<\infty \\ 0 & \text{otherwise} \end{cases} \\ &\in \mathcal{B}. \end{aligned}$$

Hence to establish (6.2), it remains only to show that $\mathcal{B}$ is closed under multiplication. But Proposition 5.10 shows that the product

$$i_X(x)i_{B_P}(1_s)i_X(y)^*i_X(z)i_{B_P}(1_t)i_X(w)^*$$

of two typical generators of $\mathcal{B}$ is contained in the closed linear span of elements of the form

$$i_X(x)i_{B_P}(1_s)i_X(f)i_X(g)^*i_{B_P}(1_t)i_X(w)^*,$$

which by (6.4) simplifies to

$$i_X(xf)i_{B_P}(1_{p(f)^{-1}(p(f)\vee s)\vee p(g)^{-1}(p(g)\vee t)})i_X(wg)^*\in\mathcal{B}.$$

Suppose the left action of A on each X_s is by compact operators; that is, $\phi_s(A)\subseteq\mathcal{K}(X_s)$ for all $s\in P$. Let $x\in X$ and $s\in P$. Since $X_{p(x)}$ is essential, we can express $x=\phi_{p(x)}(a)z$ for some $a\in A$ and $z\in X_{p(x)}$. With $\psi:=\sigma\circ i_X$, we then have

$$\begin{aligned} \sigma(i_{B_P}(1_s)i_X(x)) &= L^\psi(1_s)\psi(x)=\rho_s^\psi(1)\psi_e(a)\psi(z) \\ &= \psi^{(s)}(\phi_s(a))\psi(z) && \text{(Lemma 5.5(3))} \\ &= \sigma(i_X^{(s)}(\phi_s(a))i_X(z)), \end{aligned}$$

so $i_{B_P}(1_s)i_X(x)=i_X^{(s)}(\phi_s(a))i_X(z)\in\mathcal{T}_{\rm cov}(X)$. Since elements of the form $i_{B_P}(1_s)i_X(x)$ generate $B_P\rtimes_{\tau,X}P$, this gives $B_P\rtimes_{\tau,X}P=\mathcal{T}_{\rm cov}(X)$.

If in addition every $s,t\in P$ have a common upper bound, then by Proposition 5.4 the universal map $j_X:X\to\mathcal{O}_X$ is Nica covariant; the integrated form $(j_X)_*:\mathcal{T}_{\rm cov}(X)\to\mathcal{O}_X$ is surjective since it maps generators to generators. □

7. Faithful representations.

Our strategy for characterizing faithful representations of $B_P\rtimes_{\tau,X}P$ follows [**12**, Section 5]. First we use the dual coaction δ of G on $B_P\rtimes_{\tau,X}P$ and the canonical trace ρ on $C^*(G)$ to define a positive linear map $E_\delta:=(\mathrm{id}\otimes\rho)\circ\delta$ of norm one of $B_P\rtimes_{\tau,X}P$ onto the fixed-point algebra $(B_P\rtimes_{\tau,X}P)^\delta$. When X is compactly aligned, (B_P,P,τ,X) satisfies the spanning condition (6.2),

and E_δ is determined by

$$E_\delta(i_X(x)i_{B_P}(1_s)i_X(y)^*) = \begin{cases} i_X(x)i_{B_P}(1_s)i_X(y)^* & \text{if } p(x) = p(y) \\ 0 & \text{otherwise.} \end{cases} \tag{7.1}$$

Definition 7.1. The system (B_P, P, τ, X) is *amenable* if E_δ is faithful on positive elements.

The argument of [**17**, Lemma 6.5] shows that if G is an amenable group, then the system (B_P, P, τ, X) is amenable. In Corollary 8.2 we will show that (B_P, P, τ, X) is also amenable when X is compactly aligned and G is a free product $*(G^\lambda, P^\lambda)$ with each G^λ an amenable group.

Theorem 7.2. *Suppose (G, P) is a quasi-lattice ordered group and X is a compactly-aligned product system over P of essential Hilbert A–A bimodules such that the system (B_P, P, τ, X) is amenable. Let ψ be a Nica-covariant Toeplitz representation of X on a Hilbert space $\mathcal{H}$. Then $L^\psi \times \psi$ is a faithful representation of $B_P \rtimes_{\tau,X} P$ if and only if*

(7.2) *for every $n \geq 1$ and $s_1, \ldots, s_n \in P \setminus \{e\}$, the subrepresentation*

$$a \in A \mapsto \psi_e(a) \prod_{k=1}^{n} \big(1 - L^\psi(1_{s_k})\big) \text{ of } \psi_e \text{ is faithful.}$$

Proof of necessity of (7.2). Let $\pi : A \to B(\mathcal{H})$ be a faithful nondegenerate representation of A on a Hilbert space $\mathcal{H}$, let $l : X \to \mathcal{L}(F(X))$ be the Fock representation of X, and let $\Psi := F(X)\text{-Ind}_A^{\mathcal{L}(F(X))}\, \pi \circ l$; by Lemma 5.3, Ψ is a Nica-covariant Toeplitz representation of X on $F(X) \otimes_A \mathcal{H}$. We claim that

$$a \in A \mapsto \Psi_e(a) \prod_{k=1}^{n} \big(1 - L^\Psi(1_{s_k})\big)$$

is faithful. Since $L^\Psi(1_{s_k}) = \alpha^\Psi_{s_k}(1)$ is the orthogonal projection of $F(X) \otimes_A \mathcal{H}$ onto $\bigoplus_{t \in s_k P} X_t \otimes_A \mathcal{H}$ (see the proof of Lemma 5.3), each projection $1 - L^\Psi(1_{s_k})$ dominates the projection Q_e onto the Ψ_e-invariant subspace $X_e \otimes_A \mathcal{H}$. To establish the claim it thus suffices to show that the subrepresentation $Q_e\Psi_e$ of Ψ_e is faithful. But $\Psi_e = F(X)\text{-Ind}_A^A \pi$ decomposes as $\bigoplus_{t \in P} X_t\text{-Ind}_A^A \pi$, so $Q_e\Psi_e = A\text{-Ind}_A^A \pi$ is unitarily equivalent to π, and hence faithful.

Now suppose that $L^\psi \times \psi$ is faithful and $a \in A$. Let

$$T := i_{B_P}\left(\prod_{k=1}^{n}(1 - 1_{s_k})\right) i_X(a) \in B_P \rtimes_{\tau,X} P.$$

Then

$$\|a\| = \left\|\Psi_e(a)\prod_{k=1}^{n}\bigl(1-L^{\Psi}(1_{s_k})\bigr)\right\| = \bigl\|L^{\Psi}\times\Psi(T)\bigr\| \le \|T\|$$
$$= \bigl\|L^{\psi}\times\psi(T)\bigr\| = \left\|\psi_e(a)\prod_{k=1}^{n}\bigl(1-L^{\psi}(1_{s_k})\bigr)\right\| \le \|a\|,$$

giving (7.2). □

Our proof that (7.2) implies faithfulness of $L^{\psi}\times\psi$ is based on the argument of [**12**, Section 6]: In Proposition 7.5(1) we prove that $L^{\psi}\times\psi$ is faithful on $(B_P\rtimes_{\tau,X}P)^{\delta}$, and in Proposition 7.5(2) we construct a spatial version E_ψ of E_δ such that $(L^{\psi}\times\psi)\circ E_\delta = E_\psi\circ(L^{\psi}\times\psi)$. Faithfulness of $L^{\psi}\times\psi$ then follows easily: If $L^{\psi}\times\psi(b)=0$, then

$$0 = E_\psi\circ(L^{\psi}\times\psi)(b^*b) = (L^{\psi}\times\psi)\circ E_\delta(b^*b),$$

so by Proposition 7.5(1), $E_\delta(b^*b)=0$. The amenability hypothesis then forces $b^*b=0$, and hence $b=0$.

We begin by reviewing some notation and results from [**17**, Remark 1.5] and [**12**, Remark 5.2]. Let F be a finite subset of P. A subset C of F is an *initial segment* of F if $c:=\sigma C$ is finite and $C=\{t\in F: t\le c\}$. (Recall that σC is the least upper bound of C; we use the convention that $\sigma\emptyset=e$.) For each such C there is a nonzero projection Q_C in B_P defined by

$$Q_C := 1_c\prod_{\{t\in F: c<t\vee c<\infty\}}(1-1_t),$$

and as C ranges over the initial segements of F, these projections form a decomposition of the identity in B_P.

Lemma 7.3. *Suppose (G,P) is a quasi-lattice ordered group, X is a product system over P of essential Hilbert A–A bimodules, ψ is a Nica-covariant Toeplitz representation of X on $\mathcal{H}$, F is a finite subset of P, C is an initial segment of F, $x,y\in X$ and $s\in P$. Let $c=\sigma C$, so that $C=\{t\in F: t\le c\}$.*

(1) *If $p(x)=p(y)$, then the operator $\psi(x)L^{\psi}(1_s)\psi(y)^*$ is in the commutant of $L^{\psi}(B_P)$. In particular, it commutes with $L^{\psi}(Q_C)$.*
(2) *If $p(x)s$, $p(y)s\in F$, then*

$$L^{\psi}(Q_C)\psi(x)L^{\psi}(1_s)\psi(y)^*L^{\psi}(Q_C)$$
$$= \begin{cases} L^{\psi}(Q_C)\psi(x)L^{\psi}(1_{p(x)^{-1}c})L^{\psi}(1_{p(y)^{-1}c})\psi(y)^*L^{\psi}(Q_C) \\ \qquad\qquad\qquad\qquad \text{if } p(x)s\le c \text{ and } p(y)s\le c \\ 0 \qquad \text{otherwise.}\end{cases}$$

Proof. The proof, based on Lemma 5.11, is identical in form to the proof of [**12**, Lemma 5.3]. □

Lemma 7.4. *Suppose (G, P) is a quasi-lattice ordered group, X is a product system over P of essential Hilbert A–A bimodules, and ψ is a Nica-covariant Toeplitz representation of X which satisfies* (7.2)*. Suppose further that F is a finite subset of P and Z is a finite sum $\sum \psi(x_k)L^\psi(1_{s_k})\psi(y_k)^*$ such that $p(x_k)s_k = p(y_k)s_k \in F$ for each k. Then*

$$\|Z\| = \max\{\|T_C\| : C \text{ is an initial segment of } F\}, \tag{7.3}$$

where T_C is the adjointable operator on $X_{\sigma C}$ defined by

$$T_C := \sum_{p(x_k)s_k \le \sigma C} \Theta_{x_k,y_k} \otimes_A 1^{p(x_k)^{-1}\sigma C}. \tag{7.4}$$

Proof. Since $\{Q_C : C \text{ is an initial segment of } F\}$ is a decomposition of the identity in B_P, and since L^ψ is a unital representation of B_P, the projections $L^\psi(Q_C)$ decompose the identity operator. By Lemma 7.3(1), Z commutes with each $L^\psi(Q_C)$, and thus

$$\|Z\| = \max\left\{\left\|L^\psi(Q_C)Z\right\| : C \text{ is an initial segment of } F\right\}.$$

Fix an initial segment C, and let $c := \sigma C$. By Lemma 7.3(2) and Lemma 5.5(6),

$$\begin{aligned} L^\psi(Q_C)Z &= L^\psi(Q_C)\sum \psi(x_k)L^\psi(1_{s_k})\psi(y_k)^* \\ &= L^\psi(Q_C)\sum_{p(x_k)s_k \le c} \psi(x_k)L^\psi(1_{p(x_k)^{-1}c})\psi(y_k)^* \\ &= L^\psi(Q_C)\sum_{p(x_k)s_k \le c} \rho^\psi_c(\Theta_{x_k,y_k} \otimes_A 1) \\ &= L^\psi(Q_C)\rho^\psi_c(T_C), \end{aligned}$$

so it suffices to show that

$$\left\|L^\psi(Q_C)\rho^\psi_c(T_C)\right\| = \|T_C\|. \tag{7.5}$$

Let

$$R_C := \prod_{\{t\in F : c<t\vee c<\infty\}} (1 - 1_{c^{-1}(t\vee c)}) \in B_P. \tag{7.6}$$

Since ψ satisfies (7.2),

$$a \mapsto \psi_e(a) \prod_{\{t\in F : c<t\vee c<\infty\}} (1 - L^\psi(1_{c^{-1}(t\vee c)})) = \psi_e(a)L^\psi(R_C)$$

is a faithful representation of A. By Lemma 5.5(4), the representation $T \in \mathcal{L}(X_c) \mapsto \alpha^\psi_c(L^\psi(R_C))\rho^\psi_c(T)$ is thus also faithful. But $\alpha^\psi_c(L^\psi(R_C)) = L^\psi(\tau_c(R_C)) = L^\psi(Q_C)$, and hence (7.5) is satisfied. □

Proposition 7.5. *Suppose (G, P) is a quasi-lattice ordered group, X is a compactly-aligned product system over P of essential Hilbert A–A bimodules, and ψ is a Nica-covariant Toeplitz representation of X which satisfies* (7.2)*.*

(1) *$L^\psi \times \psi$ is isometric on $(B_P \rtimes_{\tau,X} P)^\delta$.*
(2) *There is a linear map E_ψ of norm one of $L^\psi \times \psi(B_P \rtimes_{\tau,X} P)$ onto $L^\psi \times \psi\big((B_P \rtimes_{\tau,X} P)^\delta\big)$ such that $E_\psi \circ (L^\psi \times \psi) = (L^\psi \times \psi) \circ E_\delta$.*

Proof. (1) Since X is compactly aligned, the spanning condition (6.2) holds. Since E_δ is continuous and maps onto $(B_P \rtimes_{\tau,X} P)^\delta$, we deduce that finite sums

$$z := \sum i_X(x_k) i_{B_P}(1_{s_k}) i_X(y_k)^*$$

in which $p(x_k) = p(y_k)$ for all k are dense in $(B_P \rtimes_{\tau,X} P)^\delta$. It therefore suffices to fix such a z and show that $\big\|L^\psi \times \psi(z)\big\| = \|z\|$.

Let σ be a faithful nondegenerate representation of $B_P \rtimes_{\tau,X} P$ such that $(\overline{\sigma} \circ i_{B_P}, \sigma \circ i_X)$ is a covariant representation of (B_P, P, τ, X). By Proposition 6.1, $i := \sigma \circ i_X$ is a covariant representation of X and $\overline{\sigma} \circ i_{B_P} = L^i$. Since $L^i \times i = \sigma$ is faithful, i satisfies (7.2). Hence with $F := \{p(x_k)s_k\}$, Lemma 7.4 gives

$$\begin{aligned}\Big\|L^\psi \times \psi(z)\Big\| &= \Big\|\sum \psi(x_k) L^\psi(1_{s_k}) \psi(y_k)^*\Big\| \\ &= \max\{\|T_C\| : \ C \text{ is an initial segment of } F\} \\ &= \Big\|\sum i(x_k) L^i(1_{s_k}) i(y_k)^*\Big\| = \big\|L^i \times i(z)\big\| = \|z\|\,.\end{aligned}$$

(2) Since X is compactly aligned, finite sums of the form

$$w := \sum i_X(x_k) i_{B_P}(1_{s_k}) i_X(y_k)^*$$

are dense in $B_P \rtimes_{\tau,X} P$. We will show that $\big\|L^\psi \times \psi(E_\delta(w))\big\| \le \big\|L^\psi \times \psi(w)\big\|$; it follows that E_ψ is well-defined on operators of the form $L^\psi \times \psi(w)$ and extends to the desired linear contraction.

Let $F := \{p(x_k)s_k\} \cup \{p(y_k)s_k\}$, and let $Z := L^\psi \times \psi(E_\delta(w))$; by (7.1),

$$Z = \sum_{p(x_k)=p(y_k)} \psi(x_k) L^\psi(1_{s_k}) \psi(y_k)^*.$$

By Lemma 7.4, there is an initial segment C of F such that $\|Z\| = \|T_C\|$. Let $c := \sigma C$. We will construct a projection $R \in B_P$ such that $a \in A \mapsto \psi_e(a) L^\psi(R)$ is faithful, then define $Q := L^\psi(\tau_c(R)) = \alpha_c^\psi(L^\psi(R))$, and show that $Q(L^\psi \times \psi(w))Q = Q\rho_c^\psi(T_C)$. This will complete the proof, since by Lemma 5.5(4) we then have

$$\|Z\| = \|T_C\| = \Big\|Q\rho_c^\psi(T_C)\Big\| = \Big\|Q(L^\psi \times \psi(w))Q\Big\| \le \Big\|L^\psi \times \psi(w)\Big\|\,.$$

For each $s,t \in C$ such that $s \neq t$ and $s^{-1}c \vee t^{-1}c < \infty$, define $d_{s,t} \in P$ as in [**17**, Lemma 3.2]:

$$d_{s,t} = \begin{cases} (s^{-1}c)^{-1}(s^{-1}c \vee t^{-1}c) & \text{if } s^{-1}c < s^{-1}c \vee t^{-1}c \\ (t^{-1}c)^{-1}(s^{-1}c \vee t^{-1}c) & \text{otherwise,} \end{cases}$$

noting in particular that $d_{s,t}$ is never the identity in P. Let R_C be as in (7.6), and define

$$R := R_C \prod_{\substack{s \neq t \in C \\ s^{-1}c \vee t^{-1}c < \infty}} (1 - 1_{d_{s,t}}).$$

By condition (7.2), $a \in A \mapsto L^\psi(R)\psi_e(a)$ is faithful. The proof that $Q(L^\psi \times \psi(w))Q = Q\rho_c^\psi(T_C)$ is exactly as in [**12**, Proposition 5.5], so we omit it. □

Proposition 7.6. *Suppose (G,P) is a quasi-lattice ordered group and X is a compactly-aligned product system over P of essential Hilbert A–A bimodules. Let π be a nondegenerate representation of A on a Hilbert space $\mathcal{H}$, and let Ψ be the representation $F(X)\text{-Ind}_A^{\mathcal{L}(F(X))}\,\pi \circ l$, where $l : X \to \mathcal{L}(F(X))$ is the Fock representation of X. There is a projection E_Ψ of norm one of $L^\Psi \times \Psi(B_P \rtimes_{\tau,X} P)$ onto $L^\Psi \times \Psi((B_P \rtimes_{\tau,X} P)^\delta)$ such that*

$$E_\Psi \circ (L^\Psi \times \Psi) = (L^\Psi \times \Psi) \circ E_\delta; \tag{7.7}$$

moreover, E_Ψ is faithful on positive operators.

Proof. Denote by Q_t the orthogonal projection of $F(X) \otimes_A \mathcal{H}$ onto $X_t \otimes_A \mathcal{H}$. Since the Q_t's are mutually orthogonal, the formula

$$E_\Psi(T) := \sum_{t \in P} Q_t T Q_t \qquad \text{for } T \in L^\Psi \times \Psi(B_P \rtimes_{\tau,X} P)$$

defines a completely positive projection of norm one which is faithful on positive operators. We claim that

$$E_\Psi(\Psi(x)L^\Psi(1_s)\Psi(y)^*) = \begin{cases} \Psi(x)L^\Psi(1_s)\Psi(y)^* & \text{if } p(x) = p(y) \\ 0 & \text{otherwise.} \end{cases} \tag{7.8}$$

Since X is compactly aligned the spanning condition (6.2) holds, and hence (7.7) follows from (7.8) and (7.1).

Suppose $x, y \in X$ and $s \in P$. For each $t \in P$, $\Psi(x)L^\Psi(1_s)\Psi(y)^*$ is zero on $X_t \otimes_A \mathcal{H}$ unless $p(y)s \leq t$, in which case $\Psi(x)L^\Psi(1_s)\Psi(y)^*$ maps $X_t \otimes_A \mathcal{H}$ into $X_{p(x)p(y)^{-1}t} \otimes_A \mathcal{H}$. Thus if $p(x) \neq p(y)$, $Q_t\Psi(x)L^\Psi(1_s)\Psi(y)^*Q_t = 0$ for every $t \in P$, and $E_\Psi(\Psi(x)L^\Psi(1_s)\Psi(y)^*) = 0$. If on the other hand $p(x) = p(y)$,

then $Q_t\Psi(x)L^\Psi(1_s)\Psi(y)^*Q_t = \Psi(x)L^\Psi(1_s)\Psi(y)^*Q_t$ for each $t \in P$, and thus

$$\begin{aligned} E_\Psi(\Psi(x)L^\Psi(1_s)\Psi(y)^*) &= \sum_{t\in P} Q_t\Psi(x)L^\Psi(1_s)\Psi(y)^*Q_t \\ &= \Psi(x)L^\Psi(1_s)\Psi(y)^* \sum_{t\in P} Q_t = \Psi(x)L^\Psi(1_s)\Psi(y)^*. \end{aligned}$$

□

Corollary 7.7. *Suppose π is faithful. Then the system (B_P, P, τ, X) is amenable if and only if the representation $L^\Psi \times \Psi$ of $B_P \rtimes_{\tau,X} P$ is faithful.*

Proof. Suppose $L^\Psi \times \Psi$ is faithful. By Proposition 7.6, $(L^\Psi \times \Psi) \circ E_\delta = E_\Psi \circ (L^\Psi \times \Psi)$ is faithful on positive elements, hence so is E_δ; that is, (B_P, P, τ, X) is amenable. Since Ψ satisfies (7.2) (see the proof of necessity of (7.2)), the converse follows from Theorem 7.2. □

8. Amenability.

Theorem 8.1. *Suppose $\theta : (G, P) \to (\mathcal{G}, \mathcal{P})$ is a homomorphism of quasi-lattice ordered groups such that, whenever $s \vee t < \infty$,*

$$\theta(s \vee t) = \theta(s) \vee \theta(t) \quad \text{and} \quad \theta(s) = \theta(t) \Longrightarrow s = t, \tag{8.1}$$

and suppose that $\mathcal{G}$ is amenable. If X is a compactly-aligned product system over P of essential Hilbert A–A bimodules, then the system (B_P, P, τ, X) is amenable.

Proof. Our proof is essentially that of [**12**, Theorem 6.1], suitably modified to handle Hilbert bimodules. The homomorphism $\theta : G \to \mathcal{G}$ induces a coaction $\delta_\theta = (\mathrm{id} \otimes \theta) \circ \delta$ of $\mathcal{G}$ on $B_P \rtimes_{\tau,X} P$, and hence a conditional expectation E_{δ_θ} of $B_P \rtimes_{\tau,X} P$ onto the fixed-point algebra $(B_P \rtimes_{\tau,X} P)^{\delta_\theta}$, such that

$$E_{\delta_\theta}(i_X(x)i_{B_P}(1_s)i_X(y)^*) = \begin{cases} i_X(x)i_{B_P}(1_s)i_X(y)^* & \text{if } \theta(p(x)) = \theta(p(y)) \\ 0 & \text{otherwise.} \end{cases}$$

Since $\mathcal{G}$ is amenable, E_{δ_θ} is faithful on positive elements.

Let $l : X \to \mathcal{L}(F(X))$ be the Fock representation of X, let π be a faithful nondegenerate representation of A on a Hilbert space $\mathcal{H}$, and let $\Psi := F(X)\text{-Ind}_A^{\mathcal{L}(F(X))}\, \pi \circ l$. By Proposition 7.6, for every $b \in B_P \rtimes_{\tau,X} P$ we have

$$(L^\Psi \times \Psi) \circ E_\delta(b) = E_\Psi(L^\Psi \times \Psi(E_{\delta_\theta}(b))).$$

Since E_{δ_θ} and E_Ψ are faithful on positive elements, to show that (B_P, P, τ, X) is amenable it suffices to show that $L^\Psi \times \Psi$ is faithful on $(B_P \rtimes_{\tau,X} P)^{\delta_\theta}$.

Let σ be a faithful representation of $B_P \rtimes_{\tau,X} P$ such that $(\overline{\sigma} \circ i_{B_P}, \sigma \circ i_X)$ is a covariant representation of (B_P, P, τ, X). By Proposition 6.1, $i = \sigma \circ i_X$ is

a covariant representation of X and $\overline{\sigma}\circ i_{B_P} = L^i$. Observe that i is isometric since, by Lemma 5.3,

$$\begin{aligned}\|x\| = \|\Psi(x)\| &= \left\|(L^\Psi \times \Psi)\circ i_X(x)\right\| \\ &\leq \|i_X(x)\| = \|\sigma\circ i_X(x)\| = \|i(x)\| \leq \|x\|.\end{aligned}$$

Let $\mathcal{F}$ be the set of all finite subsets F of $\mathcal{P}$ which are closed under $\vee$ in the sense that $s\vee t\in F$ whenever $s,t\in F$ and $s\vee t<\infty$. Exactly as in the proof of [**12**, Theorem 6.1], one can use Proposition 5.10 to show that, for each $F\in\mathcal{F}$,

$$\mathcal{U}_F := \overline{\operatorname{span}}\{i_X(x)i_{B_P}(1_s)i_X(y)^* : \theta(p(x)s) = \theta(p(y)s)\in F\}$$

is a C^*-subalgebra of $B_P\rtimes_{\tau,X}P$. Applying Φ_{δ_θ} to both sides of (6.2) gives

$$(B_P\rtimes_{\tau,X}P)^{\delta_\theta} = \overline{\operatorname{span}}\{i_X(x)i_{B_P}(1_s)i_X(y)^* : \theta(p(x)) = \theta(p(y))\};$$

since $\mathcal{F}$ is directed under set inclusion (see the proof of [**17**, Lemma 4.1]), we deduce that

$$(B_P\rtimes_{\tau,X}P)^{\delta_\theta} = \overline{\bigcup_{F\in\mathcal{F}}\mathcal{U}_F}.$$

By [**2**, Lemma 1.3], to prove that $L^\Psi\times\Psi$ is faithful on $(B_P\rtimes_{\tau,X}P)^{\delta_\theta}$ it is enough to prove it is faithful on each of the subalgebras $\mathcal{U}_F$. We shall accomplish this by inducting on $|F|$.

First suppose $F=\{r\}$ for some $r\in\mathcal{P}$. Let W_r be the Hilbert A–A bimodule $\bigoplus_{t\in\theta^{-1}(r)}X_t$. We claim that, for each Nica-covariant Toeplitz representation ψ of X on a Hilbert space $\mathcal{K}$, there is a linear map $\psi_r : W_r\to B(\mathcal{K})$ which satisfies $\psi_r(\oplus x_t) = \sum\psi_t(x_t)$, and that (ψ_r,ψ_e) is then a Toeplitz representation of W_r. First observe that if $x,y\in X$ satisfy $p(x)\neq p(y)$ and $\theta(p(x)) = \theta(p(y)) = r$, then by (8.1) we have $p(x)\vee p(y)=\infty$, and hence $\psi(x)^*\psi(y)=0$. Now suppose $\oplus x_t$ belongs to the algebraic direct sum $\bigodot_{t\in\theta^{-1}(r)}X_t$; such vectors are dense in W_r. Then

$$\begin{aligned}\left\|\sum_t\psi_t(x_t)\right\|^2 &= \left\|\sum_{t,t'}\psi_t(x_t)^*\psi_{t'}(x_{t'})\right\| = \left\|\sum_t\psi_t(x_t)^*\psi_t(x_t)\right\| \\ &= \left\|\sum_t\psi_e(\langle x_t,x_t\rangle_A)\right\| \leq \left\|\sum_t\langle x_t,x_t\rangle_A\right\| \\ &= \|\langle\oplus x_t,\oplus x_t\rangle_A\| = \|\oplus x_t\|^2,\end{aligned}$$

ensuring the existence of ψ_r. It is routine to check that (ψ_r,ψ_e) is a Toeplitz representation of W_r. Write α_r^ψ for the endomorphism of $\psi_e(A)'$ which corresponds to (ψ_r,ψ_e) (Proposition 4.1), and write ρ_r^ψ for the associated representation of $\mathcal{L}(W_r)$ (Lemma 5.5).

Suppose Z is a finite sum $\sum i_X(x_k)i_{B_P}(1_{s_k})i_X(y_k)^*$ such that $\theta(p(x_k)s_k) = \theta(p(y_k)s_k) = r$ for every k; to prove $L^\Psi\times\Psi$ faithful on $\mathcal{U}_{\{r\}}$ we will show

that $\|L^\Psi \times \Psi(Z)\| = \|Z\|$. For each k, let $\Theta_{x_k,y_k} \otimes_A 1^{s_k}$ denote the operator in $\mathcal{L}(W_r)$ which is the image of

$$\Theta_{x_k,y_k} \in \mathcal{K}\left(X_{p(y_k)}, X_{p(x_k)}\right) \mapsto \Theta_{x_k,y_k} \otimes_A 1^{s_k} \in \mathcal{L}\left(X_{p(y_k)s_k}, X_{p(x_k)s_k}\right) \subset \mathcal{L}(W_r).$$

Define $T := \sum \Theta_{x_k,y_k} \otimes_A 1^{s_k} \in \mathcal{L}(W_r)$. It is routine to check that

$$\rho_r^\Psi(T) = \sum \Psi(x_k) L^\Psi(1_{s_k}) \Psi(y_k)^* = L^\Psi \times \Psi(Z),$$

and similarly $\rho_r^i(T) = L^i \times i(Z) = \sigma(Z)$. Since Ψ_e and i_e are faithful representations of A, the representations ρ_r^Ψ and ρ_r^i are isometric, and thus

$$\|L^\Psi \times \Psi(Z)\| = \|\rho_r^\Psi(T)\| = \|T\| = \|\rho_r^i(T)\| = \|\sigma(Z)\| = \|Z\|.$$

For the inductive step, suppose $F \in \mathcal{F}$ and $L^\Psi \times \Psi$ is faithful on $\mathcal{U}_{F'}$ whenever $F' \in \mathcal{F}$ and $|F'| < |F|$; we aim to prove that $L^\Psi \times \Psi$ is faithful on $\mathcal{U}_F$. Since F is finite it has a minimal element; that is, there exists $r_0 \in F$ such that $r_0 < r_0 \vee r$ for each $r \in F \setminus \{r_0\}$. As in the proof of [**12**, Theorem 6.1] we have $L^\Psi \times \Psi(\mathcal{U}_{\{r\}}) P_{r_0} = \{0\}$ for each $r \in F \setminus \{r_0\}$, where P_{r_0} denotes the orthogonal projection of $F(X) \otimes_A \mathcal{H}$ onto $\bigoplus_{t \in \theta^{-1}(r_0)} X_t \otimes_A \mathcal{H}$.

On the other hand, we have already demonstrated that $L^\Psi \times \Psi$ maps $\mathcal{U}_{r_0}$ isometrically into the range of $\rho_{r_0}^\Psi$, and an easy calculation shows that $P_{r_0} = \alpha_{r_0}^\Psi(Q_e)$, where Q_e is the orthogonal projection onto $X_e \otimes_A \mathcal{H}$. Since $a \mapsto \Psi_e(a)Q_e$ is faithful, by Lemma 5.5(4) the representation $S \in \mathcal{L}(W_{r_0}) \mapsto P_{r_0}\rho_{r_0}^\Psi(S)$ is also faithful. Hence the map $Y \in \mathcal{U}_{r_0} \mapsto L^\Psi \times \Psi(Y)P_{r_0}$ is faithful.

Now suppose $Y \in \mathcal{U}_F$ and $L^\Psi \times \Psi(Y) = 0$. We will show that $Y \in \mathcal{U}_{F\setminus\{r_0\}}$, from which the inductive hypothesis implies that $Y = 0$. Let (Y_n) be a sequence in

$$\operatorname{span}\{i_X(x) i_{B_P}(1_s) i_X(y)^* : \theta(p(x)s) = \theta(p(y)s) \in F\}$$

which converges in norm to Y, and express each Y_n as a sum $\sum_{r\in F} Y_{n,r}$, where $Y_{n,r} \in \mathcal{U}_{\{r\}}$. For each n,

$$\|L^\Psi \times \Psi(Y_n)P_{r_0}\| = \|L^\Psi \times \Psi(Y_{n,r_0})P_{r_0}\| = \|Y_{n,r_0}\|,$$

and consequently $Y_{n,r_0} \to 0$. Thus $Y_n - Y_{n,r_0} \to Y$, which shows that $Y \in \mathcal{U}_{F\setminus\{r_0\}}$, as claimed. □

Corollary 8.2. *Suppose (G^λ, P^λ) is a quasi-lattice ordered group with G^λ amenable for each λ belonging to some index set Λ. If X is a compactly-aligned product system over $P := *P^\lambda$, then the system (B_P, P, τ, X) is amenable.*

Proof. The group $\bigoplus G^\lambda$ is amenable, and by [**17**, Proposition 4.3] the canonical map $\theta : *G^\lambda \to \bigoplus G^\lambda$ satisfies (8.1). □

9. Applications.

In Section 3, we associated with each twisted semigroup dynamical system (A, P, β, ω) a product system $X = X(A, P, \beta, \omega)$ of essential Hilbert A–A bimodules over the opposite semigroup P^o (Lemma 3.2), and we showed that the Cuntz-Pimsner algebra $\mathcal{O}_X$ is canonically isomorphic to the crossed product $A \rtimes_{\beta,\omega} P$; we also showed that $\mathcal{T}_X$ has the structure of a certain "Toeplitz" crossed product $\mathcal{T}(A \rtimes_{\beta,\omega} P)$ (Proposition 3.4). Suppose now that (G^o, P^o) is quasi-lattice ordered; this is equivalent to (G, P) being quasi-latticed ordered in its *right*-invariant partial order ($s \leq t \Leftrightarrow ts^{-1} \in P$). Since the left action of A on each fiber X_s is by compact operators, X is compactly aligned (Lemma 5.8) and $\mathcal{T}_{\rm cov}(X) = B_P \rtimes_{\tau,X} P$ (Theorem 6.3). Hence we can apply Theorem 7.2 to characterize the faithful representations of $\mathcal{T}_{\rm cov}(X)$. This is particularly helpful when (G^o, P^o) is a total order since $\mathcal{T}_{\rm cov}(X) = \mathcal{T}_X$; more generally, when every $s, t \in P^o$ have a common upper bound in P^o (i.e., $Ps \cap Pt \neq \emptyset$), the crossed product $A \rtimes_{\beta,\omega} P = \mathcal{O}_X$ is a quotient of $\mathcal{T}_{\rm cov}(X)$ (Theorem 6.3).

We begin by showing that $\mathcal{T}_{\rm cov}(X)$, too, has a crossed product structure:

Definition 9.1. Suppose P is a subsemigroup of a group G and (G^o, P^o) is quasi-lattice ordered. A *Nica-Toeplitz covariant representation* of (A, P, β, ω) is a Toeplitz covariant representation (π, V) such that

$$V_s^* V_s V_t^* V_t = \begin{cases} V_{s\vee t}^* V_{s\vee t} & \text{if } s \vee t < \infty \\ 0 & \text{otherwise,} \end{cases} \tag{9.1}$$

where $s \vee t$ denotes the least upper bound of s and t in the right-invariant partial order on (G, P).

The following Proposition establishes the existence of a C^*-algebra which is universal for such pairs (π, V), as in Definition 3.1. We call this algebra the *Nica-Toeplitz crossed product* of (A, P, β, ω), and denote it $\mathcal{T}_{\rm cov}(A \rtimes_{\beta,\omega} P)$. Let $i_X : X \to \mathcal{T}_{\rm cov}(X)$ be universal for Nica-covariant Toeplitz representations of X. Lemma 3.3 is easily adapted to this setting, and allows us to define $i_P : P \to M\mathcal{T}_{\rm cov}(X)$ by $i_P(s) = \lim i_X(s, \beta_s(a_i))^*$; here (a_i) is an approximate identity for A, and the convergence is strict. We also define $i_A : A \to \mathcal{T}_{\rm cov}(X)$ by $i_A(a) := i_X(e, a)$.

Proposition 9.2. *$(\mathcal{T}_{\rm cov}(X), i_A, i_P)$ is a Nica-Toeplitz crossed product for (A, P, β, ω).*

Proof. As in the proof of Proposition 3.4, i_A is nondegenerate. We verify the obvious analogues of Conditions (a), (b), and (c) in Definition 3.1. For (a), let σ be a nondegenerate representation of $\mathcal{T}_{\rm cov}(X)$ on a Hilbert space $\mathcal{H}$, let $\pi := \sigma \circ i_A$, and let $V := \overline{\sigma} \circ i_P$; we must show that (π, V) is a Nica-Toeplitz covariant representation of (A, P, β, ω). Exactly as in the proof of

Proposition 3.4, (π, V) is a Toeplitz covariant representation of (A, P, β, ω), so we need to establish (9.1). Fix $s \in P$. For any $a \in A$ and $h \in \mathcal{H}$ we have

$$\begin{aligned} V_s^*\pi(a)h &= \sigma(i_P(s)^* i_A(a))h = \sigma(\lim i_X(s, \beta_s(a_i))i_X(e, a))h \\ &= \sigma(\lim i_X(s, \beta_s(a_i)a))h = \sigma \circ i_X(s, \overline{\beta_s}(1)a)h, \end{aligned}$$

and since π is nondegenerate this shows that

$$V_s^* V_s \mathcal{H} = \overline{\operatorname{span}}\{\sigma \circ i_X(\xi)h : \xi \in X_s, h \in \mathcal{H}\} = \alpha_s^{\sigma \circ i_X}(1).$$

Since X is compactly aligned, $\sigma \circ i_X$ is Nica covariant (Theorem 6.3 and Proposition 5.9), and (9.1) follows.

For Condition (b), let (π, V) be any Nica-Toeplitz covariant representation on $\mathcal{H}$. As in the proof of Proposition 3.4, $\psi(s, x) := V_s^*\pi(x)$ defines a nondegenerate Toeplitz covariant representation $\psi : X \to B(\mathcal{H})$. To see that it is Nica-covariant, let $s \in P$, and note that for any $a \in A$ we have

$$\begin{aligned} \psi(s, \overline{\beta_s}(1)a) &= \lim \psi(s, \beta_s(a_i)a) = \lim V_s^*\pi(\beta_s(a_i)a) \\ &= \lim V_s^* V_s \pi(a_i) V_s^* \pi(a) = V_s^*\pi(a). \end{aligned}$$

Since π is nondegenerate, this implies that $\alpha_s^{\psi}(1) = V_s^* V_s$, and hence ψ is Nica covariant by (9.1). Defining $\pi \times V := \psi_* : \mathcal{T}_{\mathrm{cov}}(X) \to B(\mathcal{H})$ gives the desired representation satisfying $(\pi \times V) \circ i_A = \pi$ and $\overline{\pi \times V} \circ i_P = V$. Condition (c) is satisfied since $i_A(a) i_P(s) = i_X(s, \overline{\beta_s}(1)a^*)^*$, and elements of this form generate $\mathcal{T}_{\mathrm{cov}}(X)$. □

Let (G_i, P_i) be a collection of abelian lattice-ordered groups. Since (G_i, P_i) is quasi-lattice ordered in both its left and its right-invariant partial order, so is the free product $*(G_i, P_i)$.

Theorem 9.3. *Suppose $(G, P) = *(G_i, P_i)$ is a free product of abelian lattice-ordered groups and (π, V) is a Nica-Toeplitz covariant representation of the twisted semigroup dynamical system (A, P, β, ω) on a Hilbert space $\mathcal{H}$. Then the integrated form $\pi \times V$ is a faithful representation of $\mathcal{T}_{\mathrm{cov}}(A \rtimes_{\beta,\omega} P)$ if and only if*

for every $n \geq 1$ and $s_1, \ldots, s_n \in P \setminus \{e\}$,

$$\pi \text{ acts faithfully on the range of } \prod_{k=1}^{n} (1 - V_{s_k}^* V_{s_k}).$$

Proof. Let θ be the canonical homomorphism of $*(G_i, P_i)$ onto $\bigoplus(G_i, P_i)$. By [**17**, Proposition 4.3], θ satisfies the hypotheses of Theorem 8.1; since $X = X(A, P, \beta, \omega)$ is compactly aligned, the system (B_P, P, τ, X) is therefore amenable. Identifying $\mathcal{T}_{\mathrm{cov}}(A \rtimes_{\beta,\omega} P)$ with $\mathcal{T}_{\mathrm{cov}}(X)$ as in the previous Proposition and defining $\psi(s, x) := V_s^*\pi(x)$, the initial projection $V_s^* V_s$ is precisely $\alpha_s^{\psi}(1)$, and the result follows from Theorem 7.2. □

Nica covariance is automatic when (G, P) is totally ordered:

Corollary 9.4. *Suppose (G, P) is a totally ordered abelian group and (π, V) is a Toeplitz covariant representation of (A, P, β, ω) on a Hilbert space $\mathcal{H}$. Then the integrated form $\pi \times V$ is a faithful representation of $\mathcal{T}(A \rtimes_{\beta,\omega} P)$ if and only if π acts faithfully on $(V_s^*\mathcal{H})^\perp$ for every $s \in P \setminus \{e\}$.*

Corollary 9.5. *Suppose β is an extendible endomorphism of A. If (π, V) is a Toeplitz representation of $(A, \mathbb{N}, \beta)$, then $\pi \times V$ is a faithful representation of $\mathcal{T}(A \rtimes_\beta \mathbb{N})$ if and only if π acts faithfully on $(V^*\mathcal{H})^\perp$.*

Bicovariance. Suppose (G, P) is a quasi-lattice ordered group. Following [**17**], in [**12**] it was shown that $B_P \rtimes_{\tau,\omega} P$ is universal for isometric ω-representations of P which are Nica covariant; that is, which satisfy

$$V_s V_s^* V_t V_t^* = \begin{cases} V_{s\vee t} V_{s\vee t}^* & \text{if } s \vee t < \infty \\ 0 & \text{otherwise.} \end{cases} \tag{9.2}$$

Assuming that (G^o, P^o) is also quasi-lattice ordered, we now show that the Nica-Toeplitz crossed product $\mathcal{T}_{\rm cov}(B_P \rtimes_{\tau,\omega} P)$ is universal for partial isometric ω-representations of P which are *bicovariant* in that they satisfy both (9.2) and (9.1). Note that bicovariance is automatic when (G, P) is a totally ordered abelian group.

Proposition 9.6. *$i_P : P \to \mathcal{T}_{\rm cov}(B_P \rtimes_{\tau,\omega} P)$ is a bicovariant partial isometric ω-representation of P whose range generates $\mathcal{T}_{\rm cov}(B_P \rtimes_{\tau,\omega} P)$ as a C^*-algebra. Moreover, for every bicovariant partial isometric ω-representation V, there is a representation V_* of $\mathcal{T}_{\rm cov}(B_P \rtimes_{\tau,\omega} P)$ such that $V_* \circ i_P = V$.*

Proof. Let σ be a faithful nondegenerate representation of $\mathcal{T}_{\rm cov}(B_P \rtimes_{\tau,\omega} P)$. Then $V := \overline{\sigma} \circ i_P$ is a partial isometric ω-representation of P which satisfies (9.1), and applying $\overline{\sigma}^{-1}$ we see that i_P is as well. Since $i_P(s)i_P(s)^* = i_{B_P}(1_s)$ for every $s \in P$, i_P also satisfies (9.2), and is hence bicovariant. Since $\{1_s : s \in P\}$ generates B_P linearly and $\{i_{B_P}(a)i_P(t) : a \in B_P, t \in P\}$ generates $\mathcal{T}_{\rm cov}(B_P \rtimes_{\tau,\omega} P)$ as a C^*-algebra, elements of the form $i_{B_P}(1_s)i_P(t) = i_P(s)i_P(s)^*i_P(t)$ are also generating. If V is any bicovariant partial isometric ω-representation of P, then by [**17**, Proposition 1.3] there is a representation π_V of B_P such that $\pi_V(1_s) = V_sV_s^*$ for every $s \in P$. For any $s, t \in P$ the product $V_tV_s = \omega(t, s)V_{ts}$ is a partial isometry; hence by [**14**, Lemma 2] the projections $V_sV_s^*$ and $V_t^*V_t$ commute, and we deduce that $\pi_V(a)V_t^*V_t = V_t^*V_t\pi_V(a)$ for every $a \in B_P$ and $t \in P$. Further,

$$\begin{aligned} \pi_V(\tau_s(1_t)) &= \pi_V(1_{st}) = V_{st}V_{st}^* \\ &= (\overline{\omega(s,t)}V_sV_t)(\overline{\omega(s,t)}V_sV_t)^* = V_sV_tV_t^*V_s^* = V_s\pi_V(1_t)V_s^*, \end{aligned}$$

so $\pi_V(\tau_s(a)) = V_s\pi_V(a)V_s^*$ for every $s \in P$ and $a \in B_P$. Thus (π_V, V) is a Nica-Toeplitz covariant representation of (B_P, P, τ, ω). The representation $V_* := \pi_V \times V$ satisfies $V_* \circ i_P = V$. □

We say that a bicovariant partial isometric ω-representation V is *universal* if, for every bicovariant partial isometric ω-representation W, there is a homomorphism of $C^*\{V_s : s \in P\}$ which maps V_s to W_s for each $s \in P$.

Theorem 9.7. *Suppose $(G, P) = *(G_i, P_i)$ is a free product of abelian lattice-ordered groups and V is a bicovariant partial isometric ω-representation of P. Then V is universal if and only if*

$$\prod_{l=1}^{m}(V_r V_r^* - V_{rt_l} V_{rt_l}^*) \prod_{k=1}^{n}(1 - V_{s_k}^* V_{s_k}) \neq 0$$

whenever $r \in P$, $m, n \geq 1$, and $s_1, \ldots, s_n, t_1, \ldots, t_m \in P \setminus \{e\}$.

Proof. V is universal if and only if the representation $V_* = \pi_V \times V$ of $\mathcal{T}_{\mathrm{cov}}(B_P \rtimes_{\tau,\omega} P)$ is faithful. By Theorem 9.3, this occurs if and only if π_V acts faithfully on the range of $\prod_{k=1}^{n}(1 - V_{s_k}^* V_{s_k})$ whenever $s_1, \ldots, s_n \in P \setminus \{e\}$, and the result follows from [**17**, Proposition 1.3]. □

Let $\mathbb{F}_\infty$ be the free group on infinitely many generators $z_1, z_2, \ldots$, and let $\mathbb{F}_\infty^+$ be the subsemigroup (with identity) generated by the z_i; the pair $(\mathbb{F}_\infty, \mathbb{F}_\infty^+)$ is quasi-lattice ordered. In [**17**], Laca and Raeburn realized the Cuntz algebra $\mathcal{O}_\infty$ as the universal C^*-algebra for covariant isometric representations of $\mathbb{F}_\infty^+$, and used their characterization of the faithful representations of $B_P \rtimes_\tau P$ to derive Cuntz's simplicity result. We finish by showing that the universal C^*-algebra for bicovariant partial isometric representations of $\mathbb{F}_\infty^+$ is reminiscent of $\mathcal{O}_\infty$, and we derive a Cuntz-Krieger-type uniqueness theorem.

First some notation. For a multi-index $\mu = (\mu_1, \ldots, \mu_n)$ we write $z_\mu := z_{\mu_1} \cdots z_{\mu_n}$, and we identify $\mathbb{F}_\infty^+$ with the set of multi-indices under concatenation via $z_\mu \leftrightarrow \mu$.

Proposition 9.8. *Suppose S is a partial isometric representation of $\mathbb{F}_\infty^+$ in a C^*-algebra B; that is, S is a semigroup homomorphism and each S_μ is a partial isometry. Then $C^*\{S_\mu : \mu \in \mathbb{F}_\infty^+\}$ is generated by $\{S_n : n \in \mathbb{N}\}$, and S is bicovariant if and only if*

(a) *the range projections $s_k s_k^*$ for $k \in \mathbb{N}$ are pairwise orthogonal, and*
(b) *the initial projections $s_k^* s_k$ for $k \in \mathbb{N}$ are pairwise orthogonal.*

Proof. The first statement is obvious. In the left-invariant partial order on $\mathbb{F}_\infty$, two elements $\mu, \nu \in \mathbb{F}_\infty^+$ have a common upper bound if and only if one is an initial word of the other, and then the least upper bound is the longer of the two words. We will show that (a) holds if and only if

$$S_\mu S_\mu^* S_\nu S_\nu^* = \begin{cases} S_\mu S_\mu^* & \text{if } \nu^{-1}\mu \in \mathbb{F}_\infty^+, \\ S_\nu S_\nu^* & \text{if } \mu^{-1}\nu \in \mathbb{F}_\infty^+, \\ 0 & \text{otherwise;} \end{cases} \tag{9.3}$$

of course a similar statement holds for (b) using the right-invariant partial order, and together these prove the Proposition.

To begin with, (9.3) implies (a) since distict generators of $\mathbb{F}_\infty^+$ are not comparable. For the converse, first suppose $\nu^{-1}\mu \in \mathbb{F}_\infty^+$; since S_ν is a partial isometry, we then have

$$S_\mu S_\mu^* S_\nu S_\nu^* = S_\mu S_{\nu^{-1}\mu}^* S_\nu^* S_\nu S_\nu^* = S_\mu S_{\nu^{-1}\mu}^* S_\nu^* = S_\mu S_\mu^*.$$

The case $\mu^{-1}\nu \in \mathbb{F}_\infty^+$ is similar. Finally, suppose μ and ν are not comparable. Then there exists $\sigma, \mu', \nu' \in \mathbb{F}_\infty^+$ such that $\mu = \sigma\mu'$, $\nu = \sigma\nu'$, and $\mu_1' \neq \nu_1'$. Condition (a) implies that $S_{\mu'}^* S_{\nu'} = 0$, and by [**14**, Lemma 2] the range projection of $S_{\nu'}$ commutes with the initial projection of S_σ, so

$$S_\mu^* S_\nu = S_{\mu'}^* S_\sigma^* S_\sigma S_{\nu'} = S_{\mu'}^* S_\sigma^* S_\sigma S_{\nu'} S_{\nu'}^* S_{\nu'} = S_{\mu'}^* S_{\nu'} S_{\nu'}^* S_\sigma^* S_\sigma S_{\nu'} = 0.$$

□

Theorem 9.9. *A bicovariant partial isometric representation S of $\mathbb{F}_\infty^+$ is universal if and only if each S_μ is nonzero.*

Proof. Suppose each S_μ is nonzero. To see that S is universal, we apply Theorem 9.7. If $\nu \in \mathbb{F}_\infty^+$, $m, n \geq 1$, and $\sigma_1, \ldots, \sigma_m, \tau_1, \ldots, \tau_n \in \mathbb{F}_\infty^+ \setminus \{e\}$, then we can choose $i, j \in \mathbb{N}$ such that none of the multi-indices σ_l begins with i, and none of the multi-indices τ_k ends with j. Then

$$\prod_{l=1}^m (S_\nu S_\nu^* - S_{\nu\sigma_l} S_{\nu\sigma_l}^*) \prod_{k=1}^n (1 - S_{\tau_k}^* S_{\tau_k}) \geq S_\nu S_i S_i^* S_\nu^* S_j^* S_j = S_j^* S_{j\nu i} S_{j\nu i}^* S_j$$

is nonzero since $S_j(S_j^* S_{j\nu i} S_{j\nu i}^* S_j) S_j^* = S_{j\nu i} S_{j\nu i}^* \neq 0$. Hence S is universal.

Now define $T : \mathbb{F}_\infty^+ \to B(\ell^2(\mathbb{F}_\infty^+) \otimes \ell^2(\mathbb{F}_\infty^+))$ by

$$T_\mu(\delta_\sigma \otimes \delta_\nu) = \begin{cases} \delta_\sigma \otimes \delta_{\mu\nu} & \text{if } \sigma \text{ ends in } \mu\nu \\ 0 & \text{otherwise.} \end{cases}$$

Then T is a bicovariant partial isometric representation of $\mathbb{F}_\infty^+$ in which each T_μ is nonzero. If S is universal, then $S_\mu \mapsto T_\mu$ extends to a homomorphism of $C^*\{S_\mu\}$, and hence each S_μ must be nonzero. □

References

[1] S. Adji, *Ph.D. Thesis*, Univ. of Newcastle, 1995.

[2] S. Adji, M. Laca, M. Nilsen and I. Raeburn, *Crossed products by semigroups of endomorphisms and the Toeplitz algebras of ordered groups*, Proc. Amer. Math. Soc., **122** (1994), 1133-1141, MR 95b:46094, Zbl 0818.46071.

[3] W. Arveson, *Continuous analogues of Fock space*, Memoirs Amer. Math. Soc., **80**(**409**) (1989), MR 90f:47061, Zbl 0697.46035.

[4] S. Boyd, N. Keswani and I. Raeburn, *Faithful representations of crossed products by endomorphisms*, Proc. Amer. Math. Soc., **118** (1993), 427-436, MR 93g:46066, Zbl 0785.46051.

[5] J. Cuntz, *Simple C^*-algebras generated by isometries*, Comm. Math. Phys., **57** (1977), 173-185, MR 57 #7189, Zbl 0399.46045.

[6] ———, *The internal structure of simple C^*-algebras*, Proceedings of Symposia in Pure Mathematics, **38**(**1**) (1982), 85-115, MR 84h:46072, Zbl 0502.46039.

[7] H.T. Dinh, *Discrete product systems and their C^*-algebras*, J. Funct. Anal., **102** (1991), 1-34, MR 93d:46097, Zbl 0745.46057.

[8] ———, *On generalized Cuntz C^*-algebras*, J. Operator Theory, **30** (1993), 123-135, MR 95m:46112, Zbl 0837.46048.

[9] S. Doplicher, C. Pinzari and R. Zuccante, *The C^*-algebra of a Hilbert bimodule*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat., **8**(**1**) (1998), 263-281, MR 99i:46040, Zbl 0916.46053.

[10] N.J. Fowler, *Compactly-aligned discrete product systems, and generalizations of $\mathcal{O}_\infty$*, International J. Math., **10**(**6**) (1999), 721-738, CMP 1 715 181.

[11] N.J. Fowler, P.S. Muhly and I. Raeburn, *Representations of Cuntz-Pimsner algebras*, to appear in Indiana Univ. Math. J.

[12] N.J. Fowler and I. Raeburn, *Discrete product systems and twisted crossed products by semigroups*, J. Funct. Anal., **155** (1998), 171-204, MR 99k:46118, Zbl 0914.46054.

[13] ———, *The Toeplitz algebra of a Hilbert bimodule*, Indiana Univ. Math. J., **48**(**1**) (1999), 155-181, MR 2001b:46093, Zbl 0938.47052.

[14] P.R. Halmos and L.J. Wallen, *Powers of partial isometries*, Indiana Univ. Math. J., **19** (1970), 657-663, MR 40 #4801, Zbl 0202.42502.

[15] R. Hancock and I. Raeburn, *The C^*-algebras of some inverse semigroups*, Bull. Austral. Math. Soc., **42** (1990), 335-348, MR 91i:46061, Zbl 0716.46052.

[16] T. Kajiwara, C. Pinzari and Y. Watatani, *Ideal structure and simplicity of the C^*-algebras generated by Hilbert bimodules*, J. Funct. Anal., **159** (1998), 295-322, MR 2000a:46094, Zbl 0942.46035.

[17] M. Laca and I. Raeburn, *Semigroup crossed products and the Toeplitz algebras of nonabelian groups*, J. Funct. Anal., **139** (1996), 415-440, MR 97h:46109, Zbl 0887.46040.

[18] P.S. Muhly and B. Solel, *On the simplicity of some Cuntz-Pimsner algebras*, Math. Scand., **83** (1998), 53-73, MR 99m:46140, Zbl 0940.46034.

[19] ———, *Tensor algebras over C^*-correspondences* (*representations, dilations, and C^*-envelopes*), J. Funct. Anal., **158** (1998), 389-457, MR 99j:46066, Zbl 0912.46070.

[20] A. Nica, *C^*-algebras generated by isometries and Wiener-Hopf operators*, J. Operator Theory, **27** (1992), 17-52, MR 94m:46094, Zbl 0809.46058.

[21] W.L. Paschke, *The crossed product by an endomorphism*, Proc. Amer. Math. Soc., **80** (1980), 113-118, MR 81m:46081, Zbl 0435.46045.

[22] M.V. Pimsner, *A class of C^*-algebras generalizing both Cuntz-Krieger algebras and crossed products by $\mathbb{Z}$*, Fields Institute Communications, **12** (1997), 189-212, MR 97k:46069, Zbl 0871.46028.

[23] I. Raeburn and D.P. Williams, *Morita equivalence and continuous-trace C^*-algebras*, Math. Surveys and Monographs, **60** (1998), Amer. Math. Soc., Providence, MR 2000c:46108, Zbl 0922.46050.

[24] P.J. Stacey, *Crossed products of C^*-algebras by endomorphisms*, J. Austral. Math. Soc. (Series A), **54** (1993), 204-212, MR 94a:46077, Zbl 0785.46053.

Received April 25, 1999. The author thanks the University of Victoria, Canada, for its hospitality while this research was being completed.

3316 179th Avenue NE
Redmond WA 98052
E-mail address: duvalfowler@msn.com

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

THE PRODUCT FORMULA FOR THE SPHERICAL FUNCTIONS ON SYMMETRIC SPACES IN THE COMPLEX CASE

P. Graczyk and P. Sawyer

In this paper, we prove the existence of the product formula for the spherical functions in the complex case and we study properties of the integral kernel of this formula.

1. Introduction.

Let G be a semisimple noncompact Lie group with finite center and K a maximal compact subgroup of G and $X = G/K$ the corresponding Riemannian symmetric space of noncompact type. We have a Cartan decomposition $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ and we choose a maximal abelian subalgebra $\mathfrak{a}$ of $\mathfrak{p}$. In what follows, Σ corresponds to the root system of $\mathfrak{g}$ and Σ^+ to the positive roots. We have the root space decomposition $\mathfrak{g} = \mathfrak{g}_0 + \sum_{\alpha\in\Sigma} \mathfrak{g}_\alpha$. Let $\mathfrak{n} = \sum_{\alpha\in\Sigma^+} \mathfrak{g}_\alpha$. Denote the groups corresponding to the Lie algebras $\mathfrak{a}$ and $\mathfrak{n}$ by A and N respectively. We have the Cartan decomposition $G = K\,A\,K$ and the Iwasawa decomposition $G = K\,A\,N$. Let $\mathfrak{a}^+ = \{H \in A\colon \alpha(H) > 0 \ \forall\ \alpha \in \Sigma^+\}$ and $A^+ = \exp(\mathfrak{a}^+)$.

If λ is a complex-valued functional on $\mathfrak{a}$, the corresponding spherical function is

$$\phi_\lambda(e^H) = \int_K e^{(i\,\lambda-\rho)(\mathcal{H}(e^H\,k))}\,dk$$

where $g = k\,e^{\mathcal{H}(g)}\,n \in K\,A\,N$. A spherical function, like any K-biinvariant function, can also be considered as a K-invariant function on the Riemannian symmetric space of noncompact type $X = G/K$. Naturally, such a function is completely determined by its values on A (or on A^+). The books [**6**, **7**] constitute a standard reference on these topics.

Let us assume throughout the paper that $X,\ Y \in \mathfrak{a}^+$ and that the symmetric space G/K is irreducible.

In [**7**, (32), page 480], Helgason shows that if $X \neq 0$, $Y \neq 0$ and $Y \notin W \cdot \{-X\}$ (or equivalently that $X \notin W \cdot \{-Y\}$) then there exists a Weyl-invariant measure $\mu_{X,Y}$ on the Lie algebra $\mathfrak{a}$ such that

$$\phi_\lambda(e^X)\,\phi_\lambda(e^Y) = \int_{\mathfrak{a}} \phi_\lambda(e^H)\,d\mu_{X,Y}(H)$$

(unlike us, Helgason states his results at the group level). In fact, this is true for all X and Y.

The support of the measure $\mu_{X,Y}$ is shown to be included in $C(X)+C(Y)$ where $C(H)$ is the convex hull of the orbit of H under the action of the Weyl group W.

The measures δ_{e^X} and δ_{e^Y} are not K-invariant on G, except in the excluded cases X, $Y = 0$. If δ_K denotes the Haar measure on K, then define the K-biinvariant probability measures $\delta^\sharp_{e^X}$ and $\delta^\sharp_{e^Y}$ by convolving the Dirac masses with δ_K on both sides. Comparing the spherical Fourier transforms we see that

$$\mu_{X,Y} = \delta^\sharp_{e^X} * \delta^\sharp_{e^Y}.$$

It is known [**7**] that

$$\phi_\lambda(e^X)\,\phi_\lambda(e^Y) = \int_K \phi_\lambda(e^X\,k\,e^Y)\,dk.$$

The measure $\mu_{X,Y}$ is then to satisfy

$$\int_K f(e^X\,k\,e^Y)\,dk = \int_{\mathfrak{a}} f(e^H)\,d\mu_{X,Y}(H)$$

for all functions f which are biinvariant under the action of K.

The natural question is whether the measure $\mu_{X,Y}$ is absolutely continuous with respect to the Lebesgue measure on $\mathfrak{a}$, *i.e.,* whether we have a "product formula"

$$\phi_\lambda(e^X)\,\phi_\lambda(e^Y) = \int_{\mathfrak{a}} \phi_\lambda(e^H)\,k(H,X,Y)\,dH \tag{1}$$

where $k(H,X,Y)$ is Weyl invariant in each of the variables. Helgason also discusses this measure and some partial results in [**8**].

The question of existence of the density of the measure $\mu_{X,Y}$ is related to the question of absolute continuity of the measure ν_X on $\mathfrak{a}$ defined by

$$\int_K f(\mathcal{H}(e^X k))dk = \int_{\mathfrak{a}} f(H)d\nu_X(H), \quad f \in \mathcal{C}_c(\mathfrak{a}),$$

answered positively by Flensted-Jensen and Ragozin ([**3**]) when G/K is irreducible and $X \neq 0$.

Following the general idea of their proof one can prove the absolute continuity of $\mu_{X,Y}$ when $X, Y \in \mathfrak{a}^+$ and in some boundary cases X, $Y \in \partial\mathfrak{a}^+$ ([**5**]). This requires however considerable care due to the non-analyticity of the Cartan decomposition. Moreover, this general approach does not allow us to obtain the density explicitly or even to study its basic properties.

Koornwinder gave explicit formulae for the function $k(H,X,Y)$ for the rank one case in [**11**]. In fact, he gives a product formula for a larger class of special functions, namely the Jacobi functions. The formulae given can be

derived using an addition formula which is not currently available in higher rank situations. The reader may also wish to consult [**1**, **2**, **9**, **10**, **11**, **12**, **13**].

In this paper, we show directly the product formula (1) for symmetric spaces in the complex case, which is easy, as opposed to the general case. We also give a lot of information on the kernel k and its support.

Our formula has applications in special functions theory and multivariate statistics because it may be equivalently expressed in terms of the Schur or zonal polynomials on Hermitian positive definite matrices.

There are also important relations between product formulae for spherical functions and arithmetic of probability measures. Ostrovskii ([**14**]) and Trukhina ([**15**]) showed that the only measures without indecomposable factors (in the sense of convolution product), respectively in the set of radial measures on R^n and in the set of K-invariant measures on real hyperbolic spaces, are the Gaussian measures. Also Voit ([**16**]) studied this question on some hypergroups. The main tool of all this research is a product formula (1) with some information on its kernel. We think that our formula will give similar characterization of Gaussian measures on symmetric spaces with G complex.

Two more intrinsic applications of (1) are given in the end of Section 2.

We thank Tom Koornwinder for helpful remarks and Amos Nevo for pointing out to us the application of the product formula given in the Corollary 2.6. We thank the referee for helpful comments.

2. The product formula on complex Lie groups.

We consider the spherical functions on complex groups.

We require some preliminaries.

We first note that there exists a function $K(X, H)$ which is Weyl-invariant in both of its arguments such that

$$\phi_\lambda(e^X) = \int_{C(X)} e^{\langle i\,\lambda, H\rangle}\, K(X, H)\, dH \tag{2}$$

(K is defined for $X \neq 0$).

The existence of the kernel $K(X, H)$ in (2) is shown in [**7**, p. 479]. It is simply the kernel of the Abel transform. This is valid for every symmetric space of noncompact type.

If we use the Cartan decomposition, the integration on G can be written in polar coordinates. With suitable normalization, we have

$$\int_G f(g)\, dg = \int_K \int_K \int_{\mathfrak{a}^+} f(k_1\, e^H\, k_2)\, \delta(H)\, dH\, dk_1\, dk_2$$

where $\delta(H) = \prod_{\alpha\in\Sigma^+} \sinh^{m_\alpha} \alpha(H)$ and m_α denotes the multiplicity of the root α.

In the complex case $m_\alpha = 2$ for each α and we have

$$\delta^{1/2}(X) = \sum_{w\in W} \epsilon(w)\, e^{\langle w\rho, X\rangle}. \tag{3}$$

It is worthwhile to mention that as it is written in (3), the function $\delta^{1/2}$ is skew Weyl-invariant i.e., $\delta^{1/2}(w\cdot H) = \epsilon(w)\,\delta^{1/2}(H)$.

Still in the complex case, we have

$$\phi_\lambda(e^X) = \frac{\pi(\rho)}{\pi(i\lambda)} \frac{\sum_{w\in W} \epsilon(w)\, e^{\langle iw\cdot\lambda, X\rangle}}{\delta^{1/2}(X)}.$$

Theorem 2.1. *Suppose G is a complex Lie group. Then we have the following product formula*

$$\phi_\lambda(e^X)\,\phi_\lambda(e^Y) = \int_{\mathfrak{a}} \phi_\lambda(e^H)\, k(H,X,Y)\,\delta(H)\,dH$$

where

$$k(H,X,Y) = \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)} \frac{1}{|W|} \sum_{w\in W} \epsilon(w)\, K(X, w\cdot H - Y). \tag{4}$$

Proof. We observe first that

$$\begin{aligned}
&\int_{C(X)} \phi_\lambda(e^{H+Y}) \frac{K(X,H)\,\delta^{1/2}(H+Y)}{\delta^{1/2}(Y)}\,dH \\
&\quad= \frac{\pi(\rho)}{\pi(i\lambda)} \sum_{w\in W} \epsilon(w) \int_{C(X)} e^{\langle iw\cdot\lambda, H+Y\rangle} \frac{K(X,H)\,\delta^{1/2}(H+Y)}{\delta^{1/2}(H+Y)\,\delta^{1/2}(Y)}\,dH \\
&\quad= \frac{\pi(\rho)}{\pi(i\lambda)} \sum_{w\in W} \epsilon(w) \frac{e^{\langle iw\cdot\lambda, Y\rangle}}{\delta^{1/2}(Y)} \int_{C(X)} e^{\langle iw\cdot\lambda, H\rangle}\, K(X,H)\,dH \\
&\quad= \frac{\pi(\rho)}{\pi(i\lambda)} \sum_{w\in W} \epsilon(w) \frac{e^{\langle iw\cdot\lambda, Y\rangle}}{\delta^{1/2}(Y)}\, \phi_{w\cdot\lambda}(e^X) \\
&\quad= \phi_\lambda(e^Y)\,\phi_\lambda(e^X)
\end{aligned}$$

(we note first that $\phi_{w\cdot\lambda} = \phi_\lambda$ and then we add over w).

Hence,

$$\begin{aligned}
&\int_{C(X)+Y} \phi_\lambda(e^H) \frac{K(X,H-Y)}{\delta^{1/2}(H)\,\delta^{1/2}(Y)}\,\delta(H)\,dH \\
&\qquad= \int_{C(X)} \phi_\lambda(e^{H+Y}) \frac{K(X,H)\,\delta^{1/2}(H+Y)}{\delta^{1/2}(Y)}\,dH = \phi_\lambda(e^Y)\,\phi_\lambda(e^X).
\end{aligned}$$

We finish by ensuring that the kernel is Weyl-invariant in every argument. □

Corollary 2.2. *Suppose G is a complex group.*

1) *The support of the measure* $\mu_{X,Y}$ *is contained in*

$$(\cup_{w\in W} w\cdot(C(X)+Y))\cap(\cup_{w\in W} w\cdot(C(Y)+X))\subset C(X)+C(Y).$$

2) $0\notin \text{support}\,(\mu_{X,Y})$ *if and only if* $Y\notin W\cdot\{-X\}$.

Proof. 1) We note that $K(X,H)$ is strictly positive for $H\in C(X)^\circ$ and 0 on the complement of $C(X)$ and we use the symmetry of the product formula in X and Y.

2) Suppose that $0\in\text{support}(\mu_{X,Y})$. Then $0\in C(Y)+X$ and $0\in C(X)+Y$ which means that $-X\in C(Y)$ and $X\in -C(Y)=C(-Y)$. In the same way, $Y\in C(-X)$. This is only possible when Y belongs to the W-orbit of $-X$. The converse is clear.

□

Corollary 2.3. $X+Y\in\text{support}\,(\mu_{X,Y})$.

Proof. Without loss of generality we suppose that $X,\,Y\in\mathfrak{a}^+$.

Naturally, $X+Y\in C(X)+Y$. Suppose that $X+Y\in C(X)+w\cdot Y$ for $w\in W$. This means that $X-v=w\cdot Y-Y$ for a vector $v\in C(X)$. Let ${}^+\mathfrak{a}=\{H\in\mathfrak{a}\colon H=\sum_{i=1}^n c_i\,\alpha_i,\ c_i>0\}$ where $\alpha_1,\,\dots,\,\alpha_n$ are the simple roots. Recall that if $H\in\mathfrak{a}^+$ and $w\in W$ then $H-w\,H\in\overline{{}^+\mathfrak{a}}$ ([**7**, Chapter IV]). It follows that $X-v\in\overline{{}^+\mathfrak{a}}$ and $w\cdot Y-Y\in-\overline{{}^+\mathfrak{a}}\cap\overline{{}^+\mathfrak{a}}=\{0\}$, so $w\cdot Y=Y$. As $Y\in\mathfrak{a}^+$, we deduce that $w=\text{id}$.

The sets $C(X)+w\cdot Y$ being closed and bounded, it follows that a nonempty neighbourhood U of $X+Y$ is disjoint with all $C(X)+w\cdot Y$ except for $w=\text{id}$.

By Theorem 2.1, for any $H\in U\cap(C(Y)+X)^\circ$ the function

$$k(H,X,Y)=\frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)}\,\frac{1}{|W|}\,K(X,H-Y)>0.$$

Hence $X+Y\in\text{support}(k(\cdot,X,Y))$. □

Remark 2.4. If we convolve two uniform distributions on centered spheres of radii $0<r<s$ in $\mathbf{R}^n$, we obtain an absolutely continuous measure supported by the annulus of radii $s-r$ and $s+r$. Our results show that a similar property holds on symmetric spaces with G complex; however the description of the support of $\delta^\sharp_{e^X}*\delta^\sharp_{e^Y}$, the symmetric space analogue of the annulus, is more complicated.

Let us give two simple applications of our product formula.

Corollary 2.5. *Let* G *be a complex semisimple Lie group and let* μ, ν *be two* K*-biinvariant finite measures on* G *such that* $\mu(e\,K)=\nu(e\,K)=0$ *and* $\mu(K\,\partial A^+\,K)=0$ *or* $\nu(K\,\partial A^+\,K)=0$. *Then the measure* $\mu*\nu$ *is absolutely continuous.*

Proof. We identify K-biinvariant measures on G with W-invariant measures on $\mathfrak{a}$. Observe that the spherical Fourier transform of $\mu * \nu$ is equal to

$$\int_{\mathfrak{a}} \int_{\mathfrak{a}} \phi_\lambda(e^X)\, \phi_\lambda(e^Y) d\mu(X) d\nu(Y) = \hat{\gamma}(\lambda)$$

where γ is a K-biinvariant measure with density

$$d\gamma(H) = \int_{\mathfrak{a}} \int_{\mathfrak{a}} k(H, X, Y)\, d\mu(X)\, d\nu(Y).$$

The use of the Fubini theorem is justified by

$$\int_{\mathfrak{a}} k(H, X, Y)\, \delta(H)\, dH = 1$$

which is the product formula for $\lambda = -i\,\rho$ and by the boundedness of ϕ_λ. □

Corollary 2.6. *Let G be a simple complex Lie group and let $g \in K\,A^+\,K$. Then the orbit $K\,g\,K$ generates G.*

Proof. Let $g = k_1\, e^X\, k_2$ with $X \in \mathfrak{a}$. The existence of a continuous density of $\delta_X^\sharp * \delta_X^\sharp = \mu_{X,X}$ implies that $K\,g\,K\,g\,K$ contains a nonempty K-biinvariant open set. □

3. An explicit product formula for the complex groups.

The result [**4**, Proposition 2] give us a method to construct the Abel kernel K in (2) and therefore the product formula kernel k in (1).

Suppose $\alpha_1, \ldots, \alpha_q$ are the positive roots and $\alpha_1, \ldots, \alpha_n$ are the simple positive roots. We have integers $a_{kj} \geq 0$ such that

$$\alpha_k = \sum_{j=1}^{n} a_{kj}\, \alpha_j$$

for $k = n+1, \ldots, q$. For $y_1 \geq 0, \ldots, y_n \geq 0$, define

$$\Delta(y_1, \ldots, y_n) = \left\{ (y_{n+1}, \ldots, y_q) \colon y_{n+1}, \ldots, y_q \geq 0 \text{ and } \sum_{k=1}^{n} a_{kj}\, y_k \leq y_j,\ j = 1, \ldots, n \right\}.$$

We then define

$$\Psi(y_1, \ldots, y_n) = \int_{\Delta(y_1, \ldots, y_n)} dy_{n+1} \ldots dy_q \quad \text{and}$$
$$T(y_1\,\alpha_1 + \cdots + y_n\,\alpha_n) = \Psi(y_1, \ldots, y_n).$$

The support of T is $\overline{{}^+\mathfrak{a}} = \{H \in \mathfrak{a} \colon H = y_1\,\alpha_1 + \cdots + y_n\,\alpha_n,\ y_i \geq 0,\ i = 1, \ldots, n\}$. If the rank is 1, then T jumps from 1 (inside its support) to 0 (outside its support). When the rank is greater than 1, T is continuous.

It is not difficult to see that Ψ will be locally a polynomial of degree $q-n$ in $y_1, \ldots, y_n$.

Note that T is the distribution on $\mathfrak{a}$ which satisfies

$$(T, f) = \int_{\mathbf{R}_+^q} f\left(\sum_{\alpha\in\Sigma^+} x_k\,\alpha_k\right) dx_1 \ldots dx_q.$$

We have $\partial(\pi)\,T = \delta_0$ and, in particular, $\mathcal{L}(T)(\lambda) = \frac{1}{\pi(\lambda)}$.

Then

$$K(X,H) = \frac{\pi(\rho)}{\delta^{1/2}(X)} \sum_{w\in W} \epsilon(w)\,T(w\,X - H). \tag{5}$$

One of the drawbacks of the formula (4) is that it is not immediately clear that $k(H,X,Y) = k(H,Y,X)$ for every X and $Y \in \mathfrak{a}$ (it is clear from (1) that this should be the case). The following result makes this symmetry explicit.

Proposition 3.1. *Suppose G is a complex Lie group. Then the kernel $k(H,X,Y)$ of Theorem* 2.1 *can be written as*

$$k(H,X,Y) = \frac{\pi(\rho)}{|W|} \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(X)\,\delta^{1/2}(Y)} \sum_{v,w\in W} \epsilon(v)\,\epsilon(w)\,T(v\,X + w\,Y - H).$$

Proof. We have

$$\begin{aligned}
k(H,X,Y) &= \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)} \frac{1}{|W|} \sum_{w\in W} \epsilon(w)\,K(X, w\cdot H - Y) \\
&= \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)} \frac{1}{|W|} \sum_{w\in W} \epsilon(w)\,K(X, H - w^{-1}\cdot Y) \\
&= \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)} \frac{1}{|W|} \sum_{w\in W} \epsilon(w)\,\frac{\pi(\rho)}{\delta^{1/2}(X)} \\
&\qquad \cdot \sum_{v\in W} \epsilon(v)\,T(v\,X - (H - w^{-1}\cdot Y)) \\
&= \frac{\pi(\rho)}{|W|} \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(X)\,\delta^{1/2}(Y)} \sum_{v,w\in W} \epsilon(v)\,\epsilon(w)\,T(v\,X + w\,Y - H).
\end{aligned}$$

□

Definition 3.2. We will say that the function F is piecewise polynomial if there is a finite partition of support(F) into domains P satisfying $\overline{P^\circ} = P$ on which F is given by a fixed polynomial.

We will say that the function F is piecewise continuous if there is a finite partition of support(F) into domains P satisfying $\overline{P^\circ} = P$ on which F is given by a continuous function.

Corollary 3.3. *The function* $(H, X, Y) \to \delta^{1/2}(H)\,\delta^{1/2}(X)\,\delta^{1/2}(Y)\,k(H, X, Y)$ *is a piecewise polynomial continuous function on its support.*

Remark 3.4. It is interesting to note that $k(-H, X, -Y) = k(Y, X, H)$ (refer to (4)) and, in particular, that k is symmetric in H, X and Y if $-\mathrm{id} \in W$ which is the case when $G = \mathbf{SL}(2, \mathbf{C})$. It is not difficult to find examples that show that this symmetry is not true when $G = \mathbf{SL}(3, \mathbf{C})$.

Proposition 3.5. *Suppose G is a complex Lie group of rank greater than* 1.

1) *When $X \in \mathfrak{a}^+$, the function $H \to K(X, H)$ is continuous.*

   When $X \in \partial\mathfrak{a}^+ \setminus \{0\}$, the function $H \to K(X, H)$ is piecewise continuous. Moreover, if Δ_X denotes the set of all positive roots annihilating X then

$$K(X, H) = \frac{\pi(\rho) \prod_{\alpha\in\Delta_X} D_\alpha\, U(X, H)}{\prod_{\alpha\in\Delta_X} \|\alpha\|^2 \prod_{\beta\in\Delta^+\setminus\Delta_X} \sinh\langle\beta, X\rangle} \tag{6}$$

   where $U(X, H) = \sum_{w\in W} \epsilon(w)\, T(w\,X - H)$ and D_α denotes the derivative in the direction of α.

2) *When X, $Y \in \mathfrak{a}^+$, the function $H \to k(H, X, Y)$ is continuous on $\mathfrak{a}^+$ and piecewise continuous on $\partial\mathfrak{a}^+ \setminus \{0\}$.*

   When $X \in \partial\mathfrak{a}^+ \setminus \{0\}$ and $Y \in \mathfrak{a}^+$ (or vice-versa), the function $H \to k(H, X, Y)$ is piecewise continuous. Moreover, in the first case, when $H \in \mathfrak{a}^+$

$$k(H, X, Y) = \frac{\pi(\rho)}{|W|} \cdot \frac{\prod_{\alpha\in\Delta_X} D^X_\alpha V(H, X, Y)}{\delta^{1/2}(H) \prod_{\alpha\in\Delta_X} \|\alpha\|^2 \prod_{\beta\in\Delta^+\setminus\Delta_X} \sinh\langle\beta, X\rangle \prod_{\beta\in\Delta^+} \sinh\langle\beta, Y\rangle}$$

   where $V(H, X, Y) = \sum_{v,w\in W} \epsilon(v)\,\epsilon(w)\, T(v\,X + w\,Y - H)$.

Proof. 1) The only case to be considered is $X \in \partial\mathfrak{a}^+ \setminus \{0\}$, i.e., X belongs to a wall of $\mathfrak{a}^+$. In the formula we have for K:

$$K(X, H) = \frac{\pi(\rho)}{\delta^{1/2}(X)}\, U(X, H),$$

there is a singularity when $\delta^{1/2}(X) = 0$.

As written in [**4**, (8)], the (ordinary) Fourier transform of $H \to U(X, H)$ is equal, up to a constant $\frac{1}{\pi(i\lambda)}$, to the numerator

$$\sum_{w\in W} \epsilon(w)\, e^{\langle iw\cdot\lambda, X\rangle}$$

of the formula for the spherical function ϕ_λ which is equal to $\frac{1}{\pi(\rho)}\,\delta^{1/2}(X)\,\phi_\lambda(e^X)$.

The injectivity of Fourier transform and the properties of spherical functions imply that $U(X,H)=0$ for all H if and only if $\alpha(X)=0$ for a positive root α.

We know that T is continuous and piecewise polynomial, and therefore, so is $U(X,H)$. From this, one may deduce that in a neighbourhood of X, the function $U(\cdot,H)$ is a product of $\prod_{\alpha\in\Delta_X}\langle\alpha,\cdot\rangle$ and a piecewise polynomial function. The formula (6) then follows.

2) The proof is similar, using Proposition 3.1 and Remark 3.4.

□

The following examples are instructive.

1) Let $G=SL(3,\mathbf{C})$. For $X=A\alpha_1+B\alpha_2=[A,B-A,-B]$ and $H=u\,\alpha_1+v\,\alpha_2=[u,v-u,-v]$ in $\mathfrak{a}^+$, we have

$$K(X,H)=\frac{\min{}^+\{2\,A-B,A-u,B-v,2\,B-A\}}{\sinh(2\,A-B)\,\sinh(2\,B-A)\,\sinh(A+B)}.$$

Note also that if $H\in C(X)^\circ$, we have $u<A$ and $v<B$ (see Lemma 4.1).

Now, take any $X\neq 0$ in $\{\alpha_1=0\}\cap\overline{\mathfrak{a}^+}$. We then have $X=x\alpha_1+2\,x\,\alpha_2$ with $x>0$. If we fix $H\in\mathfrak{a}^+$ with $u<x$ and $v<2\,x$, Proposition 3.5 tells us that

$$K(X,H)=\frac{1}{\sinh^2(3\,x)}.$$

That shows that $H\to K(X,H)$ is not continuous on $\partial C(X)$ since $K(X,H)=0$ for H outside $C(X)$.

2) When $X,Y\in\mathfrak{a}^+$, $H\to k(H,X,Y)$ may not be continuous on $\overline{\mathfrak{a}^+}$ (consider for example $X=[4,3,-7]$, $Y=[6,-2,-4]$ and $H=[2,2,-4]$ on $\mathbf{SL}(3,\mathbf{C})/\mathbf{SU}(3)$).

Let us now consider an example where K and k are easy to compute. If $G=\mathbf{SL}(2,\mathbf{C})$, we have $T(X)=1$ if $X\in\mathfrak{a}^+$ and 0 otherwise. This means that for X and $H\in\mathfrak{a}^+$, we have

$$\begin{aligned}K(X,H)&=\frac{\pi(\rho)}{\delta^{1/2}(X)}\,(T(X-H)-T(-X-H)),\\&=\frac{\pi(\rho)}{\delta^{1/2}(X)}\qquad\text{if } X_2\le H_1<X_1 \text{ and 0 otherwise}\end{aligned}$$

and therefore if X, Y and $H \in \mathfrak{a}^+$,

$$k(H,X,Y) = \frac{\pi(\rho)}{\delta^{1/2}(H)\,\delta^{1/2}(X)\,\delta^{1/2}(Y)} \quad \text{if } |X_1 - Y_1| < |H_1| \le X_1 + Y_1 \text{ and 0 otherwise.}$$

This formula is given in [**8**, p. 369].

However, even for $\mathbf{SL}(n,\mathbf{C})$, the computations become quickly onerous when $n > 3$. We will discuss the case $\mathbf{SL}(3,\mathbf{C})$ in the next section.

4. The support in the case of $\mathbf{SL}(3,\mathbf{C})$.

In this section, we will assume throughout that $G = \mathbf{SL}(3,\mathbf{C})$. In this case, we have $T(X) = \min^+\{X_1, -X_3\}$ ($n = 2$ and $q = 3$) which brings

$$K(X,H) = \frac{\pi(\rho)}{\delta^{1/2}(X)} \min^+\big\{X_1 - X_2, X_2 - X_3, X_1 - H_2, X_1 - H_2, X_1 - H_3, H_1 - X_3, H_2 - X_3, X_3 - Y_3\big\}.$$

Pictures of the support of the measure $\mu_{X,Y}$ are shown in Figure 1 (two cases are shown).

Figure 1. The support of $\mu_{X,Y}$.

The following result will be used repeatedly in what follows to determine under which conditions an element H belongs to a set of the form $C(X)+Y$ with $X \in \mathfrak{a}^+$.

Lemma 4.1. *Suppose* $X \in \mathfrak{a}^+$. *Then* $C(X) = \{H \in \mathfrak{a}\colon X_3 \le H_i \le X_1,\ i = 1,2,3\}$ *and* $C(X)^\circ = \{H \in \mathfrak{a}\colon X_3 < H_i < X_1,\ i = 1,2,3\}$.

Proof. The sides of $C(X) \cap \mathfrak{a}^+$ which do not lie on the axes of symmetry belonging to W are given by $H_3 = X_3$ and $H_1 = X_1$. Since the coordinates of the origin satisfy $0 > X_3$ and $0 < X_1$, we have $\mathfrak{a}^+ \cap C(X) = \{H \in \mathfrak{a}^+ \colon H_3 \geq X_3, H_1 \leq X_1\}$. The result follows by invariance under W; the elements of W act on $H = (H_1, H_2, H_3)$ by permuting the indices. □

Lemma 4.2. *Suppose X and $Y \in \mathfrak{a}^+$. Then*

$$(\cup_{w\in W} w \cdot (C(X) + Y)) \cap (\cup_{w\in W} w \cdot (C(Y) + X)) \cap \mathfrak{a}^+ \\ = (C(X) + Y) \cap (C(Y) + X) \cap \mathfrak{a}^+.$$

Proof. Clearly, the set on the right hand side is included in the set on the left hand side.

Let $H \in ((\cup_{w\in W} w \cdot (C(X) + Y)) \cap (\cup_{w\in W} w \cdot (C(Y) + X))) \cap \mathfrak{a}^+$. We have

$$\begin{aligned} X_3 \leq \quad & H_i - Y_{w(i)} \quad \leq X_1, \\ Y_3 \leq \quad & H_i - X_{v(i)} \quad \leq Y_1 \end{aligned}$$

where $i = 1, \ldots, 3$ and w and $v \in W = S_3$. Recall that $H_1 > H_2 > H_3$, $X_1 > X_2 > X_3$ and $Y_1 > Y_2 > Y_3$. We have:

1) $H_1 - Y_1 \leq H_1 - Y_{v(1)} \leq X_1$.
2) $H_2 - Y_2 \leq H_2 - Y_{v(2)} \leq X_1$ if $v(2) = 2$ or 3. If $v(2) = 1$ then $v(1) = 2$ or 3. We then have $H_2 - Y_2 \leq H_1 - Y_{v(1)} \leq X_1$.
3) Let i be such that $v(i) = 3$. Then $H_3 - Y_3 \leq H_i - Y_{v(i)} \leq X_1$.

Using a similar approach, we show that $H_i - Y_i \geq X_3$ for each i and therefore, $H \in C(X) + Y$. In the same manner, $H \in C(Y) + X$. □

Note that

$$((C(X) + Y) \cap (C(Y) + X) \cap \mathfrak{a}^+)^\circ = (C(X)^\circ + Y) \cap (C(Y)^\circ + X) \cap \mathfrak{a}^+.$$

Lemma 4.3. *Let $X, Y \in \mathfrak{a}^+$. Suppose $H \in (C(X)^\circ + Y) \cap (C(Y)^\circ + X) \cap \mathfrak{a}^+$. Then one of the following is true.*

1) *H belongs to no other $C(X)^\circ + w \cdot Y$.*
2) *H belongs to no other $C(Y)^\circ + v \cdot X$.*
3) *H belongs to exactly one other $C(X)^\circ + w \cdot Y$, $w \in W$.*
4) *H belongs to exactly one other $C(Y)^\circ + v \cdot X$, $v \in W$.*

Proof. Suppose the result is not true. This means that we can find $H \in (C(X)^\circ + Y) \cap (C(X)^\circ + w_1 \cdot Y) \cap (C(X)^\circ + w_2 \cdot Y) \cap (C(Y)^\circ + X) \cap (C(Y)^\circ + v_1 \cdot X) \cap (C(Y)^\circ + v_2 \cdot X) \cap \mathfrak{a}^+$ with $w_1 \neq e$, $w_2 \neq e$, $w_1 \neq w_2$ and $v_1 \neq e$, $v_2 \neq e$, $v_1 \neq v_2$.

In that case, we can find $i < 3$ such that $w_1(i) = 3$ or $w_2(i) = 3$ (aside from the identity, there is only one element of $W = S_3$ that fixes any given index). In the same way, we can find $j > 1$ such that $v_1(j) = 1$ or $v_2(j) = 1$.

To simplify the notation, assume that $w_1(i) = 3$ and $v_1(j) = 1$. This means that $i \leq j$.

We have $H_i - Y_3 = H_i - Y_{w_1(i)} < X_1$ since $H \in C(X)^\circ + w_1 \cdot Y$ and $H_j - X_1 = H_j - X_{v_1(j)} > Y_3$ since $H \in C(Y)^\circ + v_1 \cdot X$. This means that $X_1 < H_j - Y_3$. Therefore $X_1 < H_j - Y_3 \leq H_i - Y_3 < X_1$ (recall that $i \leq j$) which is absurd. □

Proposition 4.4. *Suppose X, $Y \in \mathfrak{a}^+$. Let*

$$S = (C(X)^0 + Y) \cap (C(Y)^\circ + X) \cap \mathfrak{a}^+.$$

Let $H \in \mathfrak{a}^+$. Then $k(H, X, Y)$ is nonzero (and therefore strictly positive) if and only if

$$H \in S \cap \{H_3 < X_2 + Y_2\} \cap \{H_1 > X_2 + Y_2\}.$$

Note that if X and Y are both above ρ (i.e., $X_2 \geq 0$ and $Y_2 \geq 0$) then the condition $H_3 < X_2 + Y_2$ is automatically satisfied for $H \in \mathfrak{a}^+$. In the same manner, if X and Y are both below ρ (i.e., $X_2 \leq 0$ and $Y_2 \leq 0$) then the condition $H_1 > X_2 + Y_2$ is automatically satisfied for $H \in \mathfrak{a}^+$.

Proof. If we refer to Corollary 2.2 and to Lemma 4.2, we can assume that $H \in (C(X) + Y) \cap (C(Y) + X) \cap \mathfrak{a}^+$ since otherwise $k(H, X, Y) = 0$.

Let S_0 be the set consisting of $H \in (C(X)^\circ + Y) \cap (C(Y)^\circ + X) \cap \mathfrak{a}^+$ such that H belongs to no other $C(X)^\circ + w \cdot Y$ or to no other $C(Y)^\circ + v \cdot X$, v, $w \in W$. For $i = 1$ and 2, let S_i be the set consisting of $H \notin S_0$ and $H \in (C(X) + Y)^\circ \cap (C(Y) + X)^\circ \cap (C(X)^\circ + w_i \cdot Y) \cap (C(Y)^\circ + w_i \cdot X) \cap \mathfrak{a}^+$ where $w_1 = (1 \to 1, 2 \to 3, 3 \to 2)$ and $w_2 = (1 \to 2, 2 \to 1, 3 \to 3) \in W$.

We will show that for $H \in \mathfrak{a}^+$, $k(H, X, Y) > 0$ if and only if

$$H \in S_0 \cup (S_1 \cap \{H_3 < X_2 + Y_2\}) \cup (S_2 \cap \{H_1 > X_2 + Y_2\}).$$

This will prove the result once we observe the following two facts:

1) If $H \in (C(X) + Y) \cap (C(Y) + X) \cap \mathfrak{a}^+$ does not belong to S_1 then $H_3 < X_2 + Y_2$.

   It is sufficient to prove that $H \in (C(X) + Y) \cap (C(Y) + X) \cap \mathfrak{a}^+$ and $H_3 \geq X_2 + Y_2$ imply that $H \in C(Y)^\circ + w_1 \cdot X$. Then, by symmetry of the above expressions in X and Y, we will also have $H \in C(X) + w_1 \cdot Y$ and therefore $H \in S_1$.

   We note that $H \in C(Y)^\circ + w_1 \cdot X$ is equivalent to the inequalities: $Y_3 < H_1 - X_1 < Y_1$, $Y_3 < H_2 - X_3 < Y_1$ and $Y_3 < H_3 - X_2 < Y_1$.

   The first inequality is obvious since $H \in C(Y) + X^\circ$, $Y_3 < H_2 - X_3$ is true since $H_2 > H_3 \geq X_2 + Y_2 > X_3 + Y_3$. Suppose $H_2 - X_3 < Y_1$ is false. Then $-H_1 - H_3 = H_2 \geq X_3 + Y_1$ and $H_3 \leq -X_3 - Y_1 - H_1$ which combined with $H_3 \geq X_2 + Y_2$ yields $X_2 + Y_2 \leq -X_3 - Y_1 - H_1$ or $X_2 + X_3 + Y_1 + Y_2 \leq -H_1$, *i.e.*, $-X_1 - Y_3 \leq -H_1$ which contradicts $H_1 - X_1 > Y_3$ since $H \in C(Y) + X$. Finally, $H_3 - X_2 < Y_1$ holds because $H_3 - X_2 < H_2 - X_2 < Y_1$.

2) If $H \in (C(X)+Y)\cap(C(Y)+X)\cap\mathfrak{a}^+$ does not belong to S_2 then $H_1 > X_2+Y_2$.

The proof is similar.

Consider now Lemma 4.3. If Cases 1) or 2) are verified, then $H \in S_0$. In that case, we either have $k(H,X,Y) = \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)}\,\frac{1}{|W|}\,K(X,H-Y) > 0$ or $k(H,X,Y) = \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)}\,\frac{1}{|W|}\,K(Y,H-X) > 0$.

If $H \notin S_0$ then H satisfies Cases 3) and 4) of Lemma 4.3 and we have $H \in (C(X)+Y)^\circ\cap(C(Y)+X)^\circ\cap(C(X)^\circ+w\cdot Y)\cap(C(Y)^\circ+v\cdot X)\cap\mathfrak{a}^+$. Note that we cannot have $w(1)=3$ or $w(3)=1$ (and similarly for v). Indeed, if $w(1)=3$ then $H_1-Y_3<X_1$ which means that $H_1-X_1<Y_3$ which is absurd while if $w(3)=1$ then $H_3-X_1>Y_3$ which means that $H_3-Y_3>X_1$ which is absurd. Therefore, the only possibilities for w and v are w_1 and w_2. We also have $v=w$. Indeed, if we had $v\neq w$, it is not difficult to see by inspection (say by taking $w=w_1$ and $v=w_2$) that we would reach a contradiction by using a similar argument. We then have $k(H,X,Y) = \frac{1}{\delta^{1/2}(H)\,\delta^{1/2}(Y)}\,\frac{1}{|W|}\,(K(X,H-Y)-K(X,H-w_i\cdot Y)) > 0$ since $\epsilon(w_1)=\epsilon(w_2)=-1$.

It remains to show that for $H\in S_1$, $k(H,X,Y)>0$ if and only if $H_3 < X_2+Y_2$ and that for $H\in S_2$, $k(H,X,Y)>0$ if and only if $H_1 > X_2+Y_2$. Since the reasoning in the two cases are very similar, we will show only the first case.

Suppose $H\in S_1$. We deduce easily that $X_1+Y_3-H_2$ is strictly smaller than X_1-X_2, $X_1+Y_2-H_2$, $X_1+Y_3-H_3$ and $X_1+Y_2-H_3$ while $H_3-Y_2-X_3$ is strictly smaller than $X_1+Y_1-H_1$, $H_2-Y_2-X_3$, $H_3-Y_3-X_3$ and $H_2-Y_3-X_3$. This implies that $K(X,H-Y)-K(X,H-w_1\cdot Y)>0$ is equivalent to

$$\min\{X_2-X_3, H_1-Y_1-X_3\} > \min\{X_1+Y_3-H_2, H_3-Y_2-X_3\}. \tag{7}$$

Note that $X_2-X_3>H_3-Y_2-X_3$ and $H_1-Y_1-X_3>X_1+Y_3-H_2$ are both equivalent to $H_3<X_2+Y_2$. The latter inequality is therefore sufficient for (7) to be true. It remains to show that it is necessary.

Now, we get down to several cases:

1) Suppose $X_1-X_2\le X_2-X_3$ and $Y_1-Y_2\le Y_2-Y_3$ (i.e., $X_2\ge 0$ and $Y_2\ge 0$). Since $H\in\mathfrak{a}^+$, the condition $H_3<X_2+Y_2$ is satisfied and there is nothing to prove.
2) Suppose $X_1-X_2\le X_2-X_3$ and $Y_1-Y_2>Y_2-Y_3$ (i.e., $X_2\ge 0$ and $Y_2<0$).

   Suppose $H_1-Y_1-X_3>\min\{X_1+Y_3-H_2, H_3-Y_2-X_3\}$ and $H_3\ge X_2+Y_2$. That is only possible if $H_1-Y_1>H_3-Y_2$.

   We have $H_1-Y_1>H_3-Y_2\ge X_2+Y_2-Y_2=X_2\ge 0$. Now, $H_3\ge X_2+Y_2$ if and only if $-Y_2\ge -H_3+X_2$ which implies $Y_1+Y_3=-Y_2>-H_3=H_1+H_2$ which is equivalent to $Y_1>H_1+(H_2-Y_3)>H_1$

since $H_2 - Y_3 > X_2 \geq 0$ ($H_2 - X_2 > Y_3$ since $H \in C(Y)^\circ + X$). This contradicts $H_1 - Y_1 > 0$.

3) By symmetry, we do not have to consider the case $Y_1 - Y_2 \leq Y_2 - Y_3$ and $X_1 - X_2 > X_2 - X_3$.
4) Suppose $X_1 - X_2 > X_2 - X_3$ and $Y_1 - Y_2 > Y_2 - Y_3$ (i.e., $X_2 < 0$ and $Y_2 < 0$).

Suppose that (7) is true and that $H_3 \geq X_2 + Y_2$. This means that $X_2 - X_3 \leq H_3 - Y_2 - X_3$ and $H_1 - Y_1 - X_3 \leq X_1 + Y_3 - H_2$.

We consider two cases:

a) $H_1 - Y_1 - X_3 \geq X_2 - X_3 > X_1 + Y_3 - H_2$:
We have $H_1 - Y_1 - X_3 \geq X_2 - X_3$ if and only if $H_1 - Y_1 \geq X_2$. On the other hand, $H_3 \geq X_2 + Y_2$ if and only if $-X_2 + Y_1 + Y_3 \geq H_1 + H_2$. To this last inequality, we apply $X_2 - X_3 > X_1 + Y_3 - H_2$ if and only if $H_2 > X_1 + Y_3 - X_2 + X_3$. We obtain $-X_2 + Y_1 + Y_3 > H_1 + X_1 + Y_3 - X_2 + X_3$ if and only if $Y_1 > H_1 + X_1 + X_3 = H_1 - X_2$ if and only if $X_2 > H_1 - Y_1$. This contradicts $H_1 - Y_1 \geq X_2$.

b) $X_2 - X_3 > H_1 - Y_1 - X_3 > H_3 - Y_2 - X_3$:
We have $X_2 - X_3 > H_1 - Y_1 - X_3$ if and only if $X_2 > H_1 - Y_1$. On the other hand, $H_1 - Y_1 - X_3 > H_3 - Y_2 - X_3$ if and only if $-H_3 > -H_1 + Y_1 - Y_2$. We have $H_3 \geq X_2 + Y_2$ if and only if $-X_2 > -H_3 + Y_2 > -H_1 + Y_1 - Y_2 + Y_2 = -H_1 + Y_1$ which is equivalent to $X_2 < H_1 - Y_1$. This contradicts $X_2 > H_1 - Y_1$.

□

Remark 4.5. Let H, X and $Y \in \mathfrak{a}^+$. We note that computing $k(H, X, Y)$ requires one evaluation of K when $H \in S_0$ while it requires taking the difference of two values of K when $H \in S_1 \cap \{H_3 < X_2 + Y_2\}$ or $H \in S_2 \cap \{H_1 > X_2 + Y_2\}$.

Remark 4.6. When we refer to Figure 1, we can describe the support in a more informal and more concrete manner:

$$\text{support}(\mu_{X,Y}) = C \setminus (D_1 \cup D_2)$$

where $C = (\cup_{w \in W} w \cdot (C(X) + Y)) \cap (\cup_{w \in W} w \cdot (C(Y) + X))$ and D_1, D_2 are either empty or equilateral triangles in the plane $\mathfrak{a}$ such that 0 is their centre and such that a side is, respectively, on the line $-v_H = H_3 = X_2 + Y_2 < 0$ ($u_H = H_1 = X_2 + Y_2 > 0$).

Naturally, $D_1 = W \cdot (\{H_3 > X_2 + Y_2\} \cap \mathfrak{a}^+)$ and $D_2 = W \cdot (\{H_1 < X_2 + Y_2\} \cap \mathfrak{a}^+)$.

5. The function T in the case of $\mathbf{SL}(n, \mathbf{C})$.

By Proposition 3.1, in order to know the kernel $k(H, X, Y)$ of the product formula, it is sufficient to know explicitly the function T defined in Section 3.

We give here some more information available about the function T in the case $G = \mathbf{SL}(n, \mathbf{C})$.

Note that when writing $\lambda \in \mathfrak{a}^*$ in terms of the simple positive roots, *i.e.*, $\lambda = \sum_{i=1}^{n-1} a_i \alpha_i$, we find that

$$\pi(\lambda) = \prod_{\alpha>0} \langle \lambda, \alpha \rangle = \prod_{i=1}^{n-1} \left[a_i \left(a_i + a_{i+1} \right) \dots \left(a_i + \dots + a_{n-1} \right) \right]. \tag{8}$$

Using Maple, it is possible to compute the function T for $\mathbf{SL}(4, \mathbf{C})$ since it is simply a matter of computing the Laplace inverse transform of $\frac{1}{\pi(\lambda)}$.

Recall that $T(y_1 \alpha_1 + y_2 \alpha_2 + y_3 \alpha_3) = 0$ unless all y_i's are positive. Let $x_+ = \max\{0, x\}$. We find

$$T(y_1 \alpha_1 + y_2 \alpha_2 + y_3 \alpha_3) = \begin{cases} y_2^3 & 0 \le y_2 \le \min\{y_1, y_3\} \\ -2\, y_1^3 + 3\, y_1^2\, y_2 & 0 \le y_1 \le y_2 \le y_3 \\ -2\, y_3^3 + 3\, y_3^2\, y_2 & 0 \le y_3 \le y_2 \le y_1 \\ -y_1^3 + 3\, y_1^2\, y_3 - (y_1 + y_2 - y_3)_+^3 & 0 \le y_1 \le y_3 \le y_2 \\ -y_3^3 + 3\, y_3^2\, y_1 - (y_1 + y_2 - y_3)_+^3 & 0 \le y_3 \le y_1 \le y_2. \end{cases}$$

Here is a more general result for the function T:

Proposition 5.1. *The function T for $\mathbf{SL}(n, \mathbf{C})$ is given by*

$$\begin{aligned} T(y_1 \alpha_1 + \dots + y_{n-1} \alpha_{n-1}) &= \mathbf{1}_{\{0 \le y_1\}}\, \delta_0(dy_2, \dots, dy_{n-1}) * \mathbf{1}_{\{y_1 \le y_2\}}\, \delta_0(dy_3, \dots, dy_{n-1}) \\ &\quad * \mathbf{1}_{\{y_1 \le y_2 \le y_3\}}\, \delta_0(dy_4, \dots, dy_{n-1}) \\ &\quad * \dots * \mathbf{1}_{\{y_1 \le y_2 \le \dots \le y_{n-2}\}}\, \delta_0(dy_{n-1}) * \mathbf{1}_{\{y_1 \le y_2 \le \dots \le y_{n-1}\}}. \end{aligned}$$

Proof. If we consider (8), we can write $\frac{1}{\pi(\lambda)}$ as

$$\frac{1}{\pi(\lambda)} = \prod_{k=1}^{n-1} \left(\frac{1}{a_k \left(a_{k-1} + a_k \right) \left(a_{k-2} + a_{k-1} + a_k \right) \dots \left(a_1 + a_2 + \dots + a_{k-1} + a_k \right)} \right)$$

and then compute the inverse Laplace transform of each factor. □

Lemma 5.2. *If $X \in \mathfrak{a}^+$ then $C(X) = \{H\colon \sum_{i=1}^r H_{k_i} \le \sum_{i=1}^r X_i, (k_i) \in S_n, r \le n-1\}$.*

Proof. Similar to the proof of Lemma 4.1. □

Corollary 5.3. *On $\mathbf{SL}(n, \mathbf{C})$, the convex envelope of the support of $\mu_{X,Y}$ is $C(X+Y) = C(X) + C(Y)$.*

Proof. One observes easily that $C(X+Y) = C(X) + C(Y)$ using the above lemma. We then use Corollary 2.2, Corollary 2.3 and the fact that the support is Weyl invariant. □

References

[1] R. Askey, *Jacobi polynomials,* I. *New proofs of Koornwinder Laplace type integral representation and Bateman's bilinear sum*, SIAM J. Math. Anal., **5**(**1**) (1974), 119-124, MR 52 #6062, Zbl 0269.33014.

[2] M. Flensted-Jensen and T. Koornwinder, *The convolution structure for Jacobi expansions*, Ark. Mat., **10** (1973), 245-262, MR 49 #5688, Zbl 0267.42009.

[3] M. Flensted-Jensen and D.L. Ragozin, *Spherical functions are Fourier transforms of L_1-functions*, Annales scientifiques de l'École Normale Supérieure, **6** (1973), 457-458, MR 51 #807, Zbl 0293.22020.

[4] P. Graczyk and J.-J. Loeb, *Spherical analysis and central limit theorems on symmetric spaces*, Probability measures on groups and related structures, **XI** (Oberwolfach, 1994), 146-166, World Sci. Publishing, River Edge, NJ, 1995, MR 98b:43017, Zbl 0907.43011.

[5] P. Graczyk and P. Sawyer, *The product formula for the spherical functions on symmetric spaces of noncompact type*, preprint, **124**, Département de Mathématiques, Université d'Angers, January 2001.

[6] S. Helgason, *Differential Geometry, Lie Groups and Symmetric Spaces*, Academic Press, New York, 1978, MR 80k:53081, Zbl 0451.53038.

[7] ———, *Groups and Geometric Analysis*, Academic Press, New York, 1984, MR 86c:22017, Zbl 0543.58001.

[8] ———, *Geometric analysis on symmetric spaces*, Mathematical surveys and monographs, **39**, American Mathematical Society, 1994, MR 96h:43009, Zbl 0809.53057.

[9] T. Koornwinder, *The addition formula for Jacobi polynomials,* I. *Summary of results*, Indag. Math., **34** (1972), 188-191, MR 46 #7590, Zbl 0247.33017.

[10] ———, *The addition formula for Jacobi polynomials and spherical harmonics*, SIAM J. Appl. Math., **25**(**2**) (1973), 236-246, MR 49 #10938, Zbl 0276.33023.

[11] ———, *Jacobi polynomials,* II. *An analytic proof of the product formula*, SIAM J. Math. Anal., **5**(**1**) (1974), 125-137, MR 52 #6063, Zbl 0269.33015.

[12] ———, *A new proof of a Paley-Wiener type theorem for the Jacobi transform*, Archiv für Mathematik, **13** (1975), 145-159, MR 51 #11028, Zbl 0303.42022.

[13] ———, *Jacobi functions and analysis on noncompact semisimple Lie groups. Special functions: Group theoretical aspects and application*, R.A. Askey & al. (eds), Reidel, 1984, MR 86m:33018, Zbl 0584.43010.

[14] I.V. Ostrovskii, *Description of the class I_0 in a special semigroup of probability measures*, Selected Transl. in Math. Statist. and Prob., **15** (1981), 1-8, MR 47 #9680, Zbl 0292.60033.

[15] I.P. Trukhina, *Arithmetic of spherically symmetric measures on Lobatchevsky space* (in Russian), Teor. Fun'kcii, Funkc. Anal. Pril., **34** (1980), 136-146, MR 81h:60017, Zbl 0444.28011.

[16] M. Voit, *Factorization of probability measures on symmetric hypergroups*, J. Austral. Math. Soc., **A50** (1991), 417-467, MR 92i:60015, Zbl 0731.60008.

Received October 12, 2000 and revised June 15, 2001. The first author is supported by the European Commission (TMR 1998-2001 Network Harmonic Analysis). The second author is supported by a grant from NSERC.

Département de Mathématiques
Université d'Angers
2 Boulevard Lavoisier
49045 Angers cedex 01, France
E-mail address: graczyk@tonton.univ-angers.fr

Department of Mathematics & Computer Science
Laurentian University
Sudbury, Ontario
Canada P3E 5C6
E-mail address: sawyer@cs.laurentian.ca

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

DISCRETE BISPECTRAL DARBOUX TRANSFORMATIONS FROM JACOBI OPERATORS

F. Alberto Grünbaum and Milen Yakimov

We construct families of bispectral difference operators of the form $a(n)T + b(n) + c(n)T^{-1}$ where T is the shift operator. They are obtained as discrete Darboux transformations from appropriate extensions of Jacobi operators. We conjecture that along with operators previously constructed by Grünbaum, Haine, Horozov and Iliev they exhaust all bispectral regular (i.e., $a(n) \neq 0, c(n) \neq 0, \forall n \in \mathbb{Z}$) operators of the form above.

1. Introduction.

Back in 1929 S. Bochner [**6**] posed and solved the problem of isolating all families of orthogonal polynomials that are also eigenfunctions of a fixed, but arbitrary, second order differential operator. He found that they were given by what are nowadays called "the classical orthogonal polynomials", i.e., those of Jacobi, Hermite, Laguerre and (the less known) Bessel. Many developments in the last few years which establish rich links between classical function theory at one end and differential algebra at the other, can be seen as the result of looking for answers to questions that are variants of that of Bochner. Some of these developments are alluded to in the rest of the introduction. Before going into details it is probably worth noticing that while the original paper of Bochner poses and solves the problem in a few pages, the extensions that have been considered in the last 15 years or so are still awaiting complete resolution. This paper takes a step in that direction.

The bispectral problem, as originally formulated by Duistermaat and Grünbaum [**7**], asks for a description of all situations where a pair of differential operators in the variables x and z have a common eigenfunction $\Psi(x, z)$

$$L(x, \partial_x)\Psi(x, z) = \lambda(z)\Psi(x, z), \tag{1.1}$$

$$B(z, \partial_z)\Psi(x, z) = \theta(x)\Psi(x, z). \tag{1.2}$$

For simplicity we say that L, or B, or Ψ, are bispectral when the situation above holds.

The results in [**7**] already revealed a number of interesting connections with a variety of topics ranging from the Korteweg–deVries equation to

the problem of isomonodromic deformations for differential operators with rational coefficients. Later even more unexpected connections with different areas of pure mathematics were found. These include automorphisms and ideal structure of the Weyl algebra in one variable [**5**, **3**], representations of the $W_{1+\infty}$ algebra [**2**], Calogero–Moser system [**25**], Huygens' principle [**4**], traces of intertwiners for representations of (quantized) simple Lie algebras [**8**, **9**] (the last two in the multivariable case).

In [**7**] all bispectral differential operators $L(x, \partial_x)$ of second order were classified. Notice that if one insists that $B(z, \partial_z)$ should also be of order two then one is necessarily dealing with the Bessel or Airy cases. In that paper very explicit use was made of the Darboux transformation mapping a given second order differential operator into another one. When starting from an appropriate bispectral $L(x, \partial_x)$ this was shown to produce another such, with a different $B(z, \partial_z)$. Wilson [**24**] approached the problem from the viewpoint of commutative algebras of differential operators. He classified all maximal bispectral algebras of rank one (which by definition is the greatest common divisor of the orders of all operators in the algebra). In [**1**, **18**] the idea of applying Darboux transformations to commutative algebras of differential operators was developed. This allowed for a unification of the apparently unrelated methods in [**7**, **24**] and an extension of them to the higher rank case. Further interesting results in this direction were obtained in [**17**].

Grünbaum and Haine considered [**10**] a discrete–differential version of the above problem when the variable x runs over the integer lattice $\mathbb{Z}$ and accordingly one replaces the differential operator $L(x, \partial_x)$ by a difference operator

$$L(n, T) = \sum_{i=p}^{q} b_i(n) T^i, \quad b_p(n), b_q(n) \not\equiv 0$$

acting on a function $f(n)\colon \mathbb{Z} \to \mathbb{C}$ by

$$(Lf)(n) = \sum_{i=p}^{q} b_i(n) f(n+i).$$

Following [**23**, **22**], we define the support of $L(n, T)$ to be the ordered pair $[p, q]$. *Such a difference operator will be called regular if the first and the last coefficients $b_q(n)$ and $b_p(n)$ are nowhere vanishing functions on $\mathbb{Z}$.*

As indicated above, this problem is a generalization of the problem of classifying orthogonal polynomials which are eigenfunctions of differential operators. The point is that the standard three term recursion relation gives rise to a very special type of difference operator, represented by a *semiinfinite tridiagonal* matrix. In [**10**], Grünbaum and Haine showed that all instances of difference operators with support $[-1, 1]$ and second order

differential operators satisfying (1.1)-(1.2) result by replacing the variable n in the classical cases of the Hermite, Laguerre, Jacobi, and Bessel polynomials (discovered by Bochner, see [**6**]) by a variable $n+\varepsilon$ with n running over the integer lattice and ε arbitrary (see Section 2.2 for precise definitions). The differential operator in z is the celebrated hypergeometric second order differential operator of Gauss. It is worth noting that the corresponding eigenfunctions $\Psi(n,z)$ are no longer polynomials. For a recent survey of this area, see [**14**].

Recently Haine and Iliev considered the classification problem for maximal bispectral difference algebras of rank one [**15**]. Their treatment is a beautiful extension of Wilson's work [**24**] where the Grassmannians associated to Darboux transformations on differential operators are substituted with flag varieties coming from such transformations on difference operators. Among these some algebras that contain an operator with support $[-1,1]$ were isolated in [**16**], where they were conjectured to be all of this type.

The aim of this paper is to make progress in obtaining a discrete–continuous analog of the result of [**7**], namely a classification of all discrete bispectral operators of the form $a(n)T+b(n)+c(n)T^{-1}$ (referred to as the extended Bochner–Krall problem in [**14**]). In [**12**] the Darboux process was applied to a biinfinite extension of the Laguerre difference operators considered in [**10**]. A large class of bispectral difference operators of the form above was thus constructed and many properties of the resulting objects were analyzed in detail. It is fair to say that the results in [**12**] provide a general treatment of the Laguerre case. The case of Jacobi difference operators has so far not been amenable to a similar treatment and only some special cases of Darboux maps were proved to preserve the bispectral property. The goal of the present paper is to provide such a general treatment in the Jacobi case and to state a conjecture for the classification problem above.

The rest of the introduction describes our results.

We take as a starting point the following natural extensions of the Jacobi polynomials, constructed in [**10**]:

$$p_\varepsilon^{\alpha,\beta}(n,z)=\frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n}F(-(n+\varepsilon),n+\varepsilon+\alpha+\beta+1,\alpha+1,(1-z)/2).$$

Here and later we use F for Gauss' ${}_2F_1$ hypergeometric function. For negative integer values of α, see (2.17). They are no longer polynomials but are still eigenfunctions of a biinfinite difference operator $L_{\alpha,\beta;\varepsilon}(n,T)$ of the form $a_0(n)T+b_0(n)+c_0(n)T^{-1}$ and a differential operator $B_{\alpha,\beta}(z,\partial_z)$:

$$L_{\alpha,\beta;\varepsilon}(n,T)p_\varepsilon^{\alpha,\beta}(n,z)=zp_\varepsilon^{\alpha,\beta}(n,z),$$
$$B_{\alpha,\beta}(z,\partial_z)p_\varepsilon^{\alpha,\beta}(n,z)=\lambda_\varepsilon(n)p_\varepsilon^{\alpha,\beta}(n,z).$$

The operator $L_{\alpha,\beta;\varepsilon}(n,T)$ is obtained by the formal change of variables $n\mapsto n+\varepsilon$ from the standard (difference) Jacobi operator and is explicitly defined

in (2.9). The operator $B_{\alpha,\beta}(z,\partial_z)$ and the spectral function $\lambda_\varepsilon(n)$ are given in Equations (2.5) and (2.12).

The sets of difference operators that we consider are obtained by the following version of the Darboux map starting from the operators $L_{\alpha,\beta;\varepsilon}(n,T)$. Let $P(n,T)$ be a regular difference operator whose kernel is preserved by $L_{\alpha,\beta;\varepsilon}(n,T)$. Then there exists a (unique) difference operator $L(n,T)$ such that

$$L(n,T)P(n,T) = P(n,T)L_{\alpha,\beta;\varepsilon}(n,T) \tag{1.3}$$

which we refer to as *a Darboux transformation from* $L_{\alpha,\beta;\varepsilon}(n,T)$. The advantage of this version is that this $L(n,T)$ is necessarily of the same form as $L_{\alpha,\beta;\varepsilon}(n,T)$, i.e., $L(n,T) = a(n)T + b(n) + c(n)T^{-1}$ for some functions $a(n)$, $b(n)$, $c(n)$, $n \in \mathbb{Z}$.

If $q(x)$ denotes the characteristic polynomial of the endomorphism $L(n,T)$ acting on the finite dimensional space $\mathrm{Ker}P(n,T)$, then

$$\mathrm{Ker}P(n,T) \subset \mathrm{Ker}q(L_{\alpha,\beta;\varepsilon}(n,T)).$$

In view of this it is natural to parametrize the sets of operators $L(n,T)$ by the Grassmannians of special subspaces of $\mathrm{Ker}q(L_{\alpha,\beta;\varepsilon}(n,T))$ that can occur as $\mathrm{Ker}P(n,T)$. Denote the set of difference operators $L(n,T)$ corresponding to characteristic polynomial $q(x) = (x-1)^k(x+1)^l$ by

$$\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}.$$

The operators in $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ are the main objects of study in this paper. Their explicit form is given in Section 3.3. Restricting to $q(x)$ with roots at ± 1 guarantees that $L(n,T)$ will have rational coefficients. This is an important feature of bispectral operators. See [**7**] in the differential case.

It is an easy consequence of (1.3) that the function

$$\Psi(n,z) = P(n,T)p^{\alpha,\beta}_\varepsilon(n,z) \tag{1.4}$$

is an eigenfunction of the operator $L(n,T)$, namely we have

$$L(n,T)\Psi(n,z) = z\Psi(n,z).$$

Our main result is:

Theorem 1.1. *The difference operators $L(n,T)$ from the sets $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ are bispectral (or more precisely the functions $\Psi(n,z)$ (1.4) are eigenfunctions of differential operators in the variable z) in the following cases:*

1) *$\alpha \in \mathbb{Z}$ and $k \leq |\alpha|$, $l = 0$,*
2) *$\beta \in \mathbb{Z}$ and $l \leq |\beta|$, $k = 0$,*
3) *$\alpha,\ \beta \in \mathbb{Z}$ and $k \leq |\alpha|$, $l \leq |\beta|$.*

When $\varepsilon = 0$, and $k = 1$ and (or) $l = 1$ these results were established in [**19**, **27**]. All this work starts with the classical paper of H.L. Krall, [**20**].

The proof of Theorem 1.1 is based on a general result of Bakalov, Horozov and Yakimov [**3**] which guarantees that a Darboux transformation preserves the bispectral property under some conditions on the operator $P(n, T)$. We will soon see that its application to the present situation is highly nontrivial and requires, in particular, an intrinsic characterization of a space of difference operators.

We will need some notation from [**3**], see Section 4.1 for more details. Denote by $\mathcal{B}_{\alpha,\beta;\varepsilon}$ the algebra of difference operators $S(n, T)$ with rational coefficients for which there exists a differential operator $G(z, \partial_z)$ (also having rational coefficients) satisfying

$$S(n,T)p_\varepsilon^{\alpha,\beta}(n,z) = G(z,\partial_z)p_\varepsilon^{\alpha,\beta}(n,z). \tag{1.5}$$

All such operators $S(z, \partial_z)$ form a "dual" algebra $\mathcal{B}'_{\alpha,\beta;\varepsilon}$. The map

$$b\colon \mathcal{B}_{\alpha,\beta;\varepsilon} \to \mathcal{B}'_{\alpha,\beta;\varepsilon}, \quad b(R(n,T)) = S(z,\partial_z)$$

is an antiisomorphism of associative algebras. Let $\mathcal{K}_{\alpha,\beta;\varepsilon}$ and $\mathcal{K}'_{\alpha,\beta;\varepsilon}$ be the subalgebras of $\mathcal{B}_{\alpha,\beta;\varepsilon}$ and $\mathcal{B}'_{\alpha,\beta;\varepsilon}$ consisting of rational functions. Bispectrality of $p_\varepsilon^{\alpha,\beta}(n, z)$ is equivalent to $\mathcal{K}_{\alpha,\beta;\varepsilon}$ and $\mathcal{K}'_{\alpha,\beta;\varepsilon}$ being both nontrivial. Finally we arrive at the most important object for our consideration, namely the space

$$\begin{aligned}\mathcal{R}_{\alpha,\beta;\varepsilon} = \{&(\mu(n))^{-1}P_0(n,T) \mid \mu(n) \in \mathcal{K}_{\alpha,\beta;\varepsilon},\ P_0(n,T) \in \mathcal{B}_{\alpha,\beta;\varepsilon},\\ &\text{and the operator } (\mu(n))^{-1}P_0(n,T) \text{ does not have poles at } n \in \mathbb{Z}\}.\end{aligned}$$

According to Theorem 1.2 of [**3**], $\Psi(n, z)$ is an eigenfunction of a differential operator in the variable z, if

$$P(n,T) \in \mathcal{R}_{\alpha,\beta;\varepsilon}.$$

Thus to prove Theorem 1.1 we need a *good* description of the space $\mathcal{K}_{\alpha,\beta;\varepsilon}$ which can be used to check whether the operators $P(n, T)$ from (1.3) belong to $\mathcal{R}_{\alpha,\beta;\varepsilon}$. This is the hardest step in our paper. Let Δ denote the algebra of abstract difference operators with rational coefficients of the form $\sum_{i=p}^{q} b_i(n)T^i$ with rational functions $b_i(n)$ (possibly having poles in $\mathbb{Z}$). The key point of our approach is to consider the involution I of Δ acting on rational functions $h(n)$ by

$$(Ih)(n) := h(-(n + 2\varepsilon + \alpha + \beta + 1))$$

and on the shift operator T by $I(T) := T^{-1}$. In Section 4.2 we prove that $\mathcal{R}_{\alpha,\beta;\varepsilon}$ consists of those difference operators from Δ that do not have poles

in $\mathbb{Z}$ and after conjugation with the function

$$\phi(n) = \frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n}$$

become I-invariant.

The final step of the proof of Theorem 1.1 is to show that the hypothesis guarantees that the kernel of the operator $P(n,T)$ (defining $L(n,T)$) has a basis of functions $f(n)$ for which the ratio $f(n)/\phi(n)$ is an (almost) I-invariant rational function in n. This is done in Section 5.1. Finally Section 5.2 recapitulates the strategy of the proof of Theorem 1.1 for the special case of the set $\mathcal{D}^{(2,0)}_{\alpha,\beta,\varepsilon}$. The reader may find it useful to consult this section while reading the paper.

Let us also note that in the case of Laguerre polynomials the situation simplifies a lot due to a presense of a relation of the type (1.5) with a difference operator $S(n,T)$ of the form $s_1(n)T + s_0(n)$ and a first order differential operator $G(z,\partial_z)$ (see expressions (2.3) and (2.8) in [**12**]). It is not hard to show that as a consequence of this the analog of $\mathcal{R}_{\alpha,\beta;\varepsilon}$ in that case is simply the space of difference operators with rational coefficients.

Comparing with the differential case [**7**], it is natural to conjecture that *all second order regular bispectral difference operators* (*i.e., having support* $[-1,1]$) *are exhausted by the families of operators constructed in* [**10**, **12**, **16**] *and in this article.* The operators in [**16**] are obtained as Darboux transformations from the operators $L_{\alpha,\beta;\varepsilon}(n,T)$ for half integer values of the parameters α,β and are the analogs of "KdV family" in the differential case [**7**].

For later use we introduce some convenient notation. If $f(n)\colon \mathbb{Z}\to\mathbb{C}$ is a nowhere vanishing function and $D_1(n,T)$, $D_2(n,T)$ are difference operators we denote

$$\begin{aligned}\mathrm{Ad}_{f(n)}D_1(n,T) &:= f(n)D_1(n,T)f(n)^{-1},\\ \mathrm{ad}_{D_2(n,T)}D_1(n,T) &:= D_2(n,T)D_1(n,T) - D_1(n,T)D_2(n,T).\end{aligned}$$

2. Biinfinite Jacobi operators.

In the first part of this section we review some properties of the classical Jacobi polynomials $p_n^{\alpha,\beta}(z)$. The second one discusses certain functions $p_\varepsilon^{\alpha,\beta}(n,z)$ which are eigenfunctions of biinfinite analogs $L_{\alpha,\beta;\varepsilon}(n,T)$ of the Jacobi difference operators. The third part describes Darboux maps between the operators $L_{\alpha,\beta;\varepsilon}(n,T)$ with shifted indices α,β.

2.1. Jacobi polynomials. The Jacobi polynomials are the orthogonal polynomials for the measure $(1-z)^\alpha(1+z)^\beta dz$ on the interval $[-1,1]$,

$(\alpha, \beta > -1)$, normalized by

$$p_n^{\alpha,\beta}(1) = 2^{-n}\binom{n+\alpha}{n},\ n \in \mathbb{Z}_{\geq 0}.$$

They are given by

$$p_n^{\alpha,\beta}(z) = \binom{n+\alpha}{n} F(-n, n+\alpha+\beta+1; \alpha+1; (1-z)/2) \tag{2.1}$$

where $F(a, b; c; x)$ denotes the Gauss' hypergeometric function. The reader can consult [**21**, pp. 209–217] for other explicit formulas and a list of major relations for $p_n^{\alpha,\beta}(z)$. Let

$$p^{\alpha,\beta}(n, z) = \begin{cases} p_n^{\alpha,\beta}(z), & \text{for } n \in \mathbb{Z}_{\geq 0} \\ 0, & \text{for } n \in \mathbb{Z}_{<0}. \end{cases} \tag{2.2}$$

Now $p^{\alpha,\beta}(n, z)$ are functions of a discrete parameter n and a continuous parameter z. They satisfy a three term recursion relation

$$L_{\alpha,\beta}(n, T)p^{\alpha,\beta}(n, z) = zp^{\alpha,\beta}(n, z) \tag{2.3}$$

where $L_{\alpha,\beta}(n, T)$ are the difference operators

$$\begin{aligned} L_{\alpha,\beta}(n, T) = {} & \frac{2(n+1)(n+\alpha+\beta+1)}{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}T \\ & + \frac{\beta^2-\alpha^2}{(2n+\alpha+\beta)(2n+\alpha+\beta+2)} \\ & + \frac{2(n+\alpha)(n+\beta)}{(2n+\alpha+\beta)(2n+\alpha+\beta+1)}T^{-1} \end{aligned} \tag{2.4}$$

called *Jacobi operators.* In addition $p^{\alpha,\beta}(n, z)$ are eigenfunctions of the differential operators $B_{\alpha,\beta}(z, \partial_z)$ given by

$$B_{\alpha,\beta}(z, \partial_z) = (z^2-1)\partial_z^2 + (\alpha - \beta + (\alpha+\beta+2)z)\partial_z, \tag{2.5}$$

i.e.,

$$B_{\alpha,\beta}(z, \partial_z)p^{\alpha,\beta}(n, z) = \lambda(n)p^{\alpha,\beta}(n, z) \tag{2.6}$$

for

$$\lambda(n) = n(n+\alpha+\beta+1). \tag{2.7}$$

In view of (2.3) and (2.6), $p^{\alpha,\beta}(n, z)$ are discrete–continuous bispectral functions and $L_{\alpha,\beta}(n, T)$, $B(z, \partial_z)$ bispectral difference (differential) operators.

2.2. The functions $p_\varepsilon^{\alpha,\beta}(n,z)$. In a study of the relation between the so called "associated Jacobi polynomials" and the discrete–continuous bispectral problem Grünbaum and Haine, see [**10**, **11**, **13**], introduced the functions

(2.8)
$$p_\varepsilon^{\alpha,\beta}(n,z) = \frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n} F(-(n+\varepsilon), n+\varepsilon+\alpha+\beta+1; \alpha+1; (1-z)/2)$$

($n \in \mathbb{Z}, z \in \mathbb{C}, |z|<1$) defined for those $\varepsilon, \alpha, \beta \in \mathbb{C}$ such that $\alpha \notin \mathbb{Z}_{<0}$, and $\varepsilon \notin \mathbb{Z}_{<0}$, $\varepsilon+\alpha \notin \mathbb{Z}_{\geq 0}$. We will see later that the first restriction can be lifted.

The functions $p_\varepsilon^{\alpha,\beta}(n,z)$ are no longer polynomials but satisfy relations, similar to the ones for $p^{\alpha,\beta}(n,z)$. In particular they are eigenfunctions of the following difference operators with support $[-1,1]$

$$\begin{aligned} L_{\alpha,\beta;\varepsilon}(n,T) = {} & \frac{2(n+\varepsilon+1)(n+\varepsilon+\alpha+\beta+1)}{(2n+2\varepsilon+\alpha+\beta+1)(2n+2\varepsilon+\alpha+\beta+2)} T \\ & + \frac{\beta^2-\alpha^2}{(2n+2\varepsilon+\alpha+\beta)(2n+2\varepsilon+\alpha+\beta+2)} + \\ & + \frac{2(n+\varepsilon+\alpha)(n+\varepsilon+\beta)}{(2n+2\varepsilon+\alpha+\beta)(2n+2\varepsilon+\alpha+\beta+1)} T^{-1} \end{aligned} \tag{2.9}$$

and of the differential operators $B_{\alpha,\beta}(z,\partial_z)$, Equation (2.5). The corresponding relations are

$$L_{\alpha,\beta;\varepsilon}(n,T) p_\varepsilon^{\alpha,\beta}(n,z) = z p_\varepsilon^{\alpha,\beta}(n,z), \tag{2.10}$$

$$B_{\alpha,\beta}(z,\partial_z) p_\varepsilon^{\alpha,\beta}(n,z) = \lambda_\varepsilon(n) p_\varepsilon^{\alpha,\beta}(n,z), \tag{2.11}$$

where

$$\lambda_\varepsilon(n) = (n+\varepsilon)(n+\varepsilon+\alpha+\beta+1). \tag{2.12}$$

The difference operators $L_{\alpha,\beta;\varepsilon}(n,T)$ will still be called *Jacobi operators.* Further we will only deal with the case when they are regular, i.e., when their coefficients of T and T^{-1} do not vanish for $n \in \mathbb{Z}$. This amounts to the conditions

$$\varepsilon, \varepsilon+\alpha, \varepsilon+\beta, \varepsilon+\alpha+\beta, 2\varepsilon+\alpha+\beta \notin \mathbb{Z}. \tag{2.13}$$

It may be useful to stress here that these will eventually be the only restrictions on our parameters α, β, ε.

The operators $L_{\alpha,\beta;\varepsilon}(n,T)$ do satisfy certain "transformation properties". For instance the following relations hold

$$L_{-\alpha,-\beta,\varepsilon+\alpha+\beta}(n,T) = L_{\alpha,\beta;\varepsilon}(n,T), \tag{2.14}$$

$$\mathrm{Ad}_{(-1)^n} L_{\beta,\alpha,\varepsilon}(n,T) = \mathrm{Ad}_{(-1)^n} L_{-\beta,-\alpha,\varepsilon+\alpha+\beta}(n,T) = -L_{\alpha,\beta;\varepsilon}(n,T). \tag{2.15}$$

It is tempting to use (2.14) to limit attention to the case $\alpha \geq 0$. However, this would eventually bring an undesirable degree of asymmetry in the treatment of the parameters α and β. For this reason we prefer to introduce the appropriate functions $p_\varepsilon^{\alpha,\beta}(n,z)$ for $\alpha \in \mathbb{Z}_{<0}$ (and $\varepsilon, \varepsilon+\beta, \varepsilon+\alpha, \varepsilon+\alpha+\beta \notin \mathbb{Z}$) by using (2.8) and recalling, see [**21**, p. 38] that for $m \in \mathbb{Z}_{\geq 0}$

$$\begin{aligned} (2.16) \qquad & \lim_{c \to -m} \frac{1}{\Gamma(c)} F(a,b;c;z) \\ & = \frac{(a)_{m+1}(b)_{m+1}}{(m+1)!} z^{m+1} F(a+m+1, b+m+1; m+2; z). \end{aligned}$$

We see below that this leads to the following expression for $p_\varepsilon^{\alpha,\beta}(n,z)$ with $\alpha \in \mathbb{Z}_{<0}$ (as long as (2.13) is satisfied)

$$\begin{aligned} (2.17) \quad & \mathcal{C} \frac{(\varepsilon+\beta+1)_n}{(\varepsilon+\alpha+\beta+1)_n} \\ & \cdot \frac{(1-z)^{-\alpha}}{2^{-\alpha}} F(-(n+\varepsilon+\alpha), n+\varepsilon+\beta+1; -\alpha+1; (1-z)/2) \end{aligned}$$

where the constant $\mathcal{C} = \mathcal{C}(\alpha,\beta,\varepsilon)$ is explicitly given by

$$\mathcal{C} = \mathcal{C}(\alpha,\beta,\varepsilon) = \frac{(-1)^\alpha}{(-\alpha-1)!} \cdot \frac{(-\varepsilon)_{-\alpha}(\varepsilon+\alpha+\beta+1)_{-\alpha}}{(-\alpha)!}.$$

It is easy to check that the assumptions (2.13) imply that $\mathcal{C}(\alpha,\beta,\varepsilon)$ is well-defined and does not vanish.

The expression above can be derived by a continuity argument using (2.8) and (2.16) when α approaches a value in $\mathbb{Z}_{<0}$. To see this it is important to notice that for $\alpha \in \mathbb{Z}_{<0}$ the identities

$$\frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n} = \frac{(-\varepsilon)_{-\alpha}}{(-(n+\varepsilon))_{-\alpha}} \qquad \text{and}$$
$$\frac{(\varepsilon+\beta+1)_n}{(\varepsilon+\alpha+\beta+1)_n} = \frac{(n+\varepsilon+\alpha+\beta+1)_{-\alpha}}{(\varepsilon+\alpha+\beta+1)_{-\alpha}}$$

allow one to rewrite the factor

$$\frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n} \cdot \frac{(-(n+\varepsilon))_{-\alpha}(n+\varepsilon+\alpha+\beta+1)_{-\alpha}}{(-\alpha)!}$$

as

$$(2.18) \qquad \frac{(-\alpha-1)!}{(-1)^\alpha} \cdot \mathcal{C}(\alpha,\beta,\varepsilon) \cdot \frac{(\varepsilon+\beta+1)_n}{(\varepsilon+\alpha+\beta+1)_n}$$

which except for the first constant is the factor in front of (2.17).

Then conditions (2.13) guarantee that $p_\varepsilon^{\alpha,\beta}(n,z)$ is well-defined (see (2.8) and (2.17)) and satisfies (2.10) and (2.11). It was proved in [**10**] that the

space of common solutions of (2.10) and (2.11) in a domain $\Omega \subset \mathbb{C}$, not containing ± 1, is two dimensional. Notice also that (2.13) excludes, in particular, the original operators $L_{\alpha,\beta}(n,z)$ ($\varepsilon = 0$) since their leading coefficient vanishes for $n = -1$.

Finally we explain how (2.10) follows from (2.3). Conjugate the operator $L_{\alpha,\beta}(n,T)$ with $\binom{n+a}{n} = \frac{(\alpha+1)_n}{n!}$. The resulting difference operator has rational coefficients and the eigenfunction $F(-n, n+\alpha+\beta+1; \alpha+1; (1-z)/2)$, cf. (2.1). The operator obtained from it by the formal change $n \mapsto n+\varepsilon$ has the eigenfunction $F(-n-\varepsilon, n+\varepsilon+\alpha+\beta+1; \alpha+1; (1-z)/2)$ and all we need to do is conjugate it with $(\varepsilon+1)_n/(\varepsilon+\alpha+1)_n$. The result is the operator $L_{\alpha,\beta;\varepsilon}(n,T)$ which proves (2.10) in the case $\alpha \notin \mathbb{Z}_{<0}$. The case $\alpha \in \mathbb{Z}_{<0}$ follows from the definition (2.8) using the limit (2.16).

2.3. Darboux maps between Jacobi operators. There are four difference relations connecting the values of the Jacobi polynomials $p^{\alpha,\beta}(n,z)$ with shifted indices:

$$p^{\alpha-1,\beta}(n,z) = \left(\frac{n+\alpha+\beta}{2n+\alpha+\beta} - \frac{n+\beta}{2n+\alpha+\beta}T^{-1}\right) p^{\alpha,\beta}(n,z),$$
$$p^{\alpha,\beta}(n,z) = \frac{1}{z-1}\left(\frac{2(n+1)}{2n+\alpha+\beta+1}T - \frac{2(n+\alpha)}{2n+\alpha+\beta+1}\right) p^{\alpha-1,\beta}(n,z),$$

and

$$p^{\alpha,\beta-1}(n,z) = \left(\frac{n+\alpha+\beta}{2n+\alpha+\beta} + \frac{n+\alpha}{2n+\alpha+\beta}T^{-1}\right) p^{\alpha,\beta}(n,z),$$
$$p^{\alpha,\beta}(n,z) = \frac{1}{z+1}\left(\frac{2(n+1)}{2n+\alpha+\beta+1}T + \frac{2(n+\beta)}{2n+\alpha+\beta+1}\right) p^{\alpha,\beta-1}(n,z),$$

(see for instance, [**21**, Eqs. pp. 209–219]). Similarly to the proof of (2.10) at the end of the previous subsection, one shows the following analogs of these identities for $p_\varepsilon^{\alpha,\beta}(n,z)$

(2.19)
$$p_\varepsilon^{\alpha-1,\beta}(n,z) = D_-^\alpha(n,T)p_\varepsilon^{\alpha,\beta}(n,z),\ p_\varepsilon^{\alpha+1,\beta}(n,z) = \frac{1}{z-1}D_+^\alpha(n,T)p_\varepsilon^{\alpha,\beta}(n,z),$$
(2.20)
$$p_\varepsilon^{\alpha,\beta-1}(n,z) = D_-^\beta(n,T)p_\varepsilon^{\alpha,\beta}(n,z),\ p_\varepsilon^{\alpha,\beta+1}(n,z) = \frac{1}{z+1}D_+^\beta(n,T)p_\varepsilon^{\alpha,\beta}(n,z),$$

where the operators $D^{\alpha}_{\pm}(n,T)$ and $D^{\beta}_{\pm}(n,T)$ are given by

$$D^{\alpha}_{-}(n,T)=\left(\frac{\varepsilon+\alpha}{\alpha}\right)\left(\frac{n+\varepsilon+\alpha+\beta}{2n+2\varepsilon+\alpha+\beta}-\frac{n+\varepsilon+\beta}{2n+2\varepsilon+\alpha+\beta}T^{-1}\right),$$

$$D^{\alpha}_{+}(n,T)=\left(\frac{\alpha+1}{\varepsilon+\alpha+1}\right)\left(\frac{2(n+\varepsilon+1)}{2n+2\varepsilon+\alpha+\beta+2}T-\frac{2(n+\varepsilon+\alpha+1)}{2n+2\varepsilon+\alpha+\beta+2}\right),$$

$$D^{\beta}_{-}(n,T)=\left(\frac{n+\varepsilon+\alpha+\beta}{2n+2\varepsilon+\alpha+\beta}+\frac{n+\varepsilon+\alpha}{2n+2\varepsilon+\alpha+\beta}T^{-1}\right),$$

$$D^{\beta}_{+}(n,T)=\left(\frac{2(n+\varepsilon+1)}{2n+2\varepsilon+\alpha+\beta+2}T+\frac{2(n+\varepsilon+\beta+1)}{2n+2\varepsilon+\alpha+\beta+2}\right).$$

The constant in (2.17) was chosen to make the relations (2.19)-(2.20) hold for all $\alpha\in\mathbb{C}$. We show only the dependence on the index α of the operators $D^{\alpha}_{\pm}(n,T)$ because the index β is unchanged in both sides of Equation (2.19), similarly for the operators $D^{\beta}_{\pm}(n,T)$. Equations (2.19)-(2.20) and (2.10) imply the following factorizations

$$L_{\alpha,\beta;\varepsilon}(n,T)-1=D^{\alpha-1}_{+}(n,T)D^{\alpha}_{-}(n,T)=D^{\alpha+1}_{-}(n,T)D^{\alpha}_{+}(n,T), \tag{2.21}$$

$$L_{\alpha,\beta;\varepsilon}(n,T)+1=D^{\beta-1}_{+}(n,T)D^{\beta}_{-}(n,T)=D^{\beta+1}_{-}(n,T)D^{\beta}_{+}(n,T). \tag{2.22}$$

Hence the operators $L_{\alpha\pm1,\beta;\varepsilon}(n,T)$, $L_{\alpha,\beta\pm1;\varepsilon}(n,T)$ are Darboux transformations from $L_{\alpha,\beta;\varepsilon}(n,T)$ and Equations (2.19) and (2.20) represent the Darboux maps $p^{\alpha,\beta}_{\varepsilon}(n,z)\mapsto p^{\alpha\mp1,\beta}_{\varepsilon}(n,z)$ and $p^{\alpha,\beta}_{\varepsilon}(n,z)\mapsto p^{\alpha,\beta\mp1}_{\varepsilon}(n,z)$.

3. Darboux transformations from Jacobi operators.

The first part of this section contains some general facts about discrete Darboux transformations in the form in which they will be used later (see, for instance, [**26**] for the differential case). The goal of the second part is an explicit description of the kernels of the operators $(L_{\alpha,\beta;\varepsilon}-1)^k(L_{\alpha,\beta;\varepsilon}+1)^l$. Based on it, in the third part we construct Darboux transformations from $L_{\alpha,\beta;\varepsilon}(n,T)$ which are the main objects of study in the rest of the paper. The conditions (2.13) are assumed throughout Sections 3.2-3.3.

3.1. General remarks on Darboux transformations. One says that the difference operator $L(n,T)$ is obtained by a Darboux transformation from the difference operator $L_0(n,T)$ if there exists an operator $P(n,T)$ such that

$$L(n,T)P(n,T)=P(n,T)L_0(n,T). \tag{3.1}$$

Assume that $L_0(n,T)$ has an eigenfunction $\Psi_0(n,z)$, i.e.,

$$L_0(n,T)\Psi_0(n,T)=g_0(z)\Psi_0(n) \tag{3.2}$$

for some function $g_0(z)$. Then

$$\Psi(n,z) := P(n,T)\Psi_0(n,z)$$

is an eigenfunction of $L(n,T)$:

$$L(n,T)\Psi(n,T) = g_0(z)\Psi(n). \tag{3.3}$$

The map $\Psi_0(n,T) \mapsto \Psi(n,T)$ is also called a Darboux transformation.

An important feature of the transformation (3.1) *for a regular difference operator $P(n,T)$ is that the operator $L(n,T)$ has the same support as $L_0(n,T)$. Besides this $L(n,T)$ is regular if and only if $L_0(n,T)$ is regular.*

Given a difference operator $L_0(n,T)$, all transformations of the type (3.1) with a regular difference operator $P(n,T)$ can be described in terms of the kernel of $P(n,T)$.

Proposition 3.1.

(i) *For a regular difference operator $P(n,T)$ there exists an operator $L(n,T)$ for which* (3.1) *holds if and only if*

$$L_0(n,T)(\mathrm{Ker}P(n,T)) \subset \mathrm{Ker}P(n,T). \tag{3.4}$$

The operator $L(n,T)$ satisfying (3.1) *is unique.*

(ii) *Let $P(n,T)$ be a regular difference operator satisfying* (3.4) *and $q(x)$ be the characteristic polynomial of the linear map $L_0(n,T)$ acting in the space* $\mathrm{Ker}P(n,T)$. *Then* $\mathrm{Ker}P(n,T) \subset q(L_0(n,T))$ *and there exists an operator $Q(n,T)$ such that*

$$q(L_0(n,T)) = Q(n,T)P(n,T), \tag{3.5}$$

$$q(L(n,T)) = P(n,T)Q(n,T). \tag{3.6}$$

Note that the kernel of a regular difference operator $P(n,T)$ is finite dimensional. More precisely, if $P(n,T)$ has support $[m_1,m_2]$ for some $m_i \in \mathbb{Z}$, then $\dim \mathrm{Ker}P(n,T) = m_2 - m_1$. For any $j \in \mathbb{Z}$ the map

$$f \mapsto (f(j+1), \ldots, f(j+m_2-m_1)), \text{ for } f : \mathbb{Z} \to \mathbb{C} \tag{3.7}$$

provides an isomorphism between $\mathrm{Ker}P(n,T)$ and $\mathbb{C}^{m_2-m_1}$.

The transformation $Q(n,T)P(n,T) \mapsto P(n,T)Q(n,T)$ is a more traditional version of the Darboux map. Although it is a special case of the transformation $L_0(n,T) \mapsto L(n,T)$ from Equation (3.1) and Proposition 3.1 shows that there always exists a polynomial $q(x)$ for which $q(L_0(n,T)) \mapsto q(L(n,T))$ is a Darboux map in this sense.

Proof of Proposition 3.1. (i) If $P(n,T)$, $L(n,T)$ satisfy (3.1) and $f(n) \in \mathrm{Ker}P(n,T)$ then

$$P(n,T)(L_0(n,T)f(n)) = L(n,T)P(n,T)f(n) = 0$$

which proves (3.4).

In the opposite direction, let us notice that a comparison of the coefficients of the two sides of Equation (3.1) for a fixed value of n gives a finite system for the corresponding coefficients of the unknown operator $L(n,T)$ having the same support as $L_0(n,T)$. One shows that it has a solution using the standard linear algebra fact that for a finite matrix A the system $Au = b$ has a solution if and only if $v^t b = 0$, $\forall v \in \mathrm{Ker}A^t$. In the particular case which we consider the last condition is fulfilled because of (3.4).

The regularity of the difference operator $P(n,T)$ implies the uniqueness of the operator $L(n,T)$ satisfying (3.1). Indeed if there are two such operators $L(n,T)$ and $L'(n,T)$ one can subtract the resulting equalities (3.1). This gives $(L(n,T) - L'(n,T))P(n,T) = 0$ which is a contradiction.

(ii) The relation $\mathrm{Ker}P(n,T) \subset q(L_0(n,T))$ follows from the definition of $q(x)$. Similarly to Part (i), this implies the existence of an operator $Q(n,T)$ satisfying (3.5). Equations (3.1) and (3.5) imply

$$q(L(n,T))P(n,T) = P(n,T)q(L_0(n,T)) = (P(n,T)Q(n,T))P(n,T)$$

and as a consequence of this (3.6). □

A regular difference operator is reconstructed from its kernel by the following lemma.

Lemma 3.2. *Assume that $P(n,T)$ is a regular difference operator with support $[m_1, m_2]$ and leading coefficient* 1. *Let* $\mathrm{Ker}P(n,T) = \mathrm{Span}\{f^{(i)}(n)\}_{i=1}^{m}$ *where $m = m_2 - m_2$. Then the function*

$$\det(n) := \det(f^{(i)}(n-j))_{i,j=1,m_1}^{m,m_2-1}$$

does not vanish for $n \in \mathbb{Z}$ and

$$P(n,T) = \frac{1}{\det(n)} \begin{vmatrix} f^{(1)}(n+m_1) & \cdots & f^{(m)}(n+m_1) & T^{m_1} \\ \cdots & \cdots & \cdots & \cdots \\ f^{(1)}(n+m_2) & \cdots & f^{(m)}(n+m_2) & T^{m_2} \end{vmatrix} \tag{3.8}$$

where the determinant is expanded from left to right (*the shift operator T does not commute with function multiplication*).

Proof. The fact that the map (3.7) is an isomorphism between $\mathrm{Ker}P(n,T)$ and $\mathbb{C}^m$ implies that $\det(n)$ does not vanish for $n \in \mathbb{Z}$. Clearly the functions $f^{(i)}(n)$ belong to the kernel of the operator in the r.h.s. of (3.8). It has leading term 1 and the nonvanishing of $\det(n)$ implies (3.8). □

Remark 3.3. The composition of two Darboux transformations $L_0(n,T) \mapsto L_1(n,T)$ and $L_1(n,T) \mapsto L_2(n,T)$ of the type (3.1) is Darboux transformation $L_0(n,T) \mapsto L_2(n,T)$ of the same type. Indeed if

$$L_i(n,T)P_i(n,T) = P_i(n,T)L_{i-1}(n,T),\ i = 1,2,$$

then

$$L_2(n,T)P_2(n,T)P_1(n,T) = P_2(n,T)P_1(n,T)L_0(n,T).$$

3.2. Description of $\mathrm{Ker}(L_{\alpha,\beta;\varepsilon}-1)^k(L_{\alpha,\beta;\varepsilon}+1)^l$**.** The main idea is to first find some functions $\varphi(n,z)$ (depending on α,β, and ε) such that

$$L_{\alpha,\beta;\varepsilon}(n,T)\varphi(n,z)=z\varphi(n,z) \tag{3.9}$$

and then to consider the derivatives

$$\varphi_{\pm}^{(i)}(n)=\frac{1}{i!}\partial_z^i\varphi(n,z)\Big|_{z=\pm 1},\ i\in\mathbb{Z}_{\geq 0}.$$

They satisfy

$$(L_{\alpha,\beta;\varepsilon}(n,T)\mp 1)\varphi_{\pm}^{(i)}(n)=\varphi_{\pm}^{(i-1)}(n),\ \forall\, i\in\mathbb{Z}_{\geq 0} \tag{3.10}$$

with $\varphi_{\pm}^{(-1)}(n)=0$. As a consequence of this

$$(L_{\alpha,\beta;\varepsilon}(n,T)\mp 1)^i\varphi_{\pm}^{(j)}(n)=0,\ \forall\, i\in\mathbb{Z}_{>0},\ j=0,\dots,i-1.$$

Before stating the results from this subsection we recall a relation for the hypergeometric function that is a consequence of Gauss' relations between contiguous hypergeometric functions. Denote $F=F(a,b;c;(1-z)/2)$, $TF=F(a-1,b+1;c;(1-z)/2)$, and $T^{-1}F=F(a+1,b-1;c;(1-z)/2)$. Then for $c\notin\mathbb{Z}_{\leq 0}$

$$\frac{2(c-a)b}{(b-a)(b-a+1)}TF+\frac{2(a+b-1)(-2c+a+b+1)}{(b-a-1)(b-a+1)}F + \frac{2a(c-b)}{(b-a)(b-a-1)}T^{-1}F=zF. \tag{3.11}$$

This can also be checked directly using the standard expansion of $F(a,b;c,x)$ for $|x|<1$, $c\notin\mathbb{Z}_{\leq 0}$

$$F(a,b;c;x)=\sum_{j=0}^{\infty}\frac{(a)_j(b)_j}{j!(c)_j}x^j. \tag{3.12}$$

Lemma 3.4. *The four functions*

$$\varphi_+(n,z)=\frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n} \cdot F(-(n+\varepsilon),n+\varepsilon+\alpha+\beta+1;\alpha+1;(1-z)/2), \tag{3.13}$$

$$\psi_+(n,z)=\frac{(\varepsilon+\beta+1)_n}{(\varepsilon+\alpha+\beta+1)_n} \cdot F(-(n+\varepsilon+\alpha+\beta),n+\varepsilon+1;-\alpha+1;(1-z)/2), \tag{3.14}$$

$$\varphi_-(n,z) = \frac{(-1)^n(\varepsilon+\beta+1)_n}{(\varepsilon+1)_n} \cdot F(-(n+\varepsilon), n+\varepsilon+\alpha+\beta+1; \beta+1; (1+z)/2), \tag{3.15}$$

$$\psi_-(n,z) = \frac{(-1)^n(\varepsilon+\alpha+1)_n}{(\varepsilon+\alpha+\beta+1)_n} \cdot F(-(n+\varepsilon+\alpha+\beta), n+\varepsilon+1; -\beta+1; (1+z)/2) \tag{3.16}$$

satisfy

$$(L_{\alpha,\beta;\varepsilon}(n,T) - z)\varphi_\pm(n,z) = (L_{\alpha,\beta;\varepsilon}(n,T) - z)\psi_\pm(n,z) = 0, \tag{3.17}$$

provided that $\alpha \notin \mathbb{Z}_{<0}\,(\mathbb{Z}_{>0})$ *for* $\varphi_+(n,z)$ $(\psi_+(n,z))$ *and* $\beta \notin \mathbb{Z}_{<0}\,(\mathbb{Z}_{>0})$ *for* $\varphi_-(n,z)$ $(\psi_-(n,z))$.

Note that the assumptions (2.13) guarantee that the denominators of the first factors of $\varphi_\pm(n,z)$ and $\psi_\pm(n,z)$ do not vanish.

Proof. The relation (3.17) for $\varphi_+(n,z)$ holds because $\varphi_+(n,z) = p_\varepsilon^{\alpha,\beta}(n,z)$. To check the one for $\psi_+(n,z)$, we conjugate $L_{\alpha,\beta;\varepsilon}(n,T)$ by $(\varepsilon+\beta+1)_n/(\varepsilon+\alpha+\beta+1)_n$ (the factor in front of the r.h.s. of (3.14)).

The result is

$$\begin{aligned}
&\mathrm{Ad}_{(\varepsilon+\beta+1)_n/(\varepsilon+\alpha+\beta+1)_n} L_{\alpha,\beta;\varepsilon}(n,T) \\
&= \frac{2(n+\varepsilon+1)(n+\varepsilon+\beta+1)}{(2n+2\varepsilon+\alpha+\beta+1)(2n+2\varepsilon+\alpha+\beta+2)} T \\
&\quad + \frac{\beta^2-\alpha^2}{(2n+2\varepsilon+\alpha+\beta)(2n+2\varepsilon+\alpha+\beta+2)} \\
&\quad + \frac{2(n+\varepsilon+\alpha)(n+\varepsilon+\alpha+\beta)}{(2n+2\varepsilon+\alpha+\beta)(2n+2\varepsilon+\alpha+\beta+1)} T^{-1}.
\end{aligned}$$

This is the difference operator from the l.h.s. of (3.11) with $a = -(n+\varepsilon+\alpha+\beta)$, $b = n+\varepsilon+1$, and $c = -\alpha+1$ which gives the proof of (3.17) for $\psi_+(n,z)$. The cases of $\varphi_-(n,z)$ and $\psi_-(n,z)$ are handled in a similar fashion. □

Next we consider the derivatives of $\varphi_+(n,z)$, $\psi_+(n,z)$ at $z=1$ and of $\varphi_-(n,z)$, $\psi_-(n,z)$ at $z=-1$:

$$\varphi_\pm^{(i)}(n) := \frac{1}{i!}\partial_z^i \varphi_\pm(n,z)\Big|_{z=\pm 1},$$
$$\psi_\pm^{(i)}(n) := \frac{1}{i!}\partial_z^i \psi_\pm(n,z)\Big|_{z=\pm 1},$$

$i \in \mathbb{Z}_{\geq 0}$ (with the restrictions on α and β made at the end of Lemma 3.4).

Using the expansion (3.12) of the hypergeometric function, we obtain the following explicit formulas for $\varphi_{\pm}^{(i)}(n)$ and $\psi_{\pm}^{(i)}(n)$

$$\varphi_{+}^{(i)}(n) = \frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n} \cdot \frac{(-(n+\varepsilon))_i (n+\varepsilon+\alpha+\beta+1)_i}{(-2)^i i! (\alpha+1)_i} \tag{3.18}$$

$$\psi_{+}^{(i)}(n) = \frac{(\varepsilon+\beta+1)_n}{(\varepsilon+\alpha+\beta+1)_n} \cdot \frac{(-(n+\varepsilon+\alpha+\beta))_i (n+\varepsilon+1)_i}{(-2)^i i! (-\alpha+1)_i} \tag{3.19}$$

$$\varphi_{-}^{(i)}(n) = \frac{(-1)^n(\varepsilon+\beta+1)_n}{(\varepsilon+1)_n} \cdot \frac{(-(n+\varepsilon))_i (n+\varepsilon+\alpha+\beta+1)_i}{2^i i! (\beta+1)_i} \tag{3.20}$$

$$\psi_{-}^{(i)}(n) = \frac{(-1)^n(\varepsilon+\alpha+1)_n}{(\varepsilon+\alpha+\beta+1)_n} \cdot \frac{(-(n+\varepsilon+\alpha+\beta))_i (n+\varepsilon+1)_i}{2^i i! (-\beta+1)_i} . \tag{3.21}$$

We define $\varphi_{+}^{(i)}(n)$ ($\psi_{+}^{(i)}(n)$) for $\alpha \in \mathbb{Z}_{<0}$ ($\alpha \in \mathbb{Z}_{>0}$), $i < |\alpha|$ by (3.18), (3.19) and $\varphi_{-}^{(i)}(n)$ ($\psi_{-}^{(i)}(n)$) for $\beta \in \mathbb{Z}_{<0}$ ($\beta \in \mathbb{Z}_{>0}$), $i < |\beta|$ by (3.20), (3.21). (Note that these cases were excluded in Lemma 3.4.)

Theorem 3.5. *Assuming* (2.13) *the following relations*

$$(L_{\alpha,\beta;\varepsilon}(n,T) - 1)\varphi_{+}^{(i)}(n) = \varphi_{+}^{(i-1)}(n), \tag{3.22}$$

$$(L_{\alpha,\beta;\varepsilon}(n,T) - 1)\psi_{+}^{(i)}(n) = \psi_{+}^{(i-1)}(n), \tag{3.23}$$

hold for all $i \in \mathbb{Z}_{\geq 0}$ *if* $\alpha \notin \mathbb{Z}$ *and for* $i = 0, \dots, |\alpha| - 1$ *if* $\alpha \in \mathbb{Z}$. *Similarly one has*

$$(L_{\alpha,\beta;\varepsilon}(n,T) + 1)\varphi_{-}^{(i)}(n) = \varphi_{-}^{(i-1)}(n), \tag{3.24}$$

$$(L_{\alpha,\beta;\varepsilon}(n,T) + 1)\psi_{-}^{(i)}(n) = \psi_{-}^{(i-1)}(n), \tag{3.25}$$

for all $i \in \mathbb{Z}_{\geq 0}$ *if* $\beta \notin \mathbb{Z}$ *and for* $i = 0, \dots, |\beta| - 1$ *if* $\beta \in \mathbb{Z}$. (*We set* $\varphi_{\pm}^{(-1)}(n) = \psi_{\pm}^{(-1)}(n) = 0$.)

The kernels of $(L_{\alpha,\beta;\varepsilon}(n,T) - 1)^k$ *and* $(L_{\alpha,\beta;\varepsilon}(n,T) + 1)^l$ *are given by*

$$\mathrm{Ker}(L_{\alpha,\beta;\varepsilon}(n,T) - 1)^k = \mathrm{Span}\{\varphi_{+}^{(i)}(n), \psi_{+}^{(i)}(n)\}_{i=0}^{k-1}, \tag{3.26}$$

$$\mathrm{Ker}(L_{\alpha,\beta;\varepsilon}(n,T) + 1)^l = \mathrm{Span}\{\varphi_{-}^{(i)}(n), \psi_{-}^{(i)}(n)\}_{i=0}^{l-1}, \tag{3.27}$$

for $k \leq |\alpha|$ *if* $\alpha \in \mathbb{Z}$, *for* $l \leq |\beta|$ *if* $\beta \in \mathbb{Z}$, *and for all* $k, l \geq 0$ *if* $\alpha, \beta \notin \mathbb{Z}$.

Proof. In the case $\alpha \notin \mathbb{Z}$, the functions $\varphi_{+}(n,z)$ and $\psi_{+}(n,z)$ are well-defined. From the remark in the beginning of this subsection it follows that (3.17) and the definitions of $\varphi_{+}^{(j)}(n)$, $\psi_{+}^{(j)}(n)$ imply (3.22), (3.23). The case $\alpha \in \mathbb{Z}$, $i < |\alpha|$ follows by continuity on α.

The inclusion $\supset$ in (3.26), (3.27) clearly follows from (3.22)-(3.25). Because $(L_{\alpha,\beta;\varepsilon}(n,T)-1)^k$ is a regular difference operator with support $[-k,k]$, to prove (3.26) it suffices to show that the functions $\varphi_{+}^{(i)}(n)$, $\psi_{+}^{(i)}(n)$, $i = 0, \dots, k-1$ are linearly independent.

Assume that

$$\sum_{i=0}^{k_0}\left(a_i\varphi_+^{(i)}(n)+b_i\psi_+^{(i)}(n)\right)=0,\quad \forall n\in\mathbb{Z}$$

for some complex numbers $a_0,\dots,a_{k_0}$, $b_0,\dots,b_{k_0}$, such that $a_{k_0}\neq 0$ or $b_{k_0}\neq 0$ ($k_0\leq k-1$). Applying $(L_{\alpha,\beta;\varepsilon}(n,T))^{k_0-1}$ to this equality and using (3.22), (3.23), we get

$$a_{k_0}\varphi_+^{(0)}(n)+b_{k_0}\psi_+^{(0)}(n)=0,\quad \forall n\in\mathbb{Z},$$

i.e.,

$$a_{k_0}\frac{(\varepsilon+\alpha+1)_n}{(\varepsilon+1)_n}=-b_{k_0}\frac{(\varepsilon+\beta+1)_n}{(\varepsilon+\alpha+\beta+1)_n},\quad \forall n\in\mathbb{Z}. \tag{3.28}$$

For $n=0$ this gives $a_{k_0}=-b_{k_0}$ ($\neq 0$). Dividing the two sides of Equation (3.28) for two consecutive values of n, we get

$$(\varepsilon+\alpha+n)(\varepsilon+\alpha+\beta+n)=(\varepsilon+n)(\varepsilon+\beta+n),\quad \forall n\in\mathbb{Z}.$$

This gives $\alpha=0$ which is a contradiction. Equation (3.27) is proved analogously. □

Remark 3.6. It is clear that

$$\operatorname{Ker}(L_{\alpha,\beta;\varepsilon}-1)^k\cap\operatorname{Ker}(L_{\alpha,\beta;\varepsilon}+1)^l=\emptyset.$$

Therefore

$$\operatorname{Ker}(L_{\alpha,\beta;\varepsilon}-1)^k(L_{\alpha,\beta;\varepsilon}+1)^l=\operatorname{Ker}(L_{\alpha,\beta;\varepsilon}-1)^k\oplus\operatorname{Ker}(L_{\alpha,\beta;\varepsilon}+1)^l$$

and Theorem 3.5 describes the kernel of the operator $(L_{\alpha,\beta;\varepsilon}-1)^k(L_{\alpha,\beta;\varepsilon}+1)^l$ in the cases specified there.

3.3. The sets $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ of Darboux transformations from $L_{\alpha,\beta;\varepsilon}(n,z)$. Let us fix two nonnegative integers k and l and choose $2(k+l)$ complex numbers

$$A_i,B_i,\ i=0,\dots,k-1,$$
$$C_j,D_j,\ j=0,\dots,l-1.$$

If $k>0$ ($l>0$) we will assume $\alpha\neq -k+1,\dots,k-1$ ($\beta\neq -l+1,\dots,l-1$). Set

$$f^{(i)}(n)=\begin{cases}\sum_{r=0}^{i}\left(A_r\varphi_+^{(i-r)}(n)+B_r\psi_+^{(i-r)}(n)\right), & \text{for } i=0,\dots,k-1\\ \sum_{r=0}^{i-k}\left(C_r\varphi_-^{(i-k-r)}(n)+D_r\psi_-^{(i-k-r)}(n)\right), & \text{for } i=k,\dots,k+l-1\end{cases}.$$

The values of the parameters $A, B, C, D \in \mathbb{C}$ for which

$$\det(n) = \det(f^{(i)}(n+j))_{i,j=0,-k-l}^{k+l-1,-1} \neq 0, \ \forall n \in \mathbb{Z}, \tag{3.29}$$

will be called *admissible.* For such values we define the operator

(3.30)

$$P(n,T) = \frac{1}{\det(n)} \begin{vmatrix} f^{(0)}(n-k-l) & \cdots & f^{(k+l-1)}(n-k-l) & T^{-(k+l)} \\ \cdots & \cdots & \cdots & \cdots \\ f^{(0)}(n) & \cdots & f^{(k+l-1)}(n) & 1 \end{vmatrix}.$$

By expanding (3.30) along the last column one sees that the term of $T^{-(k+l)}$ is given by

$$\frac{\det(n+1)}{\det(n)} \neq 0$$

hence $P(n,T)$ is a regular difference operator. As a consequence of properties (3.22)–(3.25) we obtain

$$(L_{\alpha,\beta;\varepsilon}(n,T)-1)f^{(0)}(n) = 0, \quad (L_{\alpha,\beta;\varepsilon}(n,T)-1)f^{(i)} = f^{(i-1)}(n) \tag{3.31}$$

for $i = 1, \ldots, k-1$ and

$$(L_{\alpha,\beta;\varepsilon}(n,T)+1)f^{(k)}(n) = 0, \quad (L_{\alpha,\beta;\varepsilon}(n,T)+1)f^{(j)} = f^{(j-1)}(n) \tag{3.32}$$

for $j = k+1, \ldots, k+l-1$. Thus $\mathrm{Ker}P(n,T) = \mathrm{Span}\{f^{(i)}(n)\}_{i=0}^{k+l-1}$ is preserved by $L_{\alpha,\beta;\varepsilon}(n,T)$ and according to Proposition 3.1 there exists a difference operator $L(n,T)$ with support $[-1,1]$ such that

$$L(n,T)P(n,T) = P(n,T)L_{\alpha,\beta;\varepsilon}(n,T). \tag{3.33}$$

The set of all difference operators $L(n,T)$ for admissible values of the parameters A, B, C, D will be denoted by

$$\mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}.$$

All operators $L(n,T) \in \mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}$ are Darboux transformations from $L_{\alpha,\beta;\varepsilon}(n,T)$ and k, l refer to the multiplicity of the eigenvalues 1 and -1 of $L_{\alpha,\beta;\varepsilon}(n,T)$ in $\mathrm{Ker}P(n,T)$, see Equations (3.31) and (3.32). (Recall from part (i) of Proposition 3.1 that $L_{\alpha,\beta;\varepsilon}(n,T)$ preserves $\mathrm{Ker}P(n,T)$.) Every $L(n,T) \in \mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}$ is a *regular* difference operator with eigenfunction

$$\Psi(n,z) = P(n,T)p_{\varepsilon}^{\alpha,\beta}(n,z), \tag{3.34}$$

more precisely:

$$L(n,T)\Psi(n,z) = z\Psi(n,z). \tag{3.35}$$

The admissibility condition (3.29) holds for almost all values of $A, B, C, D \in \mathbb{Z}$. The complement of the corresponding set in $\mathbb{C}^{2(k+l)}$ consists of the zeros of countably many polynomials, obtained from $\det(n)$ for fixed $n \in$

$\mathbb{Z}$ (recall (3.29)). The latter do not vanish identically due to the linear independence of the set of functions $\{\varphi_{\pm}^{(i)}(n)\}_{i=0}^{k-1} \cup \{\psi_{\pm}^{(j)}(n)\}_{j=0}^{l-1}$ (see the proof of Theorem 3.5) and the regularity of $L_{\alpha,\beta;\varepsilon}(n,T)$.

There are in fact $k+l$ free parameters in the definition of an element $L(n,T) \in \mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}$ since the operator $P(n,T)$ (see (3.30)) only depends on the choice of the space $\mathrm{Span}\{f^{(i)}(n)\}_{i=0}^{k+l-1} (= \mathrm{Ker}P(n,T))$, and not on the choice of the individual functions $f^{(i)}(n)$. Using again the linear independence of $\{\varphi_{\pm}^{(i)}(n)\}_{i=0}^{k-1} \cup \{\psi_{\pm}^{(j)}(n)\}_{j=0}^{l-1}$, the choice of span is equivalent to a choice of flags

$$V_0 \subset V_1 \subset \ldots \subset V_{k-1} \text{ and } W_0 \subset W_1 \subset \ldots \subset W_{l-1}$$

where $V_i = \mathrm{Span}\{f^{(r)}(n)\}_{r=0}^{i}$ and $W_j = \mathrm{Span}\{f^{(r)}(n)\}_{r=k}^{k+j}$, cf. [**12**].

The relations (2.14) and (2.15) for $L_{\alpha,\beta;\varepsilon}(n,T)$ imply similar relations for the sets $\mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}$:

$$\mathcal{D}_{-\alpha,-\beta,\varepsilon+\alpha+\beta}^{(k,l)} = \mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}, \tag{3.36}$$

$$\mathrm{Ad}_{(-1)^n} \mathcal{D}_{\beta,\alpha,\varepsilon}^{(l,k)} = \mathrm{Ad}_{(-1)^n} \mathcal{D}_{-\beta,-\alpha,\varepsilon+\alpha+\beta}^{(l,k)} = -\mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}. \tag{3.37}$$

Here, in addition to (2.14) and (2.15), we use that the change of parameters $\alpha \to -\alpha$, $\beta \to -\beta$, $\varepsilon \to \varepsilon+\alpha+\beta$ exchanges $\varphi_{+}^{(i)}(n)$ with $\psi_{+}^{(i)}(n)$ and $\varphi_{-}^{(i)}(n)$ with $\psi_{-}^{(i)}(n)$. Analogously the change of parameters $\alpha \to \beta$, $\beta \to \alpha$, $\varepsilon \to \varepsilon$ exchanges $\varphi_{+}^{(i)}(n)$ with $(-1)^n \varphi_{-}^{(i)}(n)$ and $\psi_{+}^{(i)}(n)$ with $(-1)^n \psi_{-}^{(i)}(n)$.

The Darboux maps between Jacobi functions (operators) represented by the first identities in (2.19), (2.20) and Remark 3.3 imply the following inclusion relations

$$\mathrm{Ad}_{\frac{2n+2\varepsilon+\alpha+\beta}{n+\varepsilon+\alpha+\beta}} \mathcal{D}_{\alpha-1,\beta;\varepsilon}^{(k-1,l)} \subset \mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}, \tag{3.38}$$

$$\mathrm{Ad}_{\frac{2n+2\varepsilon+\alpha+\beta}{n+\varepsilon+\alpha+\beta}} \mathcal{D}_{\alpha,\beta-1;\varepsilon}^{(k,l-1)} \subset \mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}. \tag{3.39}$$

The function $(n+\varepsilon+\alpha+\beta)/(2n+2\varepsilon+\alpha+\beta)$ is the leading coefficient of the the operators $D_{-}^{\alpha}(n,T)$ and $D_{-}^{\beta}(n,T)$, see Section 2.3. Recall that the operator $P(n,T)$ is normalized to have leading coefficient 1.

Remark 3.7. Note that (3.31), (3.32) imply that for the operator $P(n,T)$ (3.30) defining an element $L(n,T)$ in $\mathcal{D}_{\alpha,\beta;\varepsilon}^{(k,l)}$ the endomorphism $L_{\alpha,\beta;\varepsilon}(n,T)$ on $\mathrm{Ker}P(n,T)$ has two Jordan blocks with eigenvalues 1 and -1 and lengths k and l, respectively. Insisting on multiple blocks with equal eigenvalues does not produce larger sets of transformations since the operator $L_{\alpha,\beta;\varepsilon}(n,T)$ has a two dimensional kernel. Allowing $k > |\alpha|$ or $l > \beta$ in the cases $\alpha \in \mathbb{Z}$ or $\beta \in \mathbb{Z}$ causes the operators $P(n,T)$ and $L(n,T)$ to have nonrational coefficients which does lead to bispectrality of $L(n,T)$ as was noted in the introduction.

At the end of this subsection we compute explicitly the coefficients of the operators $L(n,T)$ in $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$. Set

$$L(n,T) = a(n)T + b(n) + c(n)T^{-1} \tag{3.40}$$

for some functions $a(n)$, $b(n)$, and $c(n)$ (the dependence on A, B, C, D will not be shown). For convenience we denote the coefficients of the operator $L_{\alpha,\beta;\varepsilon}(n,T)$ by $a_0(n)$, $b_0(n)$, and $c_0(n)$:

$$L_{\alpha,\beta;\varepsilon}(n,T) = a_0(n)T + b_0(n) + c_0(n)T^{-1} \tag{3.41}$$

(cf. Equation (2.9) for their values). Set also

$$\det{}_{-r}(n) := \det(f^{(i)}(n+j))_{\substack{i=0,\dots,k+l-1 \\ j=-k-l,\dots,-\hat{r},\dots,0}} \quad \text{for } r = 0,\dots,k+l. \tag{3.42}$$

Note that

$$\det{}_0(n) = \det(n) \quad \text{and} \quad \det{}_{k+l}(n) = \det{}_0(n+1) = \det(n+1). \tag{3.43}$$

Expanding the determinant (3.30) defining $P(n,T)$ along the last column gives

$$P(n,T) = \sum_{r=0}^{k+l} (-1)^r \frac{\det_{-r}(n)}{\det(n)} T^{-r}. \tag{3.44}$$

Proposition 3.8. *The coefficients $a(n)$, $b(n)$, and $c(n)$ of an operator $L(n,T) \in \mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ are expressed in terms of the coefficients $a_0(n)$, $b_0(n)$, and $c_0(n)$ of $L_{\alpha,\beta;\varepsilon}(n,T)$ and the functions $f^{(i)}(n)$ (see (3.42)) by the following formulas*

$$a(n) = a_0(n), \tag{3.45}$$

$$b(n) = b_0(n) + a_0(n)\frac{\det_{-1}(n+1)}{\det(n+1)} - a_0(n-1)\frac{\det_{-1}(n)}{\det(n)}, \tag{3.46}$$

$$c(n) = c_0(n-k-l)\frac{\det(n-1)\det(n+1)}{(\det(n))^2}. \tag{3.47}$$

Proof. We compare the coefficients of T and 1 in (3.33) and use the expression (3.44) for the operator $P(n,T)$. This gives the formulas

$$a(n) = a_0(n),$$

$$b(n) - a(n)\frac{\det_{-1}(n+1)}{\det(n+1)} = b_0(n) - a_0(n-1)\frac{\det_{-1}(n)}{\det(n)},$$

which are equivalent to (3.45) and (3.46).

Similarly comparing the coefficients of T^{-k-l-1} in (3.33) gives

$$c(n)\frac{\det_{-(k+l)}(n-1)}{\det(n-1)} = c_0(n-k-l)\frac{\det_{-(k+l)}(n)}{\det(n)}$$

which implies (3.47), taking into account (3.43). □

4. Bispectral Darboux transformation and an involution.

This section is a preparation for the next one where we show that the difference operators from $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ are bispectral under some natural conditions on α and β. Our proof is based on a result of [**3**] on Darboux transformations that preserve the bispectral property. Its application to the situation under consideration is nontrivial and requires an intrinsic characterization of a certain space of difference operators. This is done in terms of an involution of the algebra of difference operators with rational coefficients.

4.1. A theorem on bispectral Darboux transformations. For a fixed choice of the parameters $\alpha, \beta, \varepsilon$ we define $\mathcal{B}_{\alpha,\beta;\varepsilon}$ as the algebra of difference operators $S(n,T)$ with *rational* coefficients for which there exists a differential operator $G(z,\partial_z)$ (also with rational coefficients) such that

$$S(n,T)p^{\alpha,\beta}_{\varepsilon}(n,z) = G(z,\partial_z)p^{\alpha,\beta}_{\varepsilon}(n,z). \tag{4.1}$$

The set of all such operators $G(z,\partial_z)$ is an algebra which will be denoted by $\mathcal{B}'_{\alpha,\beta;\varepsilon}$. It is clear that

$$b\left(S(n,T)\right) := G(z,\partial_z) \tag{4.2}$$

correctly defines a map

$$b : \mathcal{B}_{\alpha,\beta;\varepsilon} \to \mathcal{B}'_{\alpha,\beta;\varepsilon} \tag{4.3}$$

which is an antiisomorphism of algebras. In this setting Equations (2.10) and (2.11) mean that $\lambda_\varepsilon(n), L_{\alpha,\beta;\varepsilon}(n,T) \in \mathcal{B}_{\alpha,\beta;\varepsilon}$, $z, B_{\alpha,\beta}(z,\partial_z) \in \mathcal{B}'_{\alpha,\beta;\varepsilon}$, and

$$b(\lambda_\varepsilon(n)) = B_{\alpha,\beta}(z,\partial_z), \tag{4.4}$$

$$b(L_{\alpha,\beta;\varepsilon}(n,T)) = z. \tag{4.5}$$

The triple $(\mathcal{B}_{\alpha,\beta;\varepsilon}, \mathcal{B}'_{\alpha,\beta;\varepsilon}, b)$ is an example of a bispectral triple in the sense of [**3**]. Denote

$$\mathcal{K}_{\alpha,\beta;\varepsilon} = \mathcal{B}_{\alpha,\beta;\varepsilon} \cap \mathbb{C}(n), \tag{4.6}$$

$$\mathcal{K}'_{\alpha,\beta;\varepsilon} = \mathcal{B}'_{\alpha,\beta;\varepsilon} \cap \mathbb{C}(z), \tag{4.7}$$

where $\mathbb{C}(n)$ and $\mathbb{C}(z)$ stand for the algebras of rational functions in the variables n and z, respectively. Let

$$\mathcal{A}_{\alpha,\beta;\varepsilon} = b^{-1}\left(\mathcal{K}'_{\alpha,\beta;\varepsilon}\right), \tag{4.8}$$

$$\mathcal{A}'_{\alpha,\beta;\varepsilon} = b\left(\mathcal{K}_{\alpha,\beta;\varepsilon}\right). \tag{4.9}$$

It is obvious that

$$\mathcal{K}'_{\alpha,\beta;\varepsilon} = \mathbb{C}[z], \tag{4.10}$$

$$\mathcal{A}'_{\alpha,\beta;\varepsilon} = \mathbb{C}[B_{\alpha,\beta}(z,\partial_z)], \tag{4.11}$$

and

$$\mathcal{K}_{\alpha,\beta;\varepsilon} \supset \mathbb{C}[\lambda(n)], \tag{4.12}$$

$$\mathcal{A}_{\alpha,\beta;\varepsilon} \supset \mathbb{C}[L_{\alpha,\beta;\varepsilon}(n,T)]. \tag{4.13}$$

Later in Remark 4.3 we will show that the inclusions in (4.12) and (4.13) can be strengthen to give two equalities.

As was noted in Section 3.1, if a difference operator $q(L_{\alpha,\beta;\varepsilon}(n,T)) \in \mathcal{A}_{\alpha,\beta;\varepsilon}$ ($q(x) \in \mathbb{C}[x]$) is factorized as a product of two operators $Q(n,T)$ and $P(n,T)$

$$q(L_{\alpha,\beta;\varepsilon}(n,T)) = Q(n,T)P(n,T),$$

then the function

$$\Psi(n,z) = P(n,T)p_{\varepsilon}^{\alpha,\beta}(n,z)$$

is an eigenfunction of the difference operator $P(n,T)Q(n,T)$:

$$P(n,T)Q(n,T)\Psi(n,z) = q(z)\Psi(n,z).$$

We will give a version of Theorem 1.2 from [**3**] which provides general sufficient conditions on the operators $P(n,T)$ and $Q(n,T)$ under which $\Psi(n,z)$ is also an eigenfunction of a differential operator in the variable z. (The original result of [**3**] deals with "bispectral" Darboux transformations in an arbitrary associative algebra but in the form to be used, needs an additional refinement.)

Theorem 4.1. *Assume that the operator $q(L_{\alpha,\beta;\varepsilon}(n,T)) \in \mathcal{A}_{\alpha,\beta;\varepsilon}$ is factorized as*

$$q(L_{\alpha,\beta;\varepsilon}(n,T)) = (Q_0(n,T)\nu(n)^{-1})(\mu(n)^{-1}P_0(n,T)) \tag{4.14}$$

for some difference operators $P_0(n,T), Q_0(n,T) \in \mathcal{B}_{\alpha,\beta;\varepsilon}$ and rational functions $\mu(n), \nu(n) \in \mathcal{K}_{\alpha,\beta;\varepsilon}$, such that the coefficients of the operators $\mu(n)^{-1}P_0(n,T)$, $Q_0(n,T)\nu(n)^{-1}$ are correctly defined for $n \in \mathbb{Z}$. Then the function

$$\Psi(n,z) = (\mu^{-1}(n)P_0(n,T))p_{\varepsilon}^{\alpha,\beta}(n,z) \tag{4.15}$$

satisfies the relations

$$(\mu(n)^{-1}P_0(n,T))(Q_0(n,T)\nu(n)^{-1})\Psi(n,z) = q(z)\Psi(n,z), \tag{4.16}$$

$$b(P_0)(z,\partial_z)b(Q_0)(z,\partial_z)q(z)^{-1}\Psi(n,z) = \mu(n)\nu(n)\Psi(n,z), \tag{4.17}$$

i.e., it is bispectral.

Note that in Theorem 4.1 we do not assume that the rational functions $\mu(n)^{-1}$ and $\nu(n)^{-1}$ are well-defined for $n \in \mathbb{Z}$, but only that the "ratios" $\mu(n)^{-1}P_0(n,T)$ and $Q_0(n,T)\nu(n)^{-1}$ are. Because of this a small modification of the original proof from [**3**] is necessary.

First of all since the algebra $\mathcal{B}'_{\alpha,\beta;\varepsilon}$ has no zero divisors, Equation (4.14) implies (see [**3**])

$$(b\nu)(z,\partial_z)\,(b\mu)(z,\partial_z) = (bP_0)(z,\partial_z)q(z)^{-1}(bQ_0)(z,\partial_z). \tag{4.18}$$

For all values of n for which $\mu(n)$ does not vanish we have

$$\Psi(n,z) = \mu(n)^{-1}(bP_0)(z,\partial_z)p_\varepsilon^{\alpha,\beta}(n,z)$$

and (4.17) holds, as a consequence of (4.18). The validity of (4.17) for all $n \in \mathbb{Z}$ follows from the definition (4.15) of $\Psi(n,z)$ and the fact that $p_\varepsilon^{\alpha,\beta}(n,z)$ has an expansion in z around $z=1$ with coefficients that are rational functions in n (recall (3.12)).

Returning to the sets $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ of Darboux transformations from the operators $L_{\alpha,\beta;\varepsilon}(n,T)$, we need to find which of the operators $P(n,T)$ from Equation (3.30) can be expressed in the form $\mu(n)^{-1}P_0(n,T)$ with $\mu(n)$ and $P_0(n,T)$ as above. According to Theorem 4.1 the corresponding operators $L(n,T) \in \mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ will be bispectral with bispectral eigenfunction (3.34) (see also (4.15)). For this we need an invariant description of the linear space of difference operators

$$\begin{aligned}\mathcal{R}_{\alpha,\beta;\varepsilon} = \mathrm{Span}\big\{&\mu(n)^{-1}S(n,T) \mid S(n,T) \in \mathcal{B}_{\alpha,\beta;\varepsilon},\ \mu(n) \in \mathcal{K}_{\alpha,\beta;\varepsilon},\\ &\text{such that } \ \mu(n)^{-1}S(n,T) \ \text{ is well-defined for } \ n \in \mathbb{Z}\big\}.\end{aligned} \tag{4.19}$$

This will be obtained in the next subsection. Here we would like to note that the dual object – the linear space of differential operators

$$\mathcal{R}'_{\alpha,\beta;\varepsilon} = \mathrm{Span}\{g(z)^{-1}G(z,\partial_z) \mid G(z,\partial_z) \in \mathcal{B}'_{\alpha,\beta;\varepsilon},\ g(z) \in \mathcal{K}'_{\alpha,\beta;\varepsilon}\} \tag{4.20}$$

is much easier to describe. It is just the space of differential operators with rational coefficients. This is a consequence of the fact that the commutator

$$[B_{\alpha,\beta}(z,\partial_z), z] = 2(z^2-1)\partial_z + ((\alpha-\beta)+(\alpha+\beta+2)z)$$

is a first order differential operator that belongs to $\mathcal{B}'_{\alpha,\beta;\varepsilon}$ and $z \in \mathcal{K}'_{\alpha,\beta;\varepsilon}$ (see Equation (4.10)). Unfortunately for our proof of the fact that the operators from $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ are bispectral we need the space $\mathcal{R}_{\alpha,\beta;\varepsilon}$, and not the space $\mathcal{R}'_{\alpha,\beta;\varepsilon}$.

4.2. Description of $\mathcal{R}_{\alpha,\beta;\varepsilon}$. Denote by Δ the abstract algebra of difference operators $M(n,T)$ with rational coefficients; that is the algebra over $\mathbb{C}$, generated by rational functions in n, the shift operator T, and its inverse T^{-1}, subject to the relation

$$Th(n) = h(n+1)T, \quad \text{for all rational functions } \ h(n).$$

Here we do not require that the coefficients of an operator $M(n,T)$ in Δ be well-defined for $n \in \mathbb{Z}$. More explicitly these coefficients could have poles at some $n \in \mathbb{Z}$. *The subspace of Δ consisting of operators having this extra regularity property will be denoted by* Δ^{reg}. We will identify the space of

difference operators with rational coefficients acting on functions $f : \mathbb{Z} \to \mathbb{C}$ with Δ^{reg}. In particular, $\widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon} \subset \widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon} \subset \Delta^{\mathrm{reg}}$.

Define an involution I in the algebra Δ acting on rational functions $h(n)$ by

$$(Ih)(n) = h(-(n + 2\varepsilon + \alpha + \beta + 1)) \tag{4.21}$$

and on the shift operator T by

$$I(T) = T^{-1}.$$

The involution I is correctly defined since

$$I(T)\,(Ih)(n) = (Ih)(n+1)\,I(T).$$

Denote the fixed points of I in Δ by Δ^I:

$$\Delta^I = \{M(n,T) \in \Delta \mid I(M(n,T)) = M(n,T)\}. \tag{4.22}$$

Let

$$\phi(n) = \frac{(\varepsilon + \alpha + 1)_n}{(\varepsilon + 1)_n} \tag{4.23}$$

(cf. the definition (2.8) of $p_\varepsilon^{\alpha,\beta}(n,z)$ for $\alpha \notin \mathbb{Z}_{<0}$).

Theorem 4.2. *The space of difference operators $\mathcal{R}_{\alpha,\beta;\varepsilon}$ defined in* (4.19) *is characterized by*

$$\mathcal{R}_{\alpha,\beta;\varepsilon} = \mathrm{Ad}_{\phi(n)}\left(\Delta^I \cap \Delta^{\mathrm{reg}}\right), \tag{4.24}$$

i.e., after conjugation by $\phi(n)^{-1}$ all operators from $\widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon}$ are I-invariant.

Proof. Consider first the case $\alpha \notin \mathbb{Z}_{<0}$. Let

$$\widetilde{p}_\varepsilon^{\alpha,\beta}(n,z) = \phi(n)^{-1} p_\varepsilon^{\alpha,\beta}(n,z). \tag{4.25}$$

The expression (2.8) implies

$$\widetilde{p}_\varepsilon^{\alpha,\beta}(n,z) = F(-(-n+\varepsilon), n+\varepsilon+\alpha+\beta+1; \alpha+1; (1-z)/2). \tag{4.26}$$

Let $\widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon}$, $\widetilde{\mathcal{B}}'_{\alpha,\beta;\varepsilon}$, $\widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon}$, $\widetilde{\mathcal{K}}'_{\alpha,\beta;\varepsilon}$, $\widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon}$, and $\widetilde{\mathcal{R}}'_{\alpha,\beta;\varepsilon}$, denote the $\mathcal{B}$, $\mathcal{K}$ and $\mathcal{R}$ objects associated with the functions $\widetilde{p}_\varepsilon^{\alpha,\beta}(n,z)$ (see the beginning of Section 4.1 and Equations (4.6), (4.7), (4.19) and (4.20) for the appropriate definitions). Obviously

$$\widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon} = \mathrm{Ad}_{\phi(n)}\mathcal{R}_{\alpha,\beta;\varepsilon},\ \widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon} = \mathrm{Ad}_{\phi(n)}\mathcal{B}_{\alpha,\beta;\varepsilon},$$

and $\widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon} = \mathcal{K}_{\alpha,\beta;\varepsilon}$, $\widetilde{\mathcal{R}}'_{\alpha,\beta;\varepsilon} = \mathcal{R}'_{\alpha,\beta;\varepsilon}$, $\widetilde{\mathcal{K}}'_{\alpha,\beta;\varepsilon} = \mathcal{K}'_{\alpha,\beta;\varepsilon}$, $\widetilde{\mathcal{B}}'_{\alpha,\beta;\varepsilon} = \mathcal{B}'_{\alpha,\beta;\varepsilon}$. In this notation, the statement of the theorem is equivalent to

$$\widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon} = \Delta^I \cap \Delta^{\mathrm{reg}}. \tag{4.27}$$

To prove that the l.h.s. of (4.27) is contained in the r.h.s., let us fix an operator $\widetilde{R}(n,T) \in \widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon}$. There exists a difference operator $\widetilde{S}(n,T) \in$

$\widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon}$ and a function $\widetilde{\mu}(n) \in \widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon}$ such that $\widetilde{R}(n,T) = \widetilde{\mu}(n)^{-1}\widetilde{S}(n,T)$. We will prove that all operators from $\widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon}$ are I-invariant. This in particular shows that all functions from $\widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon} \subset \widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon}$ are I-invariant and so are all operators from $\widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon}$.

If $\widetilde{S}(n,T) \in \widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon}$, then there exists a differential operator $G(z,\partial_z)$ for which

$$\widetilde{S}(n,T)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z) = G(z,\partial_z)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z). \tag{4.28}$$

The fact that the hypergeometric function $F(a,b;c;x)$ is symmetric with respect to a and b, and formula (4.26) for $\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z)$ imply

$$I\left(\widetilde{S}(n,T)\right)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z) = G(z,\partial_z)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z). \tag{4.29}$$

Combining (4.28) and (4.29), we conclude that

$$\left(\widetilde{S}(n,T) - I\left(\widetilde{S}(n,T)\right)\right)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z) = 0.$$

This is only possible if

$$I\left(\widetilde{S}(n,T)\right) = \widetilde{S}(n,T).$$

The harder part of the proof of (4.27) is to show that any I-invariant difference operator from Δ^{reg} belongs to $\widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon}$. It is sufficient to prove that for any $\widetilde{R}(n,T) \in \Delta^I$ there exists $\widetilde{S}(n,T) \in \widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon}$ and $\widetilde{\mu}(n) \in \widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon}$ such that

$$\widetilde{R}(n,T) = \widetilde{\mu}(n)^{-1}\widetilde{S}(n,T).$$

First let us write formulas (2.10) and (2.11) in terms of $\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z)$. Equation (2.11) remains unchanged:

$$\lambda_{\varepsilon}(n)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z) = B_{\alpha,\beta}(z,\partial_z)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z), \tag{4.30}$$

while Equation (2.10) becomes

$$\widetilde{L}_{\alpha,\beta;\varepsilon}(n,T)\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z) = z\widetilde{p}_{\varepsilon}^{\alpha,\beta}(n,z) \tag{4.31}$$

with

$$\begin{aligned}\widetilde{L}_{\alpha,\beta;\varepsilon}(n,T) &= \phi(n)^{-1}L_{\alpha,\beta;\varepsilon}(n,T)\phi(n)\\ &= \frac{2(n+\varepsilon+\alpha+1)(n+\varepsilon+\alpha+\beta+1)}{(2n+2\varepsilon+\alpha+\beta+1)(2n+2\varepsilon+\alpha+\beta+2)}T\\ &\quad + \frac{\alpha^2-\beta^2}{(2n+2\varepsilon+\alpha+\beta)(2n+2\varepsilon+\alpha+\beta+2)}\\ &\quad + \frac{2(n+\varepsilon)(n+\varepsilon+\beta)}{(2n+2\varepsilon+\alpha+\beta)(2n+2\varepsilon+\alpha+\beta+1)}T^{-1}.\end{aligned} \tag{4.32}$$

The algebra Δ has a natural $\mathbb{Z}_{\geq 0}$ filtration where Δ_d consists of all operators from Δ with support $[-d, d]$. Denote $\Delta_d^I = \Delta^I \cap \Delta_d$.

We will prove that any difference operator $R_d^I(n,T) \in \Delta_d^I$, can be decomposed as a sum

$$R_d^I(n,T) = \widetilde{\mu}(n)^{-1}\widetilde{S}(n,T) + R_{d-1}^I(n,T) \tag{4.33}$$

where

$$\widetilde{S}(n,T) \in \widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon},\ \widetilde{\mu}(n) \in \widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon}, \tag{4.34}$$

$$R_{d-1}^I(n,T) \in \Delta_{d-1}^I. \tag{4.35}$$

Since $\Delta_0^I = \mathbb{C}(\lambda_e(n))$ (any I-invariant rational function in n is a rational function in $\lambda_e(n)$), by induction on d Equation (4.33) implies that

$$R_d^I(n,T) \in \widetilde{\mathcal{R}}_{\alpha,\beta;\varepsilon}.$$

A straightforward computation yields

$$\begin{aligned} \mathrm{ad}_{\lambda_\varepsilon(n)} T^d &= (\lambda_\varepsilon(n) - \lambda_\varepsilon(n+d))\, T^d \\ &= -d(2n + 2\varepsilon + \alpha + \beta + d + 1)T^d, \end{aligned} \tag{4.36}$$

and thus

$$\begin{aligned} &\mathrm{ad}_{\lambda_\varepsilon(n)}(\mathrm{ad}_{\lambda_\varepsilon(n)} + 1)\widetilde{L}_{\alpha,\beta;\varepsilon} \\ &= 2(\mathrm{Id} + I)((n + \varepsilon + \alpha + 1)(n + \varepsilon + \alpha + \beta + 1)T). \end{aligned}$$

So

$$\begin{aligned} &\left(\mathrm{ad}_{\lambda_\varepsilon(n)}(\mathrm{ad}_{\lambda_\varepsilon(n)} + 1)\widetilde{L}_{\alpha,\beta;\varepsilon}\right)^d \\ &= 2^d(\mathrm{Id} + I)\left(\prod_{i=1}^{d}(n + \varepsilon + \alpha + i)(n + \varepsilon + \alpha + \beta + i)T^d\right) + U_{d-1} \end{aligned} \tag{4.37}$$

for some $U_{d-1} \in \Delta_{d-1}^I$. (Here we use the I-invariance of $L_{\alpha,\beta;\varepsilon}(n,T)$.) Denote for simplicity

$$c_d(n) = 2^d \prod_{i=1}^{d}((n + \varepsilon + \alpha + i)(n + \varepsilon + \alpha + \beta + i))$$

and let

$$R_d^I(n,T) = \sum_{i=-d}^{d} \frac{a_i(n)}{b_i(n)} T^i,\ a_i(n), b_i(n) \in \mathbb{C}[n].$$

Using (4.36) and (4.37) we obtain

(4.38)

$$\begin{aligned} & a_d\left(-\frac{1}{2d}\mathrm{ad}_{\lambda_\varepsilon(n)}-\frac{\alpha+\beta+d+1}{2}-\varepsilon\right) \\ & \quad\cdot b_d\left(\frac{1}{2d}\mathrm{ad}_{\lambda_\varepsilon(n)}-\frac{\alpha+\beta-d+1}{2}-\varepsilon\right) \\ & \quad\cdot c_d\left(\frac{1}{2d}\mathrm{ad}_{\lambda_\varepsilon(n)}-\frac{\alpha+\beta-d+1}{2}-\varepsilon\right)\left(\mathrm{ad}_{\lambda_\varepsilon(n)}(\mathrm{ad}_{\lambda_\varepsilon(n)}+1)\widetilde{L}_{\alpha,\beta;\varepsilon}\right)^d \\ & = (\mathrm{Id}+I)\left(b_d(n)\,(Ib_d)(n)\,c_d(n)\,(Ic_d)(n)\,\frac{a_d(n)}{b_d(n)}T^d\right)+U_{d-1} \end{aligned}$$

for some other $U_{d-1}\in\Delta^I_{d-1}$. There exists a polynomial $q_d(n)$ for which

$$b_d(n)\,(Ib_d)(n)\,c_d(n)\,(Ic_d)(n)=q_d(\lambda_\varepsilon(n))$$

because the polynomial in the l.h.s. is clearly I-invariant. Denote by $\widetilde{S}(n,T)$ the difference operator in (4.38). The l.h.s. of (4.38) implies that $\widetilde{S}(n,T)$ belongs to $\widetilde{\mathcal{B}}_{\alpha,\beta;\varepsilon}$ and the r.h.s. implies

$$R^I_d(n,T)-(q(\lambda_\varepsilon(n)))^{-1}\widetilde{S}(n,T)\in\Delta^I_{d-1}$$

which completes the proof of Theorem 4.2. □

Remark 4.3. Any fuction $\widetilde{\mu}(n)\in\widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon}$ is I-invariant and therefore is a rational function in $\lambda_\varepsilon(n)$. In fact $\widetilde{\mu}(n)$ should be a polynomial in $\lambda_\varepsilon(n)$. Indeed if $\widetilde{\mu}(n)=p(\lambda_\varepsilon(n))/q(\lambda_\varepsilon(n))$ for two polynomials $p(x),q(x)\in\mathbb{C}[x]$ such that $q(x)\not|\,p(x)$, then there exists a differential operator $G(z,\partial_z)$ with rational coefficients such that

$$G(z,\partial_z)\widetilde{p}^{\alpha,\beta}_\varepsilon(n,z)=\frac{p(\lambda_\varepsilon(n))}{q(\lambda_\varepsilon(n))}\widetilde{p}^{\alpha,\beta}_\varepsilon(n,z)$$

which implies

$$p(B_{\alpha,\beta}(z,\partial_z))=G(z,\partial_z)q(B_{\alpha,\beta}(z,\partial_z)). \tag{4.39}$$

This is impossible; if z_0 is a root of $q(x)$ and $p(x)$ of multiplicities $d_1>d_2$, then there exist a holomorphic function $g(z)$ in a domain $\Omega\subset\mathbb{C}$ such that

$$(B_{\alpha,\beta}(z,\partial_z)-z_0)^{d_1}g(z)=0\quad\text{and}\quad q(B_{\alpha,\beta}(z,\partial_z)-z_0)g(z)\neq 0$$

which contradicts with (4.39). Since $\mathcal{K}_{\alpha,\beta;\varepsilon}=\widetilde{\mathcal{K}}_{\alpha,\beta;\varepsilon}$ we finally obtain

$$\begin{aligned} \mathcal{K}_{\alpha,\beta;\varepsilon} &= \mathbb{C}[\lambda_e(n)], \\ \mathcal{A}'_{\alpha,\beta;\varepsilon} &= \mathbb{C}[B_{\alpha,\beta}(z,\partial_z)], \end{aligned}$$

as promised following (4.12) and (4.13).

Remark 4.4. The second order bispectral differential operators of the even case of Duistermaat–Grünbaum's classification [**7**] are obtained as Darboux transformations from the Bessel operators

$$L_k(x,\partial_x)=\partial_x^2-\frac{k(k-1)}{x^2},\ k\in\mathbb{Z}+\frac{1}{2}$$

in the sense of (3.1). More precisely for each operator $L(x,\partial_x)$ of this family there exists a differential operator with rational coefficients $P(x,\partial_x)$ such that

$$L(x,\partial_x)P(x,\partial_x)=P(x,\partial_x)L_k(x,\partial_x).$$

In addition, the operator $P(x,\partial_x)$ satisfies

$$P(x,\partial_x)=P(-x,-\partial_x). \tag{4.40}$$

Let $\mathcal{I}$ denote the involution of the algebra of differential operators with rational coefficients induced by the diffeomorphism $x\mapsto -x$ of $\mathbb{C}$ (i.e., $(\mathcal{I}S)(x,\partial_x)=S(-x,-\partial_x)$). Then (4.40) means that $P(x,\partial_x)$ is invariant under $\mathcal{I}$. This gives the relation of the approach of this paper via the involution I and the space $\mathcal{R}_{\alpha,\beta;\varepsilon}$ to the construction of [**7**].

5. Bispectrality of $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$.

In this section we prove our main result: When the parameters α and β are subject to certain natural integrality conditions, the difference operators from $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ are bispectral. As an example, for each $L(n,T)\in\mathcal{D}^{(2,0)}_{2,0;\varepsilon}$ we find a dual differential operator of order 10.

The conditions (2.13) on α, β, ε are assumed throughout this section.

5.1. Proof of the main result. The conjugation by the function $\phi(n)$ (see (4.23)), used in Theorem 4.2, leads us to consider the functions $\Phi^{(i)}_{\pm}:=\varphi^{(i)}_{\pm}(n)/\phi(n)$, $\Psi^{(i)}_{\pm}:=\psi^{(i)}_{\pm}(n)/\phi(n)$. Because of Equations (3.18)–(3.21) they are explicitly given by the formulas

$$\Phi^{(i)}_{+}(n)=\frac{(-(n+\varepsilon))_i(n+\varepsilon+\alpha+\beta+1)_i}{(\alpha+1)_i(-2)^i}, \tag{5.1}$$

$$\Psi^{(i)}_{+}(n)=\frac{(\varepsilon+\beta+1)_n(\varepsilon+1)_n}{(\varepsilon+\alpha+1)_n(\varepsilon+\alpha+\beta+1)_n}\frac{(-(n+\varepsilon+\alpha+\beta))_i(n+\varepsilon+1)_i}{(-\alpha+1)_i(-2)^i}, \tag{5.2}$$

$$\Phi^{(i)}_{-}(n)=\frac{(\varepsilon+\beta+1)_n}{(-1)^n(\varepsilon+\alpha+1)_n}\frac{(-(n+\varepsilon))_i(n+\varepsilon+\alpha+\beta+1)_i}{(\beta+1)_i 2^i}, \tag{5.3}$$

(5.4)
$$\Psi_-^{(i)}(n) = \frac{(\varepsilon+1)_n}{(-1)^n(\varepsilon+\alpha+\beta+1)_n}\frac{(-(n+\varepsilon+\alpha+\beta))_i(n+\varepsilon+1)_i}{(-\beta+1)_i 2^i}.$$

Lemma 5.1. *If $\alpha \in \mathbb{Z}$, then for $i \le |\alpha| - 1$, $\Phi_+^{(i)}(n)$ and $\Psi_+^{(i)}(n)$ are I-invariant rational functions of n.*

If $\alpha \in \mathbb{Z}$ and $\beta \in \mathbb{Z}$ then for $i \le |\alpha| - 1$, $j \le |\beta| - 1$, $\Phi_+^{(i)}(n)$, $\Psi_+^{(i)}(n)$, $(-1)^n\Phi_-^{(j)}(n)$, and $(-1)^n\Psi_-^{(j)}(n)$ are rational functions of n, $\Phi_+^{(i)}(n)$, $\Psi_+^{(i)}(n)$ are I-invariant, and

$$I\left((-1)^n\Phi_-^{(j)}(n)\right) = (-1)^{\alpha+\beta}\left((-1)^n\Phi_-^{(j)}(n)\right), \tag{5.5}$$

$$I\left((-1)^n\Psi_-^{(j)}(n)\right) = (-1)^{\alpha+\beta}\left((-1)^n\Psi_-^{(j)}(n)\right). \tag{5.6}$$

Proof. First note that

$$\begin{aligned}
&(-(n+\varepsilon))_i(n+\varepsilon+\alpha+\beta+1)_i \\
&= \prod_{r=0}^{i-1}(-(n+\varepsilon)+r)(n+\varepsilon+\alpha+\beta+1+r) \\
&= (-1)^k\prod_{r=0}^{i-1}(\lambda(n) - r(\alpha+\beta+1+r))
\end{aligned}$$

and similarly

$$(-(n+\varepsilon+\alpha+\beta))_i(n+\varepsilon+1)_i = (-1)^i\prod_{r=0}^{i-1}(\lambda(n) - (\alpha+\beta-r)(r+1))$$

are I-invariant polynomials in n.

To prove the first statement of the lemma we use a similar computation. Restricting to the case $\alpha \in \mathbb{Z}_{>0}$:

$$\begin{aligned}
&\frac{(\varepsilon+\beta+1)_n(\varepsilon+1)_n}{(\varepsilon+\alpha+1)_n(\varepsilon+\alpha+\beta+1)_n} \\
&= \frac{(\varepsilon+1)_\alpha(\varepsilon+\beta+1)_\alpha}{(n+\varepsilon+1)_\alpha(n+\varepsilon+\beta+1)_\alpha} \\
&= \frac{(\varepsilon+1)_\alpha(\varepsilon+\beta+1)_\alpha}{\prod_{r=1}^{\alpha}(n+\varepsilon+r)(n+\varepsilon+\beta+\alpha+1-r)} \\
&= \frac{(\varepsilon+1)_\alpha(\varepsilon+\beta+1)_\alpha}{\prod_{r=1}^{\alpha}(\lambda(n)+r(\alpha+\beta+1-r))}.
\end{aligned}$$

The proof of the second statement is analogous. Assuming α, $\beta \in \mathbb{Z}_{>0}$ and $\beta \ge \alpha$ we obtain

$$\frac{(\varepsilon+\beta+1)_n}{(\varepsilon+\alpha+1)_n} = \frac{(\varepsilon+\alpha+n+1)_{\beta-\alpha}}{(\varepsilon+\alpha+1)_{\beta-\alpha}}$$

$$= \frac{q(n)\prod_{r=1}^{\left[\frac{\beta-\alpha}{2}\right]}(n+\varepsilon+\alpha+r)(n+\varepsilon+\beta+1-r)}{(\varepsilon+\alpha+1)_{\beta-\alpha}}$$

$$= \frac{q(n)\prod_{r=1}^{\left[\frac{\beta-\alpha}{2}\right]}(\lambda(n)+(\alpha+r)(\beta+1-r))}{(\varepsilon+\alpha+1)_{\beta-\alpha}}$$

with

$$q(n) = \begin{cases} 1, & \text{if } \beta+\alpha \text{ is even} \\ n+\varepsilon+(\alpha+\beta+1)/2, & \text{if } \beta+\alpha \text{ is odd} \end{cases}.$$

(Since $\alpha \in \mathbb{Z}$, the first condition is equivalent to $2|(\beta-\alpha)$ and the second one to $2 \not| (\beta-\alpha)$.) To finish the proof of (5.5) we just observe that

$$I(n+\varepsilon+(\alpha+\beta+1)/2) = -(n+\varepsilon+(\alpha+\beta+1)/2). \tag{5.7}$$

The remaining cases for $\alpha, \beta \in \mathbb{Z}$ are treated analogously.

The identity (5.6) follows from the analogous formula

$$\frac{(\varepsilon+1)_n}{(\varepsilon+\alpha+\beta+1)_n} = \frac{(\varepsilon+1)_{\alpha+\beta}}{q(n)\prod_{r=1}^{\left[\frac{\beta+\alpha}{2}\right]}(\lambda(n)+r(\alpha+\beta+1-r))}$$

and Equation (5.7).

Throughout this proof, for a real number x by $[x]$ we denote its integer part. □

Theorem 5.2. *Assuming* (2.13)*, the following sets consist of bispectral difference operators:*

1) $\mathcal{D}^{(k,0)}_{\alpha,\beta;\varepsilon}$ *if* $\alpha \in \mathbb{Z}$ *and* $k \leq |\alpha|$,
2) $\mathcal{D}^{(0,l)}_{\alpha,\beta;\varepsilon}$ *if* $\beta \in \mathbb{Z}$ *and* $l \leq |\beta|$,
3) $\mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ *if* $\alpha,\ \beta \in \mathbb{Z}$ *and* $k \leq |\alpha|$, $l \leq |\beta|$.

When the conditions (2.13) are not met but the operator $L_{\alpha,\beta;\varepsilon}(n,T)$ is still well-defined the arguments below can be adapted properly. We do not pursue that here.

Proof. Because of the relation (3.37) the second case follows from the first one.

Let us restrict to instances 1) and 3) of the theorem above. In each of them we can assume that $k+l$ is even using (3.38). Fix an operator $L(n,z) \in \mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$, determined by a choice of the functions $\{f^{(i)}(n)\}_{i=0}^{k+l-1}$, i.e., a choice of admissible values of the complex parameters A, B, C, D (see Section 3.3). It has the eigenfunction $\Psi(n,z)$ defined in (3.34)

$$L(n,T)\Psi(n,z) = z\Psi(n,z),$$

cf. (3.35). We need to show that there exists a differential operator $B(z, \partial_z)$ having $\Psi(n, z)$ as an eigenfunction, that is

$$B(z, \partial_z)\Psi(n, z) = \theta(n)\Psi(n, z)$$

for some function $\theta(n)$.

Define the functions

$$F^{(i)}(n) = f^{(i)}(n)/\phi(n), \quad i = 0, \dots, k + l - 1.$$

Let us put $s := (k + l)/2$ and consider the operator

$$\widetilde{P}(n, T) = (-1)^{nl} \begin{vmatrix} F^{(0)}(n-s) & \dots & F^{(k+l-1)}(n-s) & T^{-s} \\ \dots & \dots & \dots & \dots \\ F^{(0)}(n+s) & \dots & F^{(k+l-1)}(n+s) & T^{s} \end{vmatrix}. \tag{5.8}$$

It is a regular difference operator with kernel given by $\mathrm{Span}\{F^{(i)}(n)\}_{i=0}^{k+l-1}$. Hence it is related to the operator $P(n, T)$ (recall Equation (3.30)) by

$$P(n, T) = d(n)^{-1}\phi(n)^{-1}T^{-s}\widetilde{P}(n, T)\phi(n) \tag{5.9}$$

where

$$d(n) = (-1)^{nl}\det(F^{(i)}(n + j))_{i,j=0,-k-l}^{k+l-1,-1}$$

is the leading coefficient of $T^{-s}\widetilde{P}(n, T)$. Lemma 5.1 implies that $F^{(0)}(n), \dots, F^{(k-1)}(n)$, and $(-1)^n F^{(k)}(n), \dots, (-1)^n F^{(k+l-1)}(n)$ are rational functions in n. This implies that for $i = k, \dots, k+l-1$ and for all $j \in \mathbb{Z}$, $(-1)^n F^{(i)}(n+j)$ are also rational functions in n and thus $\widetilde{P}(n, T)$ has rational coefficients. In addition Lemma 5.1 gives

$$I\left(F^{(i)}(n + j)\right) = F^{(i)}(n - j),\ i = 0, \dots, k - 1,\ j \in \mathbb{Z}$$

and

$$I\left((-1)^n F^{(i)}(n + j)\right) = (-1)^{(\alpha+\beta)/2}\left((-1)^n F^{(i)}(n - j)\right), \quad i = k, \dots, k + l - 1, \ \ j \in \mathbb{Z}.$$

Taking into account that $I(T) = T^{-1}$ we obtain

$$I\left(\widetilde{P}(n, T)\right) = (-1)^s(-1)^{(\alpha+\beta)l}\widetilde{P}(n, T) \tag{5.10}$$

where the factor $(-1)^s$ comes from exchanging the pairs of rows $(1, k+l+1)$, $\dots$, $(s, s + 2)$. Set

$$q(n) = \begin{cases} 1, & \text{if } s + (\alpha + \beta)l \text{ is even} \\ (n + \varepsilon + (\alpha + \beta + 1)/2), & \text{if } s + (\alpha + \beta)l \text{ is odd} \end{cases}$$

and consider the operator

$$\overline{P}(n, T) = q(n)\widetilde{P}(n, T). \tag{5.11}$$

Because of the conditions (2.13), $q(n)$ does not vanish for $n \in \mathbb{Z}$. Taking into account (5.9) one sees that $\overline{P}(n,T)$ is related to $P(n,T)$ by

$$P(n,T) = \frac{d(n)\phi(n-s)}{q(n-s)\phi(n)} T^{-s} \phi(n)^{-1} \overline{P}(n,T)\phi(n). \tag{5.12}$$

Since $\overline{P}(n,T)$ is a regular difference operator and $\phi(n)$ does not vanish for $n \in \mathbb{Z}$ (recall (2.13)), there exists a difference operator with rational coefficients $\overline{Q}(n,T)$ such that

$$\begin{aligned}&(L_{\alpha,\beta;\varepsilon}(n,T)-1)^k (L_{\alpha,\beta;\varepsilon}(n,T)+1)^l \\ &= \left(\phi(n)^{-1}\overline{Q}(n,T)\phi(n)\right)\left(\phi(n)^{-1}\overline{P}(n,T)\phi(n)\right).\end{aligned}$$

From Equations (5.10) and (5.11) it follows that $\overline{P}(n,T)$ is I-invariant. Finally combining this with the I-invariance of $\phi(n)^{-1}L_{\alpha,\beta;\varepsilon}(n,T)\phi(n) = \widetilde{L}_{\alpha,\beta;\varepsilon}(n,T)$ (see (4.32)) implies the I-invariance of the operator $\overline{Q}(n,T)$. Theorem 4.2 now gives

$$\phi(n)^{-1}\overline{P}(n,T)\phi(n),\ \phi(n)^{-1}\overline{Q}(n,T)\phi(n) \in \mathcal{R}_{\alpha,\beta;\varepsilon}.$$

Applying Theorem 4.1 we obtain that the function

$$\overline{\Psi}(n,z) = \phi(n)\overline{P}(n,T)\phi(n)^{-1} p_\varepsilon^{\alpha,\beta}(n,z) \tag{5.13}$$

is an eigenfunction of a differential operator $B(z,\partial_z)$

$$B(z,\partial_z)\overline{\Psi}(n,z) = h(\lambda_\varepsilon(n))\overline{\Psi}(n,z), \tag{5.14}$$

for some polynomial $h(x)$. Because of (5.12) our original function $\Psi(n,z) \in \mathcal{D}^{(k,l)}_{\alpha,\beta;\varepsilon}$ is related to $\overline{\Psi}(n,z)$ by

$$\Psi(n,z) = P(n,T)p_\varepsilon^{\alpha,\beta}(n,z) = \frac{d(n)\phi(n-s)}{q(n-s)\phi(n)} T^{-(k+l)/2}\overline{\Psi}(n,z). \tag{5.15}$$

Equation (5.14) implies that $\Psi(n,z)$ is an eigenfunction of the same operator $B(z,\partial_z)$ with eigenvalue $T^{-(k+l)/2}h(\lambda_\varepsilon(n))$:

$$B(z,\partial_z)\Psi(n,z) = h(\lambda_\varepsilon(n-(k+l)/2))\Psi(n,z). \tag{5.16}$$

□

5.2. An example: The set $\mathcal{D}^{(2,0)}_{2,0,\varepsilon}$. In this final subsection we consider in detail the case $\alpha = 2$, $\beta = 0$, $k = 2$, $l = 0$ and use this example for two different purposes. First we give the reader a guided tour through the results in this paper: We start with the function $p_\varepsilon^{2,0}(n,z)$ from (2.8), give the ingredients needed to build the difference operator $P(n,T)$ (3.30) and the corresponding eigenfunction $\Psi(n,z)$ (3.34), and end with a description of the strategy used in the construction of a differential operator in the variable z giving a bispectral situation. The algebra of possible differential operators in z contains some whose order is lower than the one resulting from

this construction. We close this subsection with an explicit expression for the (essentially unique) bispectral operator of minimal order and material related to this operator.

The functions $\varphi_+^{(i)}(n)$ and $\psi_+^{(i)}(n)$ $(i = 0, 1)$ from (3.18)-(3.19) are given by

$$\varphi_+^{(0)}(n) = \frac{(n+\varepsilon+1)_2}{\kappa}, \quad \varphi_+^{(1)}(n) = \frac{(n+\varepsilon)_4}{6\kappa}, \tag{5.17}$$

$$\psi_+^{(0)}(n) = \frac{\kappa}{(n+\varepsilon+1)_2}, \quad \psi_+^{(1)}(n) = -\frac{\kappa}{2}, \tag{5.18}$$

where

$$\kappa = (\varepsilon+1)(\varepsilon+2). \tag{5.19}$$

The conditions (2.13) reduce to $\varepsilon \notin \mathbb{Z}$.

An element $L(n,T) \in \mathcal{D}_{2,0,\varepsilon}^{(2,0)}$ is determined by a choice of the functions

$$f^{(0)}(n) = A_0\varphi_+^{(0)}(n) + B_0\psi_+^{(0)}(n),$$
$$f^{(1)}(n) = A_1\varphi_+^{(0)}(n) + B_1\psi_+^{(0)}(n) + A_0\varphi_+^{(1)}(n) + B_0\psi_+^{(1)}(n),$$

cf. Section 3.3. We will restrict to the generic case when $A_0 \neq 0$. In this case we can assume that $A_0 = 1$ and $A_1 = 0$ by dividing $f^{(0)}(n)$ by A_0 and then subtracting from $f^{(1)}(n)$ the term $A_1 f^{(0)}(n)$. Recall that $L(n,T)$ depends only on $\mathrm{Span}\{f^{(0)}(n), f^{(1)}(n)\}$. Once this space has been specified by the choice of B_0, B_1 we can build the difference operator $P(n,T)$ as in (3.30) and we get the eigenfunction $\Psi(n,z)$ of $L(n,T)$ from (3.34).

The theory developed in Sections 4 and 5 makes it convenient to introduce the difference operators $\widetilde{P}(n,T)$, see (5.8), and $\overline{P}(n,T)$, see (5.11), related to $P(n,T)$ by (5.9) and (5.12).

The main point in the proof of Theorem 5.2 is that the operator $\overline{P}(n,T)$ defined in (5.11) (see also (5.8)) is I-invariant and thus $\phi(n)\overline{P}(n,T)\phi(n)^{-1} \in \mathcal{R}_{2,0;\varepsilon}$. This implies that the function

$$\overline{\Psi}(n,z) = \phi(n)\overline{P}(n,T)\phi(n)^{-1}p_\varepsilon^{\alpha,\beta}(n,z)$$

(see (5.13)) can be expressed as

$$\overline{\Psi}(n,z) = \mu(n)^{-1}G(z,\partial_z)p_\varepsilon^{\alpha,\beta}(n,z) \tag{5.20}$$

for some differential operator with rational coefficients $G(z,\partial_z)$ and some polynomial $\mu(n)$ (recall the definition (4.19) of $\mathcal{R}_{2,0;\varepsilon}$). Now any operator $B(z,\partial_z)$ that is a Darboux transformation from $h(B_{2,0}(z,\partial_z))$ for some $h(x) \in \mathbb{C}[x]$ via the operator $G(z,\partial_z)$, i.e.,

$$B(z,\partial_z)G(z,\partial_z) = G(z,\partial_z)h(B_{2,0}(z,\partial_z)) \tag{5.21}$$

will satisfy

$$B(z,\partial_z)\overline{\Psi}(n,z) = h(\lambda_\varepsilon(n))\overline{\Psi}(n,z)$$

(a differential analog of (3.3)). The function $\Psi(n,z)$ is related to $\overline{\Psi}(n,z)$ by (5.15) and is also an eigenfunction of $B(z,\partial_z)$ but with eigenvalue $h(\lambda(n-1))$

$$B(z,\partial_z)\Psi(n,z) = h(\lambda_\varepsilon(n-1))\Psi(n,z),$$

see (5.16). Combined with (3.35)

$$L(n,T)\Psi(n,z) = z\Psi(n,z)$$

this gives the desired bispectral pair $(L(n,T), B(z,\partial_z))$.

The I-invariance of the operator $\overline{P}(n,T)$ in this special case can be observed directly. Because of (5.17), (5.18) the functions $F^{(i)}(n) = f^{(i)}(n)/\phi(n)$, $i = 0,1$, see (4.23), are given in terms of

$$\lambda_\varepsilon(n) = (n+\varepsilon)(n+\varepsilon+3)$$

by

$$F^{(0)}(n) = 1 + \frac{B_0\lambda_\varepsilon(n)}{6\kappa},$$
$$F^{(1)}(n) = \frac{B_1\lambda_\varepsilon(n)}{6\kappa} + \frac{\kappa}{(\lambda_\varepsilon(n)+1)}\left(\frac{\kappa}{(\lambda_\varepsilon(n)+1)} - \frac{B_0}{2}\right).$$

The operator $\overline{P}(n,T)$ is given by

$$\overline{P}(n,T) = (n+\varepsilon+3/2)\begin{vmatrix} F^{(0)}(n-1) & F^{(1)}(n-1) & T^{-1} \\ F^{(0)}(n) & F^{(1)}(n) & 1 \\ F^{(0)}(n+1) & F^{(1)}(n+1) & T \end{vmatrix}$$

and it is I-invariant because of the I-invariance of $\lambda_\varepsilon(n)$ and the skew invariance of the factor in front compensating the effect of the exchange of first and third row. An operator $G(z,\partial_z)$ satisfying (5.20) is generated from the proof of Theorem 4.2. It is of high order and the one of minimal order 10 has the following form

$$\begin{aligned} G(z,\partial_z) = {} & (z-1)^6(z+1)^5\partial_z^{10} + (z-1)^5(z+1)^4(57z+7)\partial_z^9 \\ & + 4(z-1)^4(z+1)^3(311z^2+68z-43)\partial_z^8 \\ & + \big(3B_0\kappa^2(z-1)^2(z+1)^2 \\ & \quad + 2(18793z^4+5796z^3-15734z^2-3636z+1501)\big)\partial_z^7 + \cdots. \end{aligned}$$

Theorem 4.1 guarantees that (5.21) is satisfied for some polynomial $h(x)$. It also generates such a polynomial but it is again of high order. The one of minimal order 5 is given by

$$\begin{aligned} h(x-2) = {} & x^5 - 5x^4 + (10B_0\kappa^2+8)x^3 \\ & - (30B_1\kappa^2 + 20B_0\kappa^2 + 4)x^2 - 15B_0^2\kappa^4 x. \end{aligned}$$

Given $G(z,\partial_z)$ Equation (5.21) determines the dual bispectral operator $B(z,\partial_z)$ of $L(n,T)$ of minimal order uniquely. It is given by

$$\begin{aligned}
&B(z,\partial_z)\\
&= (z-1)^5(z+1)^5\partial_z^{10} + 50(z-1)^4z(z+1)^4\partial_z^9\\
&\quad + 5(z-1)^3(z+1)^3(11z-5)(17z+7)\partial_z^8\\
&\quad + 160(z-1)^2(z+1)^2(52z^3-7z^2-28z+1)\partial_z^7\\
&\quad + (30B_0^2\kappa^4z + 120B_1\kappa^2(z-1) + 120B_0\kappa^2)\partial_z^6\\
&\quad + (180B_0\kappa^2(z-1)^2z(z+1)^2\\
&\qquad + 240(z-1)^2(337z^3+504z^2+141z-30))\partial_z^5\\
&\quad + (-30B_1\kappa^2(z-1)^2(z+1)^2 + +120B_0\kappa^2(z-1)(z+1)(8z^2-z-3)\\
&\qquad + 120(z-1)^2(641z^2+758z+161))\partial_z^4\\
&\quad + (-240B_1\kappa^2(z-1)z(z+1) + 240B_0\kappa^2(7z^3-3z^2-7z+1)\\
&\qquad + 960(z-1)^2(26z+19))\partial_z^3\\
&\quad + (-60B_1\kappa^2(z-1)(7z+5) + 120B_0\kappa^2(2z+1)(3z-5) + 1440(z-1)^2)\partial_z^2\\
&\quad - (30B_0^2\kappa^4z + 120B_1\kappa^2(z-1) + 120B_0\kappa^2)\partial_z.
\end{aligned}$$

In the cases $k=1$, $l=0,1$ and $\varepsilon=0$ the dual bispectral operator of minimal order was determined in [**19**, **27**].

References

[1] B. Bakalov, E. Horozov and M. Yakimov, *Bispectral algebras of commuting ordinary differential operators,* Comm. Math. Phys., **190**(**2**) (1997), 331-373, MR 99c:34188, Zbl 0912.34065.

[2] ———, *Highest weight modules over the $W_{1+\infty}$ algebra and the bispectral problem,* Duke Math. J., **93**(**1**) (1998), 41-72, MR 99h:58077.

[3] ———, *General methods for constructing bispectral operators,* Phys. Lett. A, **222**(**1-2**) (1996), 59-66, MR 97i:58160, Zbl 0972.37545.

[4] Yu. Berest, *Huygens' principle and the bispectral problem,* in 'The bispectral problem (Montreal, PQ, 1997),' 11-30, CRM Proc. Lecture Notes, **14**, Amer. Math. Soc., Providence, RI, 1998, MR 99c:58154, Zbl 0897.35043.

[5] Yu. Berest and G. Wilson, *Classification of rings of differential operators on affine curves,* IMRN, **2** (1999), 105-109, MR 2000f:14025, Zbl 0957.14013.

[6] S. Bochner, *Über Sturm–Liouvillesche polynomsysteme,* Math. Z., **29** (1929), 730-736.

[7] J.J. Duistermaat and F.A. Grünbaum, *Differential equations in the spectral parameter,* Comm. Math. Phys., **103**(**2**) (1986), 177-240, MR 88j:58106, Zbl 0625.34007.

[8] P. Etingof and A. Varchenko, *Traces of intertwiners for quantum groups and difference equations*, I, Duke Math. J., **104**(**3**) (2000), 391-432, MR 2001k:17021.

[9] G. Felder, Y. Markov, V. Tarasov and A. Varchenko, *Differential equations compatible with KZ equations,* Math. Phys. Anal. Geom., **3**(**2**) (2000), 139-177, CMP 1 797 943.

[10] F.A. Grünbaum and L. Haine, *A theorem of Bochner, revisited,* in 'Algebraic aspects of integrable systems,' 143-172, Progr. Nonlinear Differential Equations Appl., **26**, Birkhäuser Boston, Boston, MA, 1997, MR 98f:58103, Zbl 0868.35116.

[11] ———, *Associated polynomials, spectral matrices and the bispectral problem,* Methods and Applications of Analysis, **6**(**2**) (1999), 209-224, CMP 1 803 891, Zbl 0956.33007.

[12] F.A. Grünbaum, L. Haine and E. Horozov, *Some functions that generalize the Krall-Laguerre polynomials,* J. Comput. Appl. Math., **106**(**2**) (1999), 271-297, MR 2000k:33017, Zbl 0926.33007.

[13] L. Haine, *Beyond the classical orthogonal polynomials,* in 'The bispectral problem (Montreal, PQ, 1997),' 47-65, CRM Proc. Lecture Notes, **14**, Amer. Math. Soc., Providence, RI, 1998, MR 99d:33008, Zbl 0943.33006.

[14] ———, *The Bochner–Krall problem: Some new perspectives,* to appear in Proceedings of the NATO Workshop on Special Functions, Tempe, Arizona, 2000.

[15] L. Haine and P. Iliev, *Commutative rings of difference operators and an adelic flag manifold,* IMRN, **6** (2000), 281-323, MR 2001d:37109.

[16] ———, *A rational analog of the Krall polynomials,* J. of Phys. A: Math. Gen., **34**(**11**) (2001), 2445-2457.

[17] E. Horozov and T. Milanov, *Fuchsian bispectral operators,* Bull. Sci. Math., **126** (2002), 161-192.

[18] A. Kasman and M. Rothstein, *Bispectral Darboux transformations: The generalized Airy case,* Phys. D, **102**(**3-4**) (1997), 159-176, MR 98g:33005, Zbl 0890.58095.

[19] J. Koekoek and R. Koekoek, *Differential equations for generalized Jacobi polynomials,* J. Comput. Appl. Math., **126**(**1-2**) (2000), 1-31, MR 2001m:34016, Zbl 0970.33004.

[20] H.L. Krall, *Certain differential equations for Tchebicheff polynomials,* Duke Math. J. **4** (1938), 705-718.

[21] W. Magnus, F. Oberhettinger and R. Soni, *Formulas and Theorems for the Special Functions of Mathematical Physics,* Springer Verlag, New York, 1966, MR 38 #1291, Zbl 0143.08502.

[22] D. Mumford, *An algebro–geometric construction of commuting operators and of solutions to the Toda lattice equation, Korteweg deVries equation and related nonlinear equation,* in 'Proceedings of the International Symposium on Algebraic Geometry (Kyoto Univ., Kyoto, 1977),' 115-153, Kinokuniya Book Store, Tokyo, 1978, MR 83j:14041, Zbl 0423.14007.

[23] P. van Moerbeke and D. Mumford, *The spectrum of difference operators and algebraic curves,* Acta Math., **143**(**1-2**) (1979), 93-154, MR 80e:58028, Zbl 0502.58032.

[24] G. Wilson, *Bispectral commutative ordinary differential operators,* J. Reine Angew. Math., **442** (1993), 177-204, MR 94m:58180, Zbl 0781.34051.

[25] ———, *Collisions of Calogero-Moser particles and an adelic Grassmannian,* with an appendix by I.G. Macdonald, Invent. Math., **133**(**1**) (1998), 1-41, MR 99f:58107, Zbl 0906.35089.

[26] P. Wright, *Darboux transformations, algebraic subvarieties of Grassmann manifolds, commuting flows and bispectrality,* Ph.D. Thesis, Univ. California, Berkeley, 1987.

[27] A. Zhedanov, *A method of constructing Krall's polynomials,* J. Comput. Appl. Math., **107** (1999), 1-20, MR 2000m:33012, Zbl 0929.33008.

Received October 18, 2000 and revised January 3, 2001. The first author was partially supported by NSF grant DMS94-00097. The second author was partially supported by NSF grants DMS94-00097 and DMS96-03239.

Department of Mathematics
University of California at Berkeley
Berkeley, CA 94720
E-mail address: grunbaum@math.berkeley.edu

Department of Mathematics
Cornell University
Ithaca, NY 14853
E-mail address: milen@math.cornell.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

PARAHORIC FIXED SPACES IN UNRAMIFIED PRINCIPAL SERIES REPRESENTATIONS

Joshua M. Lansky

Let k be a non-archimedean locally compact field and let G be the set of k-points of a connected reductive group defined over k. Let W be the relative Weyl group of G, and let $\mathcal{H}(G, B)$ be the Hecke algebra of G with respect to an Iwahori subgroup B of G. We compute the effects of $\mathcal{H}(G, B)$ and W on the B-fixed vectors of an unramified principal series representation I of G. We use this computation to determine the dimension of the space of K-fixed vectors in I, where K is a parahoric subgroup of G.

1. Introduction.

Let $\mathbf{G}$ be a reductive group defined over a non-archimedean locally compact field k and let $G = \mathbf{G}(k)$. Let P be a minimal parabolic subgroup of G with Levi decomposition $P = MN$, and let $P^- = MN^-$ be the corresponding decomposition of the opposite parabolic P^-. Let B be an Iwahori subgroup of G with an Iwahori decomposition with respect to P and M, i.e.,

$$B = (B \cap P)(B \cap M)(B \cap P^-).$$

Denote by W the relative Weyl group of G. Let χ be an unramified character of M (i.e., χ is trivial on M_0). Since $M \cong P/N$, χ extends to a character of P which we will also denote by χ. Let δ be the modulus character of P. Define $I(\chi)$ to be the unramified principal series representation of G induced by χ, i.e., the space of all locally constant functions $G \to \mathbb{C}$ such that

$$f(pg) = \chi\delta^{1/2}(p)f(g) \text{ for all } p \text{ in } P,\ g \text{ in } G$$

on which G acts by right translation. It is well-known that the space $I(\chi)^B$ of B-fixed vectors in $I(\chi)$ has dimension $\dim I(\chi)^B = |W|$ [**3**, Prop. 2.1]. In this paper, we generalize this result to the fixed space $I(\chi)^K$ where K is a parahoric subgroup of G containing B.

Let A be a maximal split torus in M and let $\mathcal{N}$ be its normalizer in G. If M_0 is the maximal compact subgroup of M and $\widetilde{W} = \mathcal{N}/M_0$, then we have a surjection $\nu : \widetilde{W} \to W = \mathcal{N}/M$. Let K be a parahoric subgroup of G containing B and let W_K be the finite Coxeter subgroup of $\widetilde{W}$ such that $K = BW_KB$ (see [**4**, §1]). We will prove the following:

Theorem 1.1. *The dimension of* $I(\chi)^K$ *is* $|W/\nu(W_K)|$.

As a Coxeter group, W_K is generated by a canonical finite set S of reflections. Thus

$$I(\chi)^K = \bigcap_{s \in S} I(\chi)^{\langle B, s \rangle}.$$

In Section 3, we explicitly determine the effects of reflections $s \in S$ on $I(\chi)^B$ (Theorem 3.1) and as a corollary the actions of the generators of the Iwahori-Hecke algebra $\mathcal{H}(G, B)$ on $I(\chi)^B$ (Corollary 3.2). We then compute the subspaces $I(\chi)^{\langle B, s \rangle}$ in terms of the usual basis of $I(\chi)$ as given in [**3**, Prop. 2.1]. Then in Section 4, we complete the proof of Theorem 1.1 by showing that the dimension of the intersection of the $I(\chi)^{\langle B, s \rangle}$ is $|W/\nu(W_K)|$.

Let $\mathcal{H}(G, K)$ be the Hecke algebra of compactly supported functions $G \to \mathbb{C}$, bi-invariant by K. Let E be a simple $\mathcal{H}(G, K)$-module. It is known that there is an irreducible admissible representation V of G such that E is isomorphic as a $\mathcal{H}(G, K)$-module to the space V^K of K-fixed vectors [**1**, 2.10]. Since $V^B \supset V^K = E \neq 0$, it follows from a well-known result that V embeds inside some unramified principal series representation I of G so that $\dim E = \dim V^K \leq \dim I^K$. Thus Theorem 1.1 has the following corollary:

Corollary 1.2. *If* K *is a parahoric subgroup of* G *and* E *is a simple module over* $\mathcal{H}(G, K)$*, then*

$$\dim E \leq |W/\nu(W_K)|.$$

Moreover, this bound is sharp.

The sharpness of this bound is a result of the fact that there exist irreducible unramified principal series representations (see e.g., [**2**, Theorem 3.3]) and that for such a representation I, the $\mathcal{H}(G, K)$-module I^K is simple [**1**, 2.10] and, by Theorem 1.1, of dimension $|W/\nu(W_K)|$.

Remark 1.3. While Theorem 1.1 is needed to prove the sharpness in Corollary 1.2, the inequality itself can be proved by a simpler argument. Indeed, it is easily demonstrated that $\dim I(\chi)^K \leq |W/\nu(W_K)|$ by noting that

$$\dim I(\chi)^K \leq |P \backslash G / K|$$

and

$$|P \backslash G / K| = |W/\nu(W_K)|.$$

I would like to express my gratitude to both Benedict Gross and David Pollack for their many helpful suggestions for this paper.

2. Preliminaries.

See [**6**] or [**3**, §1] as a reference for much of the material in this section. In the following, we let k be a non-archimedean locally compact field. We denote by $\mathbf{G}$ a connected reductive algebraic group defined over k with group of

k-points G. Similarly, throughout this section, if $\mathbf{H}$ is any algebraic group defined over k, we will denote its k-points by the corresponding non-bold letter H.

Let $\mathbf{P}$ be a fixed minimal parabolic subgroup of $\mathbf{G}$ containing a maximal split torus $\mathbf{A}$ of $\mathbf{G}$. Denote by $\mathbf{N}$ the unipotent radical of $\mathbf{P}$, and by $\mathbf{M}$ the centralizer of $\mathbf{A}$. Then $\mathbf{P}$ has Levi decomposition $\mathbf{MN}$. Let Φ' denote the set of roots of $\mathbf{G}$ relative to $\mathbf{A}$ and Φ'_{nd} the subset of non-divisible roots. Also, let W be the relative Weyl group.

Denote by $\mathcal{B} = \mathcal{B}(\mathbf{G}, k)$ the Bruhat-Tits building of $\mathbf{G}$ over k and by $\mathcal{A}$ the apartment of $\mathcal{B}$ stabilized by A. The normalizer $\mathcal{N}$ of A in G is then the stabilizer of $\mathcal{A}$ and the maximal compact subgroup M_0 of M is the kernel of the map $\mathcal{N} \to \mathrm{Aut}(\mathcal{A})$. Let $\widetilde{W} = \mathcal{N}/M_0$. Denote by Φ_{aff} the canonical affine root system on $\mathcal{A}$ and by W_{aff} the corresponding affine Weyl group. Then W_{aff} may be identified with a normal subgroup of $\widetilde{W}$.

Fix a special point x_0 in $\mathcal{B}$ and let Φ be the set of affine roots vanishing at x_0. Then Φ is a reduced root system, and we have a bijection between Φ and Φ'_{nd} corresponding to the choice of x_0 . We let Φ^+ be the subset of positive affine roots corresponding to P and Δ the subset of simple roots.

Let C be the unique chamber in $\mathcal{A}$ containing x_0 with the property that every α in Φ^+ takes positive values on C. Denote by B the Iwahori subgroup of G fixing C pointwise and by K_0 the special maximal compact subgroup fixing x_0. Then $W = \mathcal{N}/M \cong (\mathcal{N} \cap K_0)/M_0$, which is the stabilizer of x_0 in $\widetilde{W}$. We will identify these groups throughout. We denote by ν the surjection $\widetilde{W} \to W$. The kernel of ν is the group of translations in $\widetilde{W}$.

For each α in Φ_{aff}, denote by $N(\alpha)$ the pointwise stabilizer of the half-apartment $\{x \in \mathcal{A} \mid \alpha(x) \geq 0\}$. We note that

$$B = M_0 \cdot \prod_{\alpha \in \Phi^+} N(\alpha) \cdot \prod_{\alpha \in \Phi^-} N(\alpha + 1).$$

Let $P_0 \subset P$ be the compact subgroup

$$P \cap K_0 = M_0 \cdot \prod_{\alpha \in \Phi^+} N(\alpha).$$

Let $\Phi = \bigcup \Phi_i$ be the decomposition of Φ into irreducible root systems. Denote by $\widetilde{\Delta}$ the set containing the highest root $\widetilde{\alpha}_i$ of Φ_i for each i. Let

$$\Delta_{\mathrm{aff}} = \{\alpha \in \Phi_{\mathrm{aff}} \mid \alpha \in \Delta \text{ or } \ \alpha = \widetilde{\alpha} - 1 \text{ for some } \ \widetilde{\alpha} \in \widetilde{\Delta}\}.$$

For α in Δ_{aff}, let w_α be the reflection in $\mathrm{Aut}(\mathcal{A})$ through the vanishing hyperplane of α. Then $S_{\mathrm{aff}} = \{w_\alpha \mid \alpha \in \Delta_{\mathrm{aff}}\}$ is a set of involutive generators for the Coxeter group W_{aff}.

For α in Φ, let a_α be the translation $w_\alpha w_{\alpha-1}$ on $\mathcal{A}$. We note that

$$a_{-\alpha} = a_\alpha^{-1} \text{ for any } \alpha \text{ in } \Phi.$$

We let K be a fixed parahoric subgroup of G containing B. Since the triple $(G, B, \mathcal{N})$ is a generalized Tits system (see [**4**, §1]), there exists a special subgroup W_K of W_{aff} such that $K = BW_KB$; W_K is finite as K is compact. We denote by S the subset of S_{aff} generating W_K.

For any w in $\widetilde{W}$, we denote by $q(w)$ the index $[BwB : B]$. Also for α in Φ_{aff}, we let q_α be the index $[N(\alpha - 1) : N(\alpha)]$. We note that $q_{\alpha+2} = q_\alpha$. Since (cf. [**5**, Cor. 2.7])

$$Bw_\alpha B = N(\alpha)w_\alpha B \text{ for } \alpha \text{ in } \Delta, \tag{1}$$

$$Bw_{\widetilde{\alpha}-1}B = N(-\widetilde{\alpha}+1)w_{\widetilde{\alpha}-1}B \text{ for } \widetilde{\alpha} \text{ in } \widetilde{\Delta}, \tag{2}$$

it follows that

$$q(w_\alpha) = q_{\alpha+1} \text{ for } \alpha \text{ in } \Delta, \qquad q(w_{\widetilde{\alpha}-1}) = q_{\widetilde{\alpha}+2} = q_{\widetilde{\alpha}} \text{ for } \widetilde{\alpha} \text{ in } \widetilde{\Delta}.$$

If $\alpha \in \Delta$, we denote by B_α the group $B \cap w_\alpha B w_\alpha$, and if $\widetilde{\alpha} \in \widetilde{\Delta}$, $B_{\widetilde{\alpha}-1}$ denotes the group $B \cap w_{\widetilde{\alpha}-1} B w_{\widetilde{\alpha}-1}$.

Let dx be the Haar measure on G for which B has volume 1. We denote by $\mathcal{H}(G, B)$ the Iwahori-Hecke algebra of compactly supported functions $G \to \mathbb{C}$ bi-invariant by B. The product on $\mathcal{H}(G, B)$ is given by convolution with respect to dx. Fix an unramified character χ of M and let δ be the modulus character of P. Denote by $I(\chi)$ the induced representation $\text{Ind}_P^G(\chi\delta^{1/2})$, i.e., the unramified principal series representation induced by χ as described in Section 1. If x is an element of G, we will denote the action of x on $u \in I(\chi)$ by $u \mapsto x \cdot u$. Note that if $w \in \widetilde{W}$ then the expression $w \cdot u$ is well-defined for $u \in I(\chi)^B$ as w is determined modulo $M_0 \subset B$. A function $h \in \mathcal{H}(G, B)$ acts on $I(\chi)^B$ by the formula

$$h \cdot u = \int_G (x \cdot u)\, h(x)\, dx,$$

where $v \in I(\chi)^B$.

Let $C_c^\infty(G)$ be the space of locally constant, compactly supported functions $G \to \mathbb{C}$. The map $\mathcal{P}_\chi : C_c^\infty(G) \to I(\chi)$ defined by

$$\mathcal{P}_\chi(f)(g) = \int_P \chi^{-1}\delta^{1/2}(p)\, f(pg)\, dp$$

(where dp is the left Haar measure on P giving P_0 measure 1) is a G-equivariant surjection. The functions $\phi_{w,\chi} = \mathcal{P}_\chi(\text{ch}_{BwB})$ (w in W) form a basis of the subspace of B-fixed vectors $I(\chi)^B$ [**3**, Prop. 2.1]. Concretely, for $p \in P$,$w' \in W$ and $b \in B$, $\phi_{w,\chi}(pw'b)$ equals $\chi\delta^{1/2}(p)$ if $w' = w$ and is zero otherwise.

3. The effect of $W_{\mathbf{aff}}$ on $I(\chi)^B$.

The goal of this section is to compute the effect of $s \in S_{\text{aff}}$ on $I(\chi)^B$. This will be important for the proof in the following section since we will need to determine the space $I(\chi)^{\langle B,s\rangle}$ of vectors in $I(\chi)^B$ fixed by s.

Theorem 3.1. *Suppose that $w \in W$, $\alpha \in \Delta$ and $\widetilde{\alpha} \in \widetilde{\Delta}$. Then*

$$w_\alpha \cdot \phi_{w,\chi} = \begin{cases} \mathrm{ch}_{Pww_\alpha B_\alpha}\phi_{ww_\alpha,\chi} & \text{if } w\alpha \in \Phi^+ \\ \phi_{ww_\alpha,\chi} + \mathrm{ch}_{Pw(B-B_\alpha)}\phi_{w,\chi} & \text{if } w\alpha \in \Phi^-, \end{cases}$$

$$w_{\widetilde{\alpha}-1} \cdot \phi_{w,\chi} = \begin{cases} \chi\delta^{1/2}(a_{w\widetilde{\alpha}})\mathrm{ch}_{Pww_{\widetilde{\alpha}}B_{\widetilde{\alpha}-1}}\phi_{ww_{\widetilde{\alpha}},\chi} & \text{if } w\widetilde{\alpha} \in \Phi^- \\ \chi\delta^{1/2}(a_{w\widetilde{\alpha}})\phi_{ww_{\widetilde{\alpha}},\chi} + \mathrm{ch}_{Pw(B-B_{\widetilde{\alpha}-1})}\phi_{w,\chi} & \text{if } w\widetilde{\alpha} \in \Phi^+. \end{cases}$$

Proof. For any s in S_{aff}, $g \in G$,

$$(s \cdot \phi_{w,\chi})(g) = \phi_{w,\chi}(gs).$$

The Iwasawa decomposition enables us to write $g = p'w'b'$ for some p' in P, w' in W, and b' in B. We will evaluate $\phi_{w,\chi}(gs) = \phi_{w,\chi}(p'w'b's)$ by determining the double coset in which $p'w'b's$ lies.

We first consider $s = w_\alpha$ for $\alpha \in \Delta$. Now if $w'\alpha \in \Phi^+$ then by (1)

$$\begin{aligned} p'w'b'w_\alpha &\in p'w'Bw_\alpha B \\ &= p'w'N(\alpha)w_\alpha B \\ &= p'N(w'\alpha)w'w_\alpha B \\ &\subset (p'N)w'w_\alpha B. \end{aligned}$$

Since $\chi\delta^{1/2}$ is trivial on N, it follows that $\phi_{w,\chi}(p'w'b'w_\alpha)$ equals $\chi\delta^{1/2}(p')$ if $w = w'w_\alpha$ and 0 otherwise.

If, on the other hand, $w'\alpha \in \Phi^-$ then suppose first that $b' \in B_\alpha$. Then

$$p'w'b'w_\alpha \in p'w'b'w_\alpha B = p'w'w_\alpha B$$

since $w_\alpha B_\alpha w_\alpha \subset B$. Thus $\phi_{w,\chi}(p'w'b'w_\alpha)$ equals $\chi\delta^{1/2}(p')$ if $w = w'w_\alpha$ and 0 otherwise.

Lastly, suppose that $w'\alpha \in \Phi^-$ and $b' \in B - B_\alpha$. It is easily deduced from $w'\alpha \in \Phi^-$ that

$$Pw'Bw_\alpha B = Pw'w_\alpha B \cup Pw'B.$$

Moreover, one can show that $p'w'b'w_\alpha \in Pw'B$ if and only if b' is an element of $B - B_\alpha$. Thus $p'w'b'w_\alpha = pw'b$ for some $p \in P$, $b \in B$. Since

$$p^{-1}p' = w'bw_\alpha b'^{-1}w'^{-1} \in P \cap K_0 = P_0$$

and since $\chi\delta^{1/2}$ is trivial on P_0, we have that $\chi\delta^{1/2}(p) = \chi\delta^{1/2}(p')$. Therefore, $\phi_{w,\chi}(p'w'b'w_\alpha)$ equals $\chi\delta^{1/2}(p')$ if $w = w'$ and 0 otherwise.

Note that $w'\alpha \in \Phi^{\pm}$ if and only if $w'w_\alpha\alpha = -w'\alpha \in \Phi^{\mp}$. Using this, we assemble the preceding cases to obtain that

$$(w_\alpha \cdot \phi_{w,\chi})(p'w'b') = \begin{cases} \chi\delta^{1/2}(p') & \text{if } w\alpha \in \Phi^+,\ w' = ww_\alpha,\ b' \in B_\alpha \\ \chi\delta^{1/2}(p') & \text{if } w\alpha \in \Phi^-,\ w' = ww_\alpha \\ \chi\delta^{1/2}(p') & \text{if } w\alpha \in \Phi^-,\ w' = w,\ b' \in B - B_\alpha \\ 0 & \text{otherwise.} \end{cases}$$

This immediately implies the first result of the theorem.

We now prove the second formula by calculating $w_{\widetilde{\alpha}-1} \cdot \phi_{w,\chi}$ for $\widetilde{\alpha} \in \widetilde{\Delta}$. Assume first that $w'\widetilde{\alpha} \in \Phi^-$. Then by (2)

$$\begin{aligned} p'w'b'w_{\widetilde{\alpha}-1} &\in p'w'Bw_{\widetilde{\alpha}-1}B \\ &= p'w'N(-\widetilde{\alpha}+1)w_{\widetilde{\alpha}-1}B \\ &= p'N(-w'\widetilde{\alpha}+1)w'w_{\widetilde{\alpha}}a_{\widetilde{\alpha}}B \\ &\subset (p'a_{-w'\widetilde{\alpha}}N)w'w_{\widetilde{\alpha}}B. \end{aligned}$$

Since χ is trivial on N, it follows that $\phi_{w,\chi}(p'w'b'w_{\widetilde{\alpha}-1})$ equals $\chi\delta^{1/2}(p'a_{-w'\widetilde{\alpha}})$ if $w = w'w_{\widetilde{\alpha}}$ and 0 otherwise.

Now suppose that $w'\widetilde{\alpha} \in \Phi^+$ and that $b' \in B_{\widetilde{\alpha}-1}$. Then

$$p'w'b'w_{\widetilde{\alpha}-1} \in p'w'b'w_{\widetilde{\alpha}-1}B = p'w'w_{\widetilde{\alpha}-1}B = (p'a_{-w'\widetilde{\alpha}})w'w_{\widetilde{\alpha}}B$$

since $w_{\widetilde{\alpha}-1}B_{\widetilde{\alpha}-1}w_{\widetilde{\alpha}-1} \subset B$. It follows that $\phi_{w,\chi}(p'w'b'w_{\widetilde{\alpha}-1})$ is equal to $\chi\delta^{1/2}(p'a_{-w'\widetilde{\alpha}})$ if $w = w'w_{\widetilde{\alpha}}$ and 0 otherwise.

Finally, suppose that $b' \in B - B_{\widetilde{\alpha}-1}$. As before, it can be shown that

$$Pw'Bw_{\widetilde{\alpha}-1}B = Pw'w_{\widetilde{\alpha}}B \cup Pw'B,$$

and furthermore that $p'w'b'w_{\widetilde{\alpha}-1} \in Pw'B$ if and only if b' is an element of $B - B_{\widetilde{\alpha}-1}$. Hence $p'w'b'w_{\widetilde{\alpha}-1} = pw'b$ for some $p \in P$, $b \in B$. It is easily shown that this forces $p^{-1}p' \in NP_0$ so that $\chi\delta^{1/2}(p) = \chi\delta^{1/2}(p')$. Thus $\phi_{w,\chi}(p'w'b'w_{\widetilde{\alpha}-1})$ equals $\chi\delta^{1/2}(p')$ if $w = w'$ and 0 otherwise.

Noting that $w'\widetilde{\alpha} \in \Phi^{\pm}$ if and only if $w'w_{\widetilde{\alpha}}\widetilde{\alpha} = -w'\widetilde{\alpha} \in \Phi^{\mp}$, we obtain

$$\begin{aligned} &(w_\alpha \cdot \phi_{w,\chi})(p'w'b') \\ &= \begin{cases} \chi\delta^{1/2}(a_{w\widetilde{\alpha}})\chi\delta^{1/2}(p') & \text{if } w\widetilde{\alpha} \in \Phi^-,\ w' = ww_{\widetilde{\alpha}},\ b' \in B_{\widetilde{\alpha}-1} \\ \chi\delta^{1/2}(a_{w\widetilde{\alpha}})\chi\delta^{1/2}(p') & \text{if } w\widetilde{\alpha} \in \Phi^+,\ w' = ww_{\widetilde{\alpha}} \\ \chi\delta^{1/2}(p') & \text{if } w\widetilde{\alpha} \in \Phi^+,\ w' = w,\ b' \in B - B_{\widetilde{\alpha}-1} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The second result follows. □

Theorem 3.1 has the following corollary giving the action of ch_{BsB} for s in S_{aff}.

Corollary 3.2. *Suppose that $w \in W$, $\alpha \in \Delta$ and $\widetilde{\alpha} \in \widetilde{\Delta}$. Then*

$$\mathrm{ch}_{Bw_\alpha B} \cdot \phi_{w,\chi} = \begin{cases} \phi_{ww_\alpha,\chi} & \text{if } w\alpha \in \Phi^+ \\ q_{\alpha+1}\phi_{ww_\alpha,\chi} + (q_{\alpha+1}-1)\phi_{w,\chi} & \text{if } w\alpha \in \Phi^-, \end{cases}$$

$$\mathrm{ch}_{Bw_{\widetilde{\alpha}-1}B} \cdot \phi_{w,\chi} = \begin{cases} \chi\delta^{1/2}(a_{w\widetilde{\alpha}})\phi_{ww_{\widetilde{\alpha}},\chi} & \text{if } w\widetilde{\alpha} \in \Phi^- \\ \chi\delta^{1/2}(a_{w\widetilde{\alpha}})q_{\widetilde{\alpha}}\phi_{ww_{\widetilde{\alpha}},\chi} + (q_{\widetilde{\alpha}}-1)\phi_{w,\chi} & \text{if } w\widetilde{\alpha} \in \Phi^+. \end{cases}$$

Proof. We prove the first formula in the case $w\alpha \in \Phi^-$. The other cases are handled similarly. For $g \in G$ we have

$$\begin{aligned} (\mathrm{ch}_{Bw_\alpha B} \cdot \phi_{w,\chi})(g) &= \int_G \phi_{w,\chi}(gx)\mathrm{ch}_{Bw_\alpha B}(x)dx \\ &= \int_{Bw_\alpha B} \phi_{w,\chi}(gx)dx \\ &= \sum_n \phi_{w,\chi}(gnw_\alpha) \\ &= \sum_n (w_\alpha \cdot \phi_{w,\chi})(gn), \end{aligned}$$

where n ranges over a set of representatives in $N(\alpha)$ for $N(\alpha)/N(\alpha+1)$.

If $g \in Pww_\alpha B$ then so is gn for each of the $q_{w_\alpha} = q_{\alpha+1}$ representatives n. On the other hand, if $g \in PwB$, then $gn \in Pw(B - B_\alpha)$ for precisely $q_{\alpha+1} - 1$ of the representatives n. Thus

$$\begin{aligned} (\mathrm{ch}_{Bw_\alpha B} \cdot \phi_{w,\chi})(g) &= \sum_n (w_\alpha \cdot \phi_{w,\chi})(gn) \\ &= \sum_n \left[\phi_{ww_\alpha,\chi}(gn) + \mathrm{ch}_{Pw(B-B_\alpha)}(gn)\phi_{w,\chi}(gn)\right] \\ &= q_{\alpha+1}\phi_{ww_\alpha,\chi}(g) + (q_{\alpha+1}-1)\phi_{w,\chi}(g). \end{aligned}$$

□

The following corollary of Theorem 3.1 gives a basis for $I(\chi)^{\langle B,s\rangle}$, $s \in S_{\mathrm{aff}}$.

Corollary 3.3. *Suppose $\alpha \in \Delta$ and $\widetilde{\alpha} \in \widetilde{\Delta}$. Then*

(i) $\{\phi_{w,\chi} + \phi_{ww_\alpha,\chi} \mid w \in W, w\alpha \in \Phi^+\}$ *is a basis for the fixed space* $I(\chi)^{\langle B,w_\alpha\rangle}$.

(ii) $\{\phi_{w,\chi} + \chi\delta^{1/2}(a_{w\widetilde{\alpha}})\phi_{ww_{\widetilde{\alpha}},\chi} \mid w \in W, w\widetilde{\alpha} \in \Phi^+\}$ *is a basis for the fixed space* $I(\chi)^{\langle B,w_{\widetilde{\alpha}-1}\rangle}$.

Proof. Let $s \in S_{\mathrm{aff}}$. Note that

$$s \cdot I(\chi)^B \cap I(\chi)^B = I(\chi)^{sBs} \cap I(\chi)^B = I(\chi)^{\langle sBs,B\rangle} = I(\chi)^{\langle B,s\rangle}.$$

Thus $I(\chi)^{\langle B,s\rangle}$ is precisely the set of vectors in $I(\chi)^B$ sent to $I(\chi)^B$ by s. It is clear from Theorem 3.1 that if $s = w_\alpha$ this set is spanned by

$$\{\phi_{w,\chi} + \phi_{ww_\alpha,\chi} \mid w \in W, w\alpha \in \Phi^+\},$$

and if $s = w_{\widetilde{\alpha}-1}$ this set is spanned by

$$\{\phi_{w,\chi} + \chi\delta^{1/2}(a_{w\widetilde{\alpha}})\phi_{ww_{\widetilde{\alpha}},\chi} \mid w \in W, w\widetilde{\alpha} \in \Phi^+\}.$$

□

4. Proof of Theorem 1.1.

We now prove that the dimension of

$$I(\chi)^K = I(\chi)^{BW_KB} = \bigcap_{s\in S} I(\chi)^{\langle B,s\rangle}$$

is equal to $|W/\nu(W_K)|$.

Suppose that $f = \sum_{w\in W} c(w)\phi_{w,\chi}$ is a vector in $I(\chi)^B$ with the $c(w) \in \mathbb{C}$. Then it is easily deduced from Corollary 3.3 that $f \in \bigcap_{s\in S} I(\chi)^{\langle B,s\rangle}$ if and only if for all $w \in W$,

$$c(ww_\alpha) = c(w) \text{ for all } \alpha \in \Delta \text{ with } w_\alpha \in S \tag{3}$$

$$c(ww_{\widetilde{\alpha}}) = \chi\delta^{1/2}(a_{w\widetilde{\alpha}})c(w) \text{ for all } \widetilde{\alpha} \in \widetilde{\Delta} \text{ with } w_{\widetilde{\alpha}-1} \in S. \tag{4}$$

Let V be the space of functions $c : W \to \mathbb{C}$ satisfying (3) and (4). Then $\dim I(\chi)^K = \dim V$. Since $\nu(w_{\beta-1}) = \nu(w_\beta) = w_\beta$ for all $\beta \in \Phi$, it follows that $c(w)$ determines $c(ww')$ for all $w' \in \langle \nu(s) \mid s \in S\rangle = \nu(W_K)$ so

$$\dim V \le |W/\nu(W_K)|.$$

We will prove that $\dim V = |W/\nu(W_K)|$.

Remark 4.1. We note that if $W_K \subset W$ (i.e., if $K \subset K_0$) then it is clear that $\dim V = \dim I(\chi)^K = |W/\nu(W_K)|$ since in this case only the relations in (3) appear.

Since W_K is finite, it contains no non-trivial translations so ν is injective on W_K. Thus $\nu(W_K) \cong W_K$, and $\nu(W_K)$ is generated as a Coxeter group by $\nu(S)$. We will denote the element of W_K corresponding to $t \in \nu(S)$ by $\nu^{-1}(t)$. Define recursively a function $[\ ]$ from the set of finite sequences of elements of $\nu(S)$ to W_{aff}. Let $t_1, \dots, t_n \in \nu(S)$. For the empty sequence $\emptyset$, let $[\emptyset] = e$. Define

$$[t_1] = \begin{cases} e & \text{if } \nu^{-1}(t_1) = w_\alpha,\ \alpha \in \Delta \\ a_{\widetilde{\alpha}} & \text{if } \nu^{-1}(t_1) = w_{\widetilde{\alpha}-1},\ \widetilde{\alpha} \in \widetilde{\Delta}, \end{cases}$$

and then set

$$[t_1, \dots, t_n] = \begin{cases} [t_1, \dots, t_{n-1}] & \text{if } \nu^{-1}(t_n) = w_\alpha,\ \alpha \in \Delta \\ [t_1, \dots, t_{n-1}]\, a_{t_1\cdots t_{n-1}\widetilde{\alpha}} & \text{if } \nu^{-1}(t_n) = w_{\widetilde{\alpha}-1},\ \widetilde{\alpha} \in \widetilde{\Delta}. \end{cases}$$

It follows easily from the definition of [ ] that

$$[t_1,\dots,t_k](t_1\cdots t_k)[t_{k+1},\dots,t_n](t_1\cdots t_k)^{-1}=[t_1,\dots,t_n]. \tag{5}$$

We claim that the element $[t_1,\dots,t_n]$ of W_{aff} depends only on the product $t_1\cdots t_n$ and not on the particular sequence $t_1,\dots,t_n$.

Lemma 4.2. *Let $t_1,\dots,t_n,u_1,\dots,u_m$ be elements of $\nu(S)$ such that*

$$t_1\cdots t_n=u_1\cdots u_m.$$

Then $[t_1,\dots,t_n]=[u_1,\dots,u_m]$.

Proof. Since $(\nu(W_K),\nu(S))$ is a Coxeter group, the word $t_1\cdots t_n$ is obtainable from $u_1\cdots u_m$ via the basic Coxeter group relations among the elements of $\nu(S)$, i.e., those of the form $(tu)^{m(t,u)}=e$, where $t,u\in\nu(S)$ and $m(t,u)$ is some number in $\{1,2,3,4,6\}$ (see e.g. [**5**, 1.6]). Therefore, it suffices to show that [ ] remains unchanged when a subsequence of consecutive terms in a sequence $t_1,\dots,t_n$ is deleted according to such a relation. In fact, due to (5) one need only show that

$$\underbrace{[t,u,t,u,\dots,t,u]}_{m(t,u)}=[\emptyset]=e \tag{6}$$

for each basic relation $(tu)^{m(t,u)}=e$ among the elements of $\nu(S)$.

It is clear that (6) holds if $\nu^{-1}(t),\nu^{-1}(u)\in W$. Therefore we shall consider only those relations which involve some reflection $t\in\nu(S)$ such that $\nu^{-1}(t)\notin W$. Such a t is necessarily of the form $w_{\widetilde{\alpha}}=\nu(w_{\widetilde{\alpha}-1})$ for some $\widetilde{\alpha}\in\widetilde{\Delta}$. The basic relations involving $w_{\widetilde{\alpha}}$ are of the form

$$(w_{\widetilde{\alpha}}u)^m=e \tag{7}$$

where $u\in\nu(S)$ and $m\in\{1,2,3,4\}$. (It is never the case that $m=6$.)

First consider the case $m=1$. Here u must equal $w_{\widetilde{\alpha}}$ so (6) holds as

$$[w_{\widetilde{\alpha}},w_{\widetilde{\alpha}}]=a_{\widetilde{\alpha}}a_{w_{\widetilde{\alpha}}\widetilde{\alpha}}=a_{\widetilde{\alpha}}a_{-\widetilde{\alpha}}=e.$$

Now suppose that $m>1$ and $\nu^{-1}(u)\in W$ in (7). Then

$$\underbrace{[w_{\widetilde{\alpha}},u,\dots,w_{\widetilde{\alpha}},u]}_{m}=a_{\widetilde{\alpha}}\dots a_{(w_{\widetilde{\alpha}}u)^{m-1}\widetilde{\alpha}}.$$

Since $w_{\widetilde{\alpha}}u$ is a rotation of order m, $\widetilde{\alpha}+\dots+(w_{\widetilde{\alpha}}u)^{m-1}\widetilde{\alpha}=0$ so (6) holds as

$$a_{\widetilde{\alpha}}\dots a_{(w_{\widetilde{\alpha}}u)^{m-1}\widetilde{\alpha}}=e.$$

Finally, suppose $m>1$ and $\nu^{-1}(u)\notin W$ in (7). In this case, it follows that $m=2$ and $u=w_{\widetilde{\beta}}$ for some $\widetilde{\beta}\in\widetilde{\Delta}$. Then $w_{\widetilde{\beta}}(\widetilde{\alpha})=\widetilde{\alpha}$ and $w_{\widetilde{\alpha}}(\widetilde{\beta})=\widetilde{\beta}$. It follows that (6) holds again as

$$[w_{\widetilde{\alpha}},w_{\widetilde{\beta}},w_{\widetilde{\alpha}},w_{\widetilde{\beta}},]=a_{\widetilde{\alpha}}a_{w_{\widetilde{\alpha}}\widetilde{\beta}}a_{w_{\widetilde{\alpha}}w_{\widetilde{\beta}}\widetilde{\alpha}}a_{w_{\widetilde{\alpha}}w_{\widetilde{\beta}}w_{\widetilde{\alpha}}\widetilde{\beta}}=a_{\widetilde{\alpha}}a_{\widetilde{\beta}}a_{-\widetilde{\alpha}}a_{-\widetilde{\beta}}=e.$$

□

Let $t_1, \dots, t_n \in \nu(S)$. Since $[t_1, \dots, t_n]$ depends only on the product $t_1 \cdots t_n$, $[\ ]$ gives a function $\nu(W_K) \to W_{\text{aff}}$, which we will also denote by $[\ ]$. Explicitly, for $w \in \nu(W_K)$, $[w] = [t_1, \dots, t_n]$ for any $t_1, \dots, t_n \in \nu(S)$ with $w = t_1 \cdots t_n$. Note that $[\ ]$ is a 1-cocycle from $\nu(W_K)$ to the group of translations in W_{aff}.

Proposition 4.3. *The space V of functions $W \to \mathbb{C}$ satisfying* (3) *and* (4) *has dimension $|W/\nu(W_K)|$.*

Proof. Let R be a set of representatives for the left cosets of $\nu(W_K)$ in W. For each $\sigma \in R$, define the function $c_\sigma : W \to \mathbb{C}$ by setting

$$c_\sigma(w) = \begin{cases} \chi\delta^{1/2}([w']) & \text{if } w = \sigma w' \in \sigma\nu(W_K) \\ 0 & \text{if } w \notin \sigma\nu(W_K). \end{cases}$$

The c_σ are clearly linearly independent and are $|W/\nu(W_K)|$ in number. It suffices then to show that the c_σ are in V.

Fix $\sigma \in R$. Let α be an element of Δ such that $w_\alpha \in S$. If $w \notin \sigma\nu(W_K)$ then $ww_\alpha \notin \sigma\nu(W_K)$ so

$$c_\sigma(w) = 0 = c_\sigma(ww_\alpha).$$

If $w = \sigma w' \in \sigma\nu(W_K)$ then

$$c_\sigma(ww_\alpha) = c_\sigma(\sigma w' w_\alpha) = \chi\delta^{1/2}([w'w_\alpha]) = \chi\delta^{1/2}([w']) = c_\sigma(w).$$

Thus (3) holds for c_σ.

Now let $\widetilde{\alpha}$ be an element of $\widetilde{\Delta}$ such that $w_{\widetilde{\alpha}-1} \in S$. As before, if $w \notin \sigma\nu(W_K)$ then

$$c_\sigma(w) = 0 = \chi\delta^{1/2}(a_{w\widetilde{\alpha}})c_\sigma(ww_{\widetilde{\alpha}}).$$

And if $w = \sigma w' \in \sigma\nu(W_K)$ then

$$\begin{aligned} c_\sigma(ww_{\widetilde{\alpha}}) &= c_\sigma(\sigma w' w_{\widetilde{\alpha}}) \\ &= \chi\delta^{1/2}([w'w_{\widetilde{\alpha}}]) \\ &= \chi\delta^{1/2}([w']a_{w'\widetilde{\alpha}}) \\ &= \chi\delta^{1/2}([w'])\chi\delta^{1/2}(a_{w'\widetilde{\alpha}}) \\ &= \chi\delta^{1/2}(a_{w'\widetilde{\alpha}})c_\sigma(w). \end{aligned}$$

Thus c_σ satisfies (4) and lies in V. □

It follows that $\dim I(\chi)^K = \dim V = |W/\nu(W_K)|$.

References

[1] I.N. Bernstein and A.V. Zelevinskii, *Representations of the group* $\mathrm{GL}(n, F)$*, where* F *is a non-Archimedean local field*, Russian Math. Surveys, **31**(**3**) (1976), 1-66, MR 54 #12988, Zbl 0348.43007.

[2] P. Cartier, *Representations of* $\mathfrak{p}$*-adic groups: A survey*, Automorphic Forms, Representations, and L-functions (Providence, RI) (Armand Borel and William Casselman, eds.), Proc. Symp. Pure Math., **33**, Amer. Math. Soc., (1977), 111-155, MR 81e:22029, Zbl 0421.22010.

[3] W. Casselman, *The unramified principal series of* $\mathfrak{p}$*-adic groups* I, Compositio Math., **40** (1980), 387-406, MR 83a:22018, Zbl 0472.22004.

[4] N. Iwahori, *Generalized Tits system* (*Bruhat decomposition*) *on* $\mathfrak{p}$*-adic semisimple groups*, Algebraic Groups and Discontinuous Subgroups (Providence, RI) (Armand Borel and George D. Mostow, eds.), Proc. Symp. Pure Math., **9**, Amer. Math. Soc., (1966), 71-89, MR 35 #6693, Zbl 0199.06901.

[5] N. Iwahori and H. Matsumoto, *On some Bruhat decompositions and the structure of the Hecke ring of* $\mathfrak{p}$*-adic Chevalley groups*, Inst. Hautes Études Sci. Publ. Math., **25** (1965), 237-280, MR 32 #2486.

[6] J. Tits, *Reductive groups over local fields*, Automorphic Forms, Representations, and L-functions (Providence, RI) (Armand Borel and William Casselman, eds.), Proc. Symp. Pure Math., **33**, Amer. Math. Soc., (1977), 29-69, MR 80h:20064, Zbl 0415.20035.

Received August 1, 2000 and revised January 9, 2001.

Department of Mathematics
University of Rochester
Rochester, New York 14627
E-mail address: lansky@math.rochester.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

A SPLITTING THEOREM FOR ALEXANDROV SPACES

Yukihiro Mashiko

We use a notion of differentiability for functions on Alexandrov spaces and prove a splitting theorem for Alexandrov spaces admitting affine functions with such differentiability.

0. Introduction.

A classical result of Toponogov [**12**] states that if a complete Riemannian manifold M with nonnegative sectional curvature contains a straight line, then M is isometric to the metric product of a nonnegatively curved manifold and a line. We then know that the Busemann function associated with the straight line is an *affine function*, namely, a function which is affine on each unit speed geodesic in the one variable sense. After the theorem, many generalizations were proved. Cheeger-Gromoll's theorem [**2**] is the most excellent one among them.

An *Alexandrov space with curvature bounded below by* $\kappa \in R$ is a locally compact, complete and path connected inner metric space on which the triangle comparison theorem holds (see [**1**]). For simplicity, we denote by $curv \geq \kappa$ the lower curvature bound. The direct generalization of the Toponogov theorem for Alexandrov spaces with $curv \geq 0$ was proved early in 1967 by A. Milka [**8**]. We see that this is essentially implied by the rigidity of geodesic triangles and hinges in the Global Comparison Theorem (see Fact 1.0).

The author has shown in [**7**] that if a 2-dimensional Alexandrov space X with $curv \geq -\kappa^2$ without boundary admits a nontrivial affine function, then X is isometric to flat R^2 or flat $S^1 \times R$. In the present paper, we extend this to higher dimensional Alexandrov spaces, possibly with nonempty boundary, admitting affine functions with a new notion of differentiability. Innami [**6**] showed that every complete Riemannian manifold admitting a nontrivial affine function splits isometrically into the metric product of a line and a Riemannian manifold. Affine functions on complete Riemannian manifolds naturally possess the differentiability introduced in this paper.

We shall define some notion needed to state our main theorem. Let X be an n-dimensional Alexandrov space with $curv \geq -\kappa^2$ and $n \geq 2$, $\kappa > 0$. We denote by pq a minimal geodesic from p to q and by $|p, q|$ the distance

between p and q. We put

$$\widetilde{\Sigma}_pX := \{pq | q \in X - \{p\}\}/\sim,$$

where the equivalence relation $\sim$ is defined such that $pq \sim pr$ iff $pq \subset pr$ or $pq \supset pr$. Then the *space of directions* Σ_pX at p in X is defined to be the metric completion of $\widetilde{\Sigma}_pX$. For $u \in \Sigma_pX$ we denote by γ_u the geodesic tangent to u with $\gamma_u(0) = p$ if it exists. For a function $f : X \longrightarrow R$, we define the directional derivative $\widetilde{d}_pf : \widetilde{\Sigma}_pX \longrightarrow R$ at p by $\widetilde{d}_pf(u) := (f \circ \gamma_u)'_+(0)$ if the right-hand derivative exists. We denote by $d_pf : \Sigma_pX \longrightarrow R$ the continuous extension of $\widetilde{d}_pf$ if it exists. Remarking that Σ_pX is an Alexandrov space with $curv \geq 1$, we consider the composition $\widetilde{d}_{u_1} \circ d_p$ of the two operators $\widetilde{d}_{u_1}$ and d_p for $p \in X$ and $u_1 \in \Sigma_pX$. We put $\widetilde{d}f_{p,u_1} := \widetilde{d}_{u_1} \circ d_pf$ and denote by $df_{p,u_1} : \Sigma_{u_1}\Sigma_pX \longrightarrow R$ the cotinuous extension of $\widetilde{d}f_{p,u_1} : \widetilde{\Sigma}_{u_1}\Sigma_pX \longrightarrow R$. Repeating the procedure, we define for k with $1 \leq k < n = \dim X$,

$$\Sigma^kX := \{(p, u_1, u_2, \ldots, u_k) |\ p \in X, u_1 \in \Sigma_pX, \\ u_i \in \Sigma_{u_{i-1}}\Sigma_{u_{i-2}} \cdots \Sigma_{u_1}\Sigma_pX\ (i = 2, \ldots, k)\},$$

$$\widetilde{d}f_{p,u_1,u_2,\ldots,u_k} := \widetilde{d}_{u_k} \circ \cdots \circ d_{u_2} \circ d_{u_1} \circ d_pf : \widetilde{\Sigma}_{u_k}\Sigma_{u_{k-1}} \cdots \Sigma_{u_1}\Sigma_pX \longrightarrow R$$

for a function $f : X \longrightarrow R$ and for $(p, u_1, u_2, \ldots, u_k) \in \Sigma^kX$.

Definition 0.1. A function $f : X \longrightarrow R$ belongs to the *class* D^r $(1 \leq r \leq n = \dim X)$, or simply, *$f$ is of D^r class* iff $\widetilde{d}f_{p,u_1,u_2,\ldots,u_{k-1}}$ is defined and has the continuous extension

$$df_{p,u_1,u_2,\ldots,u_{k-1}} : \Sigma_{u_{k-1}}\Sigma_{u_{k-2}} \cdots \Sigma_{u_1}\Sigma_pX \longrightarrow R$$

for all $(p, u_1, u_2, \ldots, u_{k-1}) \in \Sigma^{k-1}X$ and for all k with $1 \leq k \leq r$. We agree that a function $f : X \longrightarrow R$ is of D^1 class if and only if $\widetilde{d}f_p$ has the continuous extension df_p.

To control the behavior of the directional derivatives, we introduce a quantity associated with $f : X \longrightarrow R$ as follows:

$$\Delta_1 f(p) := \frac{1}{\mathcal{H}^{n-1}(\Sigma_pX)} \int_{\Sigma_pX \ni u} df_p(u)\, d\mathcal{H}^{n-1}(u), \quad \text{for } p \in X,$$

where $\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure on Σ_pX. This means the total flow of the gradient of f at the point p. If f is a diffenrentiable function on a Riemannian manifold, then f is of D^n class and $\Delta_1 f(p) = 0$ at every point p. On the other hand, there exists an Alexandrov space on which $\Delta_1 f(p) \neq 0$ at some singular point p for an affine function f of D^3 class and the whole space does not split (see Example 1.4). To avoid this case, we need the following definition.

Definition 0.2. A function $f : X \longrightarrow R$ is *of $D^{r,s}$ class*, $s \leq r$, iff f is of D^r class and

$$\Delta_1 df_{p,u_1,u_2,\ldots,u_{k-2}}(u_{k-1}) = 0$$

for all $(p, u_1, u_2, \ldots, u_{k-1}) \in \Sigma^{k-1}X$ and for all k with $1 \leq k \leq s$.

With these definitions, we now state our main theorem of this paper:

Theorem A. *Let X be an n-dimensional Alexandrov space with curv $\geq -\kappa^2$. Then, X admits a nontrivial affine function $\varphi : X \longrightarrow R$ of $D^{2,2}$ class if and only if X is isometric to the metric product $\widetilde{X} \times R$, where $\widetilde{X}$ is an $(n-1)$-dimensional Alexandrov space with curv $\geq -\kappa^2$.*

Since every constant function on X is an affine function of $D^{2,2}$ class, the dimension of the space of all affine functions on X of $D^{2,2}$ class is at least one. Thus we obtain the following corollary:

Corollary B. *The linear space of all affine functions on X of $D^{2,2}$ class is of dimension $k+1$ if and only if X is isometric to $\widetilde{X} \times R^k$, where $\widetilde{X}$ does not admit any nontrivial affine function of $D^{2,2}$ class.*

For the proof of Theorem A, it suffices to show that, for each minimal geodesic γ in X, there exists a totally geodesic and flat strip including γ. Under the assumption that X admits an affine function of $D^{2,2}$ class, we will show in Proposition 4.3 that the strip is spanned by the gradient curves of φ. Recently, G. Perelman and A. Petrunin [**10**] considered the existence of gradient curves in more general situation. The arguments in this paper is more elementary than theirs, and the author believes that his arguments will be shortend by their existence theorem.

1. Preliminaries and examples.

Throughout this paper, let X be an Alexandrov space with *curv* $\geq -\kappa^2$ for $\kappa > 0$.

1.0. Global Comparison Theorem. The most basic tool in Alexandrov geometry is the following theorem.

Fact 1.0 (Global Comparison Theorem)**.** (See [**1**, §3], [**5**, Theorem 1.1] and [**4**, Appendix].) If Z is an n-dimensional Alexandrov space, $n \geq 2$, with *curv* $\geq k$, then the following holds:

(i) For any triple (p_0, p_1, p_2) in Z there is a unique (up to isometry) triple $(\bar{p}_0, \bar{p}_1, \bar{p}_2)$ in $M^2(k)$ with $|p_i, p_j| = |\bar{p}_i, \bar{p}_j|$. Moreover, for any segment p_1p_2 and $0 \leq t \leq |p_1, p_2|$

(a) $$|p_0, p_1p_2(t)| \geq |\bar{p}_0, \bar{p}_1\bar{p}_2(t)|.$$

(If $k > 0$ we must also assume that $|p_0, p_1| + |p_1, p_2| + |p_2, p_0| < 2\pi/\sqrt{k}$.)

(ii) If equality holds in (a) for some $0 < t_0 < |p_1, p_2|$ and c_{t_0} is a segment from p_0 to $p_1p_2(t_0)$, then $c_{t_0}(s)$, $0 < s < |p_0, p_1p_2(t_0)|$, is joined their limit segments from p_0 to p_1 and p_2, form a surface which has totally geodesic interior and which is isometric to the triangular surface in $M^2(k)$ with vertices $\bar{p}_0, \bar{p}_1, \bar{p}_2$.

(iii) For any hinge (p_0p_1, p_0p_2) in Z with $0 < \angle(p_0p_1, p_0p_2) < \pi$ we have

$$|p_1, p_2| \leq |\bar{p}_1, \bar{p}_2|, \tag{b}$$

where $(\bar{p}_0\bar{p}_1, \bar{p}_0\bar{p}_2)$ is the corresponing hinge in $M^2(k)$.

(iv) If equality holds in (b), then (p_0p_1, p_0p_2) spans a surface which has totally geodesic interior and which is isometric to the triangular surface in $M^2(k)$ spanned by $(\bar{p}_0\bar{p}_1, \bar{p}_0\bar{p}_2)$. In fact, any such surface is determined uniquely by a segment in Z between interior points of p_0p_1 and p_0p_2.

1.1. Drectional derivative Df and the tangent cone. We denote by $K(\,\cdot\,)$ the Euclidean cone over a metric space (see [**1**, §4] for the definition of the Euclidean cone). The following fact is well-known:

Fact 1.1. The pointed Hausdorff limit $\lim_{\varepsilon\to 0}(\varepsilon^{-1}X, p)$ of the (ε^{-1})-scaling of the metric around p is isometric to the Euclidean cone $K(\Sigma_pX)$ for every $p \in X$.

We set $K_pX := K(\Sigma_pX)$ and call it the *tangent cone* at p in X. Let p^* denote the vertex of K_pX and αu, for $\alpha \geq 0$ and $u \in \Sigma_pX$, the point in K_pX such that $|\alpha u, p^*| = \alpha$ and $pr(\alpha u) = u$, where $pr : K_pX\backslash\{p^*\} \longrightarrow \Sigma_pX$ is the projection.

Let $f : X \longrightarrow R$ be a function of D^r class and $1 \leq k \leq r$. Then we obtain the extension

$$Df_{p,u_1,u_2,\ldots,u_{k-1}} : K_{u_{k-1}}K_{u_{k-2}}\cdots K_{u_1}K_pX \longrightarrow R$$

of $df_{p,u_1,u_2,\ldots,u_{k-1}} : \Sigma_{u_{k-1}}\Sigma_{u_{k-2}}\cdots\Sigma_{u_1}\Sigma_pX \longrightarrow R$ with the condition described as follows. We see that $K_{u_{k-1}}K_{u_{k-2}}\cdots K_{u_1}K_pX$ splits isometrically into the product

$$K(\Sigma_{u_{k-1}}\Sigma_{u_{k-2}}\cdots\Sigma_{u_1}\Sigma_pX) \times \langle u_1, u_2, \ldots, u_{k-1}\rangle$$

for every $(p, u_1, u_2, \ldots, u_{k-1}) \in \Sigma^{k-1}X$, where $\langle\cdot\rangle$ denotes the linear span. Under this identification, we have for $u = (\alpha_k u_k,\ \alpha_{k-1}u_{k-1} + \alpha_{k-2}u_{k-2} + \cdots + \alpha_1u_1) \in K(\Sigma_{u_{k-1}}\Sigma_{u_{k-2}}\cdots\Sigma_{u_1}\Sigma_pX) \times \langle u_1, u_2, \ldots, u_{k-1}\rangle$

$$Df_{p,u_1,u_2,\ldots,u_{k-1}}(u) = \sqrt{\alpha_k^2 df_{p,u_1,u_2,\ldots,u_{k-1}}(u_k)^2 + \alpha_{k-1}^2 + \cdots + \alpha_1^2}. \tag{†}$$

In particular for $k = 1$ in (†), we agree that $Df_p(\alpha_1u_1) = \alpha_1 df_p(u_1)$ for all $\alpha_1u_1 \in K_pX$ and $u_1 \in \Sigma_pX$.

1.2. Generalized gradient and gradient curves. Let $f : X \longrightarrow R$ be a function of D^1 class. Then by the compactness of $\Sigma_p X$, $df_p : \Sigma_p X \longrightarrow R$ attains the maximum and minimum values on $\Sigma_p X$ for each $p \in X$. We denote by $M(df_p)$ the maximum level set of df_p and by $m(df_p)$ the minimum level set of df_p. If $M(df_p)$ consists of only one element, we express it by $\widehat{\nabla} f(p)$ and put $\nabla f(p) := |\nabla f|(p) \cdot \widehat{\nabla} f(p) \in K_p X$, where $|\nabla f|(p) := \max_{u \in \Sigma_p X} df_p(u)$. We call $\nabla f(p)$ the *generalized gradient of f at p*. In this notation, if $m(df_p)$ consists of only one element, $\widehat{\nabla}(-f)(p)$ coincides with the element of $m(df_p)$.

A curve $c : [a, b] \longrightarrow X$ by definition has the right (left) tangent direction $v \in \Sigma_{c(t)}$ at $t \in [a, b)$ (resp. $t \in (a, b]$) if any initial direction of any minimal segment from $c(t + h)$ to $c(t)$ converges to v as $h \downarrow 0$ (resp. $h \uparrow 0$). The *gradient curve* $c : [a, b] \longrightarrow X$ of a function g on X is defined such that c has the right tangent $\widehat{\nabla} g(c(t))$ for every $t \in [a, b)$.

Example 1.3. Let $\widetilde{X}$ be an Alexandrov space with $curv \geq -\kappa^2$. Then the metric product $X := \widetilde{X} \times R$ is an Alexandrov space with $curv \geq -\kappa^2$. Define $\eta : X \longrightarrow R$ by $\eta((\widetilde{p}, t)) := t$. Then η is a nontrivial affine function of $D^{2,2}$ class (see Proposition 3.2). Then $\Sigma_p X$ is the spherical suspension of $\Sigma_p \widetilde{X}$ with its suspension points $\nabla \eta(p)$ $(= \widehat{\nabla} \eta(p))$ and $\nabla(-\eta)(p)$ $(= \widehat{\nabla}(-\eta)(p))$ for every $p \in X$. If $\widetilde{X}$ has singular points, then so does $X = \widetilde{X} \times R$.

Example 1.4. Let C be an unbounded convex body in R^n with nonempty interior and with boundary. Then C is a noncompact n-dimensional Alexandrov space with $curv \geq 0$ (with boundary). We take an arbitrary unit vector $z \in R^n$ and denote by $h_z : C \longrightarrow R$ the height function in the direction z, i.e., $h_z(p) := \langle z, p \rangle$, where $\langle\ ,\ \rangle$ is the cannonical inner product in R^n. Then h_z is affine. If there is a point on the boundary of C such that the diameter of Σ_p is less than π, then C does not split into the product of a line and a space, and then h_z is of D^n class but not of $D^{n,1}$ class.

2. Affine functions of D^1 class.

Throughout this section we assume that $\varphi : X \longrightarrow R$ is an affine function of D^1 class. We first prove the following lemma, which will frequently be used in this paper.

Lemma 2.1. *Fix an arbitrary point $p \in X$. Then $D\varphi_p : K_p X \longrightarrow R$ becomes an affine function again. In other words, we have*

$$(*) \qquad (\sin |u, v|)\, d\varphi_p(\sigma(t)) = \sin(|u, v| - t)\, d\varphi_p(u) + \sin t\, d\varphi_p(v)$$

for all $u, v \in \Sigma_p X$, for every minimal geodesic $\sigma : [0, |u, v|] \longrightarrow \Sigma_p X$ from u to v and for every $t \in [0, |u, v|]$.

Differentiating $(*)$ in t at $t = 0$ yields the following:

Corollary 2.2. *We have the directional derivative of second order at* $(p, u) \in \Sigma X$,

$$(**) \qquad (\sin|u,v|)\,\widetilde{d}\varphi_{p,u}(\dot{\sigma}(0)) = d\varphi_p(v) - d\varphi_p(u)\cos|u,v|,$$

where $\dot{\sigma}(0)$ *is the initial direction of* σ.

Proof of Lemma 2.1. From the continuity of $d\varphi_p$, it suffices to show $(*)$ for all $u, v \in \widetilde{\Sigma}_p X$ with $0 < |u,v| < \pi$ and for $t \in (0, |u,v|)$. Identifying u, v with two unit vectors in R^2 which makes angle $|u,v|_{\Sigma_p X}$, we define a number $\lambda = \lambda(t) \in (0,1)$ such that $\angle(u, (1-\lambda)u + \lambda v) = t$. Let γ_u and γ_v be the geodesics tangent to u and v respectively and $\{s_i\}$ a sequence of numbers such that $s_i \searrow 0$ as $i \to \infty$. We can choose an appropriate subsequence $\{s_j\} \subset \{s_i\}$ so that a sequence $\{\tau_j : [0,1] \longrightarrow (1/s_j)X\}$ of minimal geodesics in $(1/s_j)X$ from $\gamma_u(s_j)$ to $\gamma_v(s_j)$ tends to a segment τ on K_pX as $j \to \infty$. Here each τ_j is parameterized proportionally to arclength, and the segment τ is projected to a minimal geodesic $\sigma : [0, |u,v|] \longrightarrow \Sigma_p X$ from u to v.

Let $\alpha_j : [0, |p, \tau_j(\lambda)|] \longrightarrow X$ be a minimal geodesic from p to $\tau_j(\lambda)$. By the continuity of $d\varphi_p$ and Fact 1.1, we have $d\varphi_p(\dot{\alpha}_j(0)) \to d\varphi_p(\sigma(t))$ as $j \to \infty$. Using Fact 1.1 and the definition of affine functions, we obtain

$$d\varphi_p(\sigma(t)) = \lim_{j\to\infty} d\varphi_p(\dot{\alpha}_j(0)) = [(1-\lambda)d\varphi_p(u) + \lambda d\varphi_p(v)]\frac{\sin t}{\lambda \sin|u,v|}.$$

In elementary Euclidean geometry, we have $(1-\lambda)/\lambda = \sin(|u,v| - t)/\sin t$. Thus we obtain $(*)$.

For the first assertion, we need to prove that $d\varphi_p(u) = -d\varphi_p(v)$ for $u, v \in \Sigma_p X$ with $|u,v| = \pi$. This is obvious from the continuity of $d\varphi_p$ and $(*)$. Hence this completes the proof. □

Lemma 2.3. *If* $\max_{u\in\Sigma_p X} d\varphi_p(u) > 0$ $(\min_{u\in\Sigma_p X} d\varphi_p(u) < 0)$ *at some point* $p \in X$, *then the maximum level set* $M(d\varphi_p)$ *of* $d\varphi_p$ (*resp. the minimum level set* $m(d\varphi_p)$ *of* $d\varphi_p$) *consists of only one element. In particular, the generalized gradient* $\nabla\varphi(p)$ (*resp.* $\nabla(-\varphi)(p)$) *is defined for all* $p \in X$ *with* $\varphi(p) < \sup_X \varphi$ (*resp.* $\varphi(p) > \inf_X \varphi$).

Proof. We prove this lemma only in the case $\max_{u\in\Sigma_p X} d\varphi_p(u) > 0$. Suppose that $M(d\varphi_p)$ contains two elements u_1 and u_2 under the assumption $\max d\varphi_p > 0$. If $|u_1, u_2| = \pi$, then we have $d\varphi_p(u_1) = -d\varphi_p(u_2)$ by Lemma 2.1. Hence $d\varphi_p(u_1) = \max d\varphi_p = 0$, a contradiction to the assumption. Otherwise, if $|u_1, u_2| \neq \pi$, we have, for $t = |u_1, u_2|/2$ in $(*)$ along some minimal geodesic $\sigma : [0, |u_1, u_2|] \longrightarrow \Sigma_p X$ from u_1 to u_2, $d\varphi_p(u_1) = \max d\varphi_p < d\varphi_p(\sigma(|u_1, u_2|/2))$. This contradicts the choice of u_1. □

We discuss the zero level set $(d\varphi_p)^{-1}(0)$ of $d\varphi_p$. For simplicity, we put

$$O(d\varphi_p) := (d\varphi_p)^{-1}(0).$$

If $o_1, o_2 \in O(d\varphi_p)$ satisfy $0 < |o_1, o_2| < \pi$, then Lemma 2.1 implies that all minimal geodesics from o_1 to o_2 are contained entirely in $O(d\varphi_p)$. Thus we obtain the following lemma:

Lemma 2.4. *$O(d\varphi_p)$ is locally convex.*

We deal with a point $p \in X$ such that $\varphi(p) < \sup_X \varphi$, or equivalently, $\max d\varphi_p > 0$. Then by Lemma 2.3, the generalized gradient $\nabla\varphi(p)$ is defined. For such point p, we investigate $\widehat{\nabla}\varphi(p)$ and $O(d\varphi_p)$ in the following proposition:

Proposition 2.5. *Let p be a point of X with $\varphi(p) < \sup_X \varphi$. Then the following* (i) *and* (ii) *hold:*

(i) *For every $o \in O(d\varphi_p)$, we have $|\widehat{\nabla}\varphi(p), o| \le \pi/2$.*
(ii) *If there are elements o_1 and o_2 of $O(d\varphi_p)$ such that $0 < |o_1, o_2| < \pi$ and $|\widehat{\nabla}\varphi(p), o_1| = |\widehat{\nabla}\varphi(p), o_2| = \pi/2$, then the triple $(\widehat{\nabla}\varphi(p), o_1, o_2)$ spans a totally geodesic triangular surface of constant curvature* 1.

Remark 2.6. Since $-\varphi$ is also affine, the same assertions as above hold for $p \in X$ with $\varphi(p) > \inf_X \varphi$ and for $\widehat{\nabla}(-\varphi)(p)$ instead of $\widehat{\nabla}\varphi(p)$.

Proof of Proposition 2.5. (i) We want to use $(**)$ in Corollary 2.2 for $u = \widehat{\nabla}\varphi(p)$ and $v = o \in O(d\varphi_p)$. If $|\widehat{\nabla}\varphi(p), o| = \pi$, then it follows from Lemma 2.1 that $\max_{u\in\Sigma_p X} d\varphi_p(u) = d\varphi_p(\widehat{\nabla}\varphi(p)) = -d\varphi_p(o) = 0$, a contradiction. Thus $|\widehat{\nabla}\varphi(p), o| \ne \pi$. Applying $(**)$ to $\widehat{\nabla}\varphi(p)$ and o along some minimal geodesic from $\widehat{\nabla}\varphi(p)$ to o and using $\widetilde{d\varphi}_{p,\widehat{\nabla}\varphi(p)} \le 0$, we have

$$d\varphi_p(o) - |\nabla\varphi|(p)\cos|\widehat{\nabla}\varphi(p), o| \le 0.$$

Hence $|\widehat{\nabla}\varphi(p), o| \le \pi/2$.

(ii) Draw a comparison triangle $\triangle(\overline{\widehat{\nabla}\varphi(p)}, \bar{o}_1, \bar{o}_2)$ in the unit sphere $S^2(1)$ corresponding to a geodesic triangle $\triangle(\widehat{\nabla}\varphi(p), o_1, o_2)$ in $\Sigma_p X$. Take a point o_3 in the interior of the edge o_1o_2 and a point $\bar{o}_3$ in the edge $\bar{o}_1\bar{o}_2$ corresponding to o_3. Then the Global Comparison Theorem implies that $|\widehat{\nabla}\varphi(p), o_3| \ge |\overline{\widehat{\nabla}\varphi(p)}, \bar{o}_3| = \pi/2$ (see Fact 1.0). Since $o_3 \in O(d\varphi_p)$ by Lemma 2.4, it follows from (i) that $|\widehat{\nabla}\varphi(p), o_3| \le \pi/2$. Hence $|\widehat{\nabla}\varphi(p), o_3| = |\overline{\widehat{\nabla}\varphi(p)}, \bar{o}_3|$. Therefore (ii) follows from the rigidity of geodesic triangle in the Global Comparison Theorem. □

3. Affine functions of $D^{2,2}$ class.

Proposition 3.1. *If $\varphi : X \longrightarrow R$ is a nontrivial affine function of $D^{2,2}$ class, then the following* (i)-(v) *hold:*

(i) *$\Sigma_p X$ is the spherical suspension of $O(d\varphi_p)$ with its suspension points $\widehat{\nabla}\varphi(p)$ and $\widehat{\nabla}(-\varphi)(p)$ for every $p \in X$.*

(ii) *We have*

$$d\varphi_p(u) = |\nabla\varphi|(p)\cos|\widehat{\nabla}\varphi(p), u|_{\Sigma_p X}$$

for every $p \in X$ and for every $u \in \Sigma_p X$, or equivalentlly,

$$D\varphi_p(u) = |\nabla\varphi|(p)|u|\cos\angle_{p*}(\nabla\varphi(p), u)$$

for every $u \in K_p X$. Here $\angle_{p*}$ denotes the angle distance at the vertex $p*$.

(iii) *Fix two arbitrary numbers $a, b \in \varphi(X)$ with $a < b$. Then for every $q \in \varphi^{-1}(a)$ and for every minimal geodesic $\sigma_q : [0, l(q)] \longrightarrow X$ from q to $\varphi^{-1}(b)$, we have*

$$\dot{\sigma}_q(0) = \widehat{\nabla}\varphi(q) \quad \text{and} \quad (\dot{\sigma_q^{-1}})(0) = \widehat{\nabla}(-\varphi)(\sigma_q^{-1}(0)),$$

where $\sigma_q^{-1} : [0, l(q)] \longrightarrow X$ is defined by $\sigma_q^{-1}(t) := \sigma_q(l(q) - t)$, $t \in [0, l(q)]$.

(iv) *There is a unique complete gradient curve $\phi : R \longrightarrow X$ of φ passing through p parameterized by $\varphi \circ \phi(t) = \varphi(p) + t$, $t \in R$, for every $p \in X$. Moreover, it satisfies*

$$\frac{|\varphi(\phi(t_1)) - \varphi(\phi(t_2))|}{|\phi(t_1), \phi(t_2)|} = |\nabla\varphi|(\phi(t)) = |\nabla\varphi|(p)$$

for all $t, t_1, t_2 \in R$, and in partucular ϕ is a straight line.

(v) *$|\nabla\varphi|(p)$ is constant for all $p \in X$.*

Proof. (i) Fix a point $p \in X$ arbitrarily. Since φ is nontrivial, either $\max d\varphi_p > 0$ or $\min d\varphi_p < 0$ holds. If $\max d\varphi_p > 0$, then $\min d\varphi_p < 0$ since φ is particularly of $D^{2,1}$ class. Similarly, we have $\max d\varphi_p > 0$ if $\min d\varphi_p < 0$. Thus $\widehat{\nabla}\varphi(p)$ and $\widehat{\nabla}(-\varphi)(p)$ at p are defined.

By Proposition 2.5 (ii), it suffices to show the following:

Assertion. $|\widehat{\nabla}\varphi(p), o|_{\Sigma_p X} = |\widehat{\nabla}(-\varphi)(p), o|_{\Sigma_p X} = \pi/2$ for all $o \in O(d\varphi_p)$.

Suppose that there is $o \in O(d\varphi_p)$ such that $|\widehat{\nabla}\varphi(p), o| \neq \pi/2$. Then by Proposition 2.5 (i), $|\widehat{\nabla}\varphi(p), o| < \pi/2$. Apply $(**)$ to $\widehat{\nabla}\varphi(p)$ and o along a minimal geodesic $\sigma : [0, |\widehat{\nabla}\varphi(p), o|] \longrightarrow \Sigma_p X$ from $\widehat{\nabla}\varphi(p)$ to o. Then we have $d\varphi_{p,\widehat{\nabla}\varphi(p)}(\dot{\sigma}(0)) < 0$. Since φ is particularly of D^2 class, there is a neighborhood W of $\dot{\sigma}(0)$ in $\Sigma_{\widehat{\nabla}\varphi(p)}\Sigma_p X$ such that $d\varphi_{p,\widehat{\nabla}\varphi(p)}(w) < 0$ for every

$w \in \overline{W}$. Since $d\varphi_{p,\widehat{\nabla}\varphi(p)} \leq 0$ and $\overline{W}$ is of positive measure with respect to $(n-2)$-dimensional Hausdorff measure, we obtain

$$\int_{\Sigma_{\widehat{\nabla}\varphi(p)}\Sigma_p X \ni w} d\varphi_{p,\widehat{\nabla}\varphi(p)}(w)\, d\mathcal{H}^{n-2}(w) < 0.$$

This contradicts the assumption that φ is of $D^{2,2}$ class.

(ii) It suffices to show the equation for $u \in \Sigma_p X \backslash \{\widehat{\nabla}\varphi(p), \widehat{\nabla}(-\varphi)(p)\}$. By (i), any such u is contained in a minimal geodesic $\tau : [0,\pi] \longrightarrow \Sigma_p X$ joining two suspension points. Applying $(**)$ to $\widehat{\nabla}\varphi(p)$ and $\tau(\pi/2)$ along τ, we have $d\varphi_{p,\widehat{\nabla}\varphi(p)}(\dot{\tau}(0)) = 0$. Using again $(**)$ for $\widehat{\nabla}\varphi(p)$ and u along τ, we obtain $d\varphi_p(u) = |\nabla\varphi|(p)\cos|\widehat{\nabla}\varphi(p), u|$.

(iii) Suppose that $\dot{\sigma}_q(0) \neq \nabla\varphi(q)$ for some minimal geodesic $\sigma_q : [0, l(q)] \longrightarrow X$ from q to $\varphi^{-1}(b)$. Then we can find a broken geodesic

$$\xi = \bigcup_i \gamma_i : [0, l(\xi)] \longrightarrow X$$

such that $(\varphi \circ \xi)'_+(s) > d\varphi_q(\dot{\sigma}_q(0))$ for every $s \in [0, l(\xi))$ and $\xi(0) = q$, $\xi(l(\xi)) \in \varphi^{-1}(b)$. The construction of ξ is achieved in the same way as in 2-dimensional Alexandrov space (see [**7**, Lemma 2(2)]). Since $\varphi \circ \xi$ is almost everywhere differentiable, we conclude that $l(q) > l(\xi)$. This contradicts the minimizing property of σ_q.

(iv) Choose a double-ended sequence $\{a_j\}_{j\in Z}$ such that $a_0 = \varphi(p)$, $a_j \nearrow \sup_X \varphi$ as $j \to \infty$ and $a_j \searrow \inf_X \varphi$ as $j \to -\infty$. We start from $p \in \varphi^{-1}(a_0)$ and repeat the same construction by minimal projections as in (iii). That is, let $p_0 := p$ and p_{j+1} denote the foot of the (unique) minimal geodesic from p_j to $\varphi^{-1}(a_{j+1})$ for $j \geq 0$. For $j \leq 0$, let p_{j-1} denote the foot of the (unique) minimal geodesic from p_j to $\varphi^{-1}(a_{j-1})$. Then we obtain the curve

$$\phi := \bigcup_{j\in Z} p_j p_{j+1} : \left(\inf_X \varphi - \varphi(p), \sup_X \varphi - \varphi(p)\right) \longrightarrow X$$

parameterized by $\varphi \circ \phi(t) = \varphi(p) + t$. By the construction, we see that every subarc from $\phi(t_1)$ to $\phi(t_2)$ of ϕ is a minimal geodesic and

$$\frac{|\varphi(\phi(t_1)) - \varphi(\phi(t_2))|}{|\phi(t_1), \phi(t_2)|} = |\nabla\varphi|(\phi(t)) = |\nabla\varphi|(p)$$

for all $t, t_1, t_2 \in (\inf_X \varphi - \varphi(p), \sup_X \varphi - \varphi(p))$.

Once $\sup_X \varphi = \infty$ and $\inf_X \varphi = -\infty$ are established, the proof of (iv) is completed. Suppose that $\sup_X \varphi < \infty$. Then the sequence $\{p_j\}_{j=0,1,\ldots}$ accumulates to some point p_∞. (i) implies that $\max d\varphi_{p_\infty} > 0$ also at p_∞. This is a contradiction to $\varphi(p_\infty) = \sup_X \varphi$. Therefore $\sup_X \varphi = \infty$. On the other hand, $\inf_X \varphi = -\infty$ follows from which $-\varphi$ is also affine.

(v) Choose two points p_0 and p_1 arbitrarily. Let $\phi_0, \phi_1 : R \longrightarrow X$ be two gradient curves passing through p_0, p_1 respectively obtained by (iv). We may assume from (iv) that $\varphi(p_0) = \varphi(p_1)$ and $\phi_0 \neq \phi_1$. It is easily seen from (i) and (iv) that $\angle_{p_1}(p_1p_0, p_1\phi_1(t)) = \pi/2$, $|p_1, \phi_1(t)| = t/|\nabla\varphi|(p_1)$ and $|\nabla\varphi|(p_0) \geq (\varphi(\phi_1(t)) - \varphi(p_0))/|p_0, \phi_1(t)| = t/|p_0, \phi_1(t)|$ for all $t \in [0, \infty)$.

For every $t \in [0, \infty)$, draw a comparison hinge $(\bar{p}_1\bar{p}_0, \bar{p}_1\overline{\phi_1(t)})$ of a hinge $(p_1p_0, p_1\phi_1(t))$ in hyperbolic surface $H^2(-\kappa^2)$ of constant curvature $-\kappa^2$. Then the Global Comparison Theorem and the cosine formula in $H^2(-\kappa^2)$ implies that

$$|p_0, \phi_1(t)| \leq |\bar{p}_0, \overline{\phi_1(t)}| = \frac{1}{\kappa}\cosh^{-1}\Big[\cosh(\kappa|p_0, p_1|)\cosh(\kappa|p_1, \phi_1(t)|)\Big].$$

Therefore we have

$$|\nabla\varphi|(p_0) \geq \frac{\kappa \cdot t}{\cosh^{-1}\Big[\cosh(\kappa|p_0, p_1|)\cosh(\kappa \cdot t/|\nabla\varphi|(p_1))\Big]}.$$

Taking $t \to \infty$ and applying L'Hospital's formula, we obtain $|\nabla\varphi|(p_0) \geq |\nabla\varphi|(p_1)$. The symmetric property of the above discussion implies the reverse inequality. This completes the proof. □

We now assume that $\varphi : X \longrightarrow R$ is a nontrivial affine function of D^1 class satisfying the condition of the assertion (i) of Proposition 3.1. Then all other assertions (ii)-(v) follow. More precisely, the following holds:

Proposition 3.2. *Let $\varphi : X \longrightarrow R$ be a nontrivial affine function of D^1 class. If φ satisfies the condition that Σ_pX forms the spherical suspension with its suspension points $\widehat{\nabla}\varphi(p)$ and $\widehat{\nabla}(-\varphi)(p)$ for every $p \in X$, then φ is of $D^{n,n}$ class, $n = \dim X$.*

Remark 3.3. It is easily seen that η in Example 1.3 is of D^1 class and satisfies the assumption of Proposition 3.2. Hence η is of $D^{2,2}$ class.

Proof of Proposition 3.2. We first prove that φ is of $D^{1,1}$ class. Fix a point $p \in X$ arbitrarily. Since Σ_pX is a spherical suspension with its suspension points $\widehat{\nabla}\varphi(p)$ and $\widehat{\nabla}(-\varphi)(p)$, there is a unique point $\underline{u} \in \Sigma_pX$ for every $u \in \Sigma_pX$ such that $|\widehat{\nabla}\varphi(p), u| = |\underline{u}, \widehat{\nabla}(-\varphi)(p)|$ and that u, $\underline{u}$ are lying on a common minimal geodesic joining suspension points. The correspondence $u \mapsto \underline{u}$ is a isometry between $d\varphi_p^+ := \{v \in \Sigma_pX | d\varphi_p(v) \geq 0\}$ and $d\varphi_p^- := \{v \in \Sigma_pX | d\varphi_p(v) \leq 0\}$. Note that Proposition 3.1 (ii) is valid under the assumption of Proposition 3.2. Hence by Proposition 3.1 (ii), we have

$d\varphi_p(\underline{u}) = -d\varphi_p(u)$ for every $u \in \Sigma_p X$. A direct computation implies that

$$\int_{\Sigma_p X \ni u} d\varphi_p(u)\, d\mathcal{H}^{n-1}(u)$$
$$= \int_{d\varphi_p^+ \ni u} d\varphi_p(u)\, d\mathcal{H}^{n-1}(u) + \int_{d\varphi_p^- \ni \underline{u}} d\varphi_p(\underline{u})\, d\mathcal{H}^{n-1}(\underline{u}) = 0.$$

That is, φ is of $D^{1,1}$ class.

We next show that φ is of $D^{2,2}$ class. For every $w \in \Sigma_{\widehat{\nabla}\varphi(p)}\Sigma_p X$, there is a minimal geodesic $\sigma_w : [0,\pi] \longrightarrow \Sigma_p X$ from $\widehat{\nabla}\varphi(p)$ to $\widehat{\nabla}(-\varphi)(p)$ tangent to w. Apply $(**)$ to $\widehat{\nabla}\varphi(p)$ and $\sigma_w(\pi/2)$ along σ_w. Then we have $d\varphi_{p,\widehat{\nabla}\varphi(p)}(w) = 0$ for all $w \in \Sigma_{\widehat{\nabla}\varphi(p)}\Sigma_p X$. Therefore we conclude that φ is of $D^{2,2}$ class at $(p, \widehat{\nabla}\varphi(p)) \in \Sigma X$ for all $p \in X$. Similarly, φ is of $D^{2,2}$ class at $(p, \widehat{\nabla}(-\varphi)(p)) \in \Sigma X$ for all $p \in X$.

Therefore it suffices to show the $D^{2,2}$ condition for all $u \in \Sigma_p X \backslash \{\widehat{\nabla}\varphi(p), \widehat{\nabla}(-\varphi)(p)\}$. Such u is contained in a minimal geodesic from $\widehat{\nabla}\varphi(p)$ to $\widehat{\nabla}(-\varphi)(p)$. Letting $d_{\widehat{\nabla}\varphi} : \Sigma_p X \longrightarrow R$ denote the distance function from $\widehat{\nabla}\varphi(p)$, we see that $\Sigma_u\Sigma_p X$ is also a spherical suspension with its suspension points $\nabla(-d_{\widehat{\nabla}\varphi})(u)$ and $\nabla d_{\widehat{\nabla}\varphi}(u)$. For every $w \in \widetilde{\Sigma}_u\Sigma_p X$ let $\sigma_w : [0, l(w)] \longrightarrow \Sigma_p X$ be a geodesic tangent to w and put $\theta_w := \angle(\nabla(-d_{\widehat{\nabla}\varphi}(u), w)$. Then by a direct computation, we have

$$\widetilde{d}\varphi_{p,u}(w) = \frac{d}{dt}(d\varphi_p(\sigma_w(t)))\big|_{t=0} = |\nabla\varphi|(p) \sin|\widehat{\nabla}\varphi(p), u| \cos\theta_w.$$

This means that $\widetilde{d}\varphi_{p,u}$ has the continuous extension $d\varphi_{p,u} : \Sigma_u\Sigma_p X \longrightarrow R$. Therefore φ is of $D^{2,1}$ class. Moreover, the $D^{2,2}$ condition at $(p,u) \in \Sigma X$ is implied by the same computation as in the proof of the $D^{1,1}$ condition of φ.

Repeating the above computation, we see that φ is of $D^{n,n}$ class. □

4. Totally geodesic flat strip spanned by gradient curves.

In this section we prove Theorem A. Throughout this section let $\varphi : X \longrightarrow R$ be a nontrivial affine function of $D^{2,2}$ class.

Let p_0 and p_1 be two arbitrary points with $\varphi(p_0) = \varphi(p_1) =: a$ and $\gamma : [0,1] \longrightarrow X$ a minimal geodesic from p_0 to p_1 parameterized to be proportional to arclength. By Proposition 3.1 (iv), there is a unique gradient curve $\phi_\lambda : R \longrightarrow X$ passing through $\gamma(\lambda)$ for every $\lambda \in [0,1]$.

We will prove in Proposition 4.3 that

$$S := \bigcup_{\lambda\in[0,1]} \phi_\lambda(R)$$

is totally geodesic and flat. Once this is established, Theorem A easily follows.

We first prove the following lemma:

Lemma 4.1. *Fix $\lambda_1, \lambda_2 \in [0,1]$ and $s_0 \in R$ arbitrarily. Setting $l(s,t) := |\phi_{\lambda_1}(s), \phi_{\lambda_2}(t)|$ for all $s, t \in R$, we have*

$$\lim_{h \to 0} \frac{l(s_0+h, s_0+h) - l(s_0, s_0)}{h} = 0.$$

Proof. Let $\tau_{st} : [0, l(s,t)] \longrightarrow X$ be a minimal geodesic from $\phi_{\lambda_1}(s)$ to $\phi_{\lambda_2}(t)$ and $\theta_i(s,t) := \angle(\tau_{st}, \nabla\varphi(\phi_{\lambda_i}))$, $i = 1, 2$. We note that $\theta_i(s,t)$, $i = 1, 2$, is independent of the choice of the minimal geodesic τ_{st}. Actually, Proposition 3.1 (ii) implies that $|\nabla\varphi| \cos\theta_1(s,t) = d\varphi_{\phi_{\lambda_1}(s)}(\dot{\tau}_{st}(0)) = d\varphi_{\phi_{\lambda_1}(s)}(\dot{\tau}'_{st}(0)) = |\nabla\varphi| \cos\theta'_1(s,t)$ for every other minimal geodesic τ'_{st} from $\phi_{\lambda_1}(s)$ to $\phi_{\lambda_2}(t)$ and for $\theta'_1(s,t) = \angle(\tau'_{st}, \nabla\varphi(\phi_{\lambda_1}))$. Similarly, we have $\theta'_2(s,t) = \theta_2(s,t)$ for $\theta'_2(s,t) := \angle(\tau'_{st}, \nabla\varphi(\phi_{\lambda_2}))$. Therefore it follows from the first variation formula ([**9**, Theorem 3.5]) that for all $s, t \in R$, the partial derivatives $\frac{\partial l}{\partial s}(s,t)$ and $\frac{\partial l}{\partial t}(s,t)$ exist and equal $(-1/|\nabla\varphi|)\cos\theta_1(s,t)$ and $(-1/|\nabla\varphi|)\cos\theta_2(s,t)$ respectively. Thus we obtain for every $h \in R$,

$$|l(s_0+h, s_0+h) - l(s_0+h, s_0)| = \frac{1}{|\nabla\varphi|} \left| \int_0^{|h|} \cos\theta_2(s_0+h, s_0+t)dt \right|$$

and

$$|l(s_0+h, s_0) - l(s_0, s_0)| = \frac{1}{|\nabla\varphi|} \left| \int_0^{|h|} \cos\theta_1(s_0+t, s_0)dt \right|.$$

Proposition 3.1 implies that for every $\varepsilon > 0$ there is $\delta = \delta(\varepsilon) > 0$ such that $|\theta_1(s,t) - \pi/2|, |\theta_2(s,t) - \pi/2| \leq \varepsilon$ for all $s, t \in [s_0 - \delta, s_0 + \delta]$. Therefore we have for all $h \in [-\delta, \delta]$,

$$|l(s_0+h, s_0+h) - l(s_0, s_0)| \leq \frac{2}{|\nabla\varphi|}|h| \left| \cos\left(\frac{\pi}{2} \pm \varepsilon\right) \right|.$$

Since $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$, we obtain the desired equality. □

Define $c_s : [0,1] \longrightarrow S$ by $c_s(\lambda) := \phi_\lambda(s)$ for an arbitrarily fixed $s \in R$. Then the following holds:

Corollary 4.2. *The curve c_s is minimal for every $s \in R$.*

Proof. Let $pr_a : X \longrightarrow \varphi^{-1}(a)$ be the minimal projection to $\varphi^{-1}(a)$ and τ a minimal geodesic from $\phi_0(s)$ to $\phi_1(s)$. Then Lemma 4.1 implies that

$$L(c_s) \geq L(\tau) = L(pr_a \circ \tau) \geq L(\gamma) = L(pr_a \circ c_s) = L(c_s),$$

where $L(\cdot)$ means the length of a curve. This completes the proof. □

The proof of Theorem A is completed by the following:

Proposition 4.3. *The strip S is totally geodesic and flat.*

Proof. Let y_0, y_1 be two points in S and $\tau : [0,1] \longrightarrow X$ a minimal geodesic from y_0 to y_1 parameterized proportionally to arclength. We may assume that $y_0 \in \phi_0(R)$ and $y_1 \in \phi_1(R)$. Moreover, we may assume from Corollary 4.2 that $\varphi(y_0) < \varphi(y_1)$. We define a curve $c : [0,1] \longrightarrow S$ in S by $c(\lambda) := \phi_\lambda((1-\lambda)\varphi(y_0) + \lambda\varphi(y_1))$. Then $\varphi \circ c$ is an affine function. We will calculate the length of c. We denote by $pr_{\phi_\lambda} : X \longrightarrow \phi_\lambda(R)$ the minimal projection to $\phi_\lambda(R)$ and put $y_1' := pr_{\phi_1}(y_0)$. Fix $\lambda \in (0,1)$ arbitrarily and let $h \in R\backslash\{0\}$ be a number such that $|h|$ is sufficiently small. We consider the triangle $\triangle(c(\lambda), c(\lambda+h), pr_{\phi_{\lambda+h}}(c(\lambda)))$ whose sides are all minimal if $h > 0$ (if $h < 0$ consider $\triangle(c(\lambda), c(\lambda+h), pr_{\phi_\lambda}(c(\lambda+h)))$). This is a right triangle on X. By Lemma 4.1 and the parameterization of c, we have for $h > 0$,

$$|c(\lambda), pr_{\phi_{\lambda+h}}(c(\lambda))| = L(\gamma)|h| \quad \text{and} \quad |pr_{\phi_{\lambda+h}}(c(\lambda)), c(\lambda+h)| = |y_1, y_1'||h|,$$

and for $h < 0$, $|c(\lambda), pr_{\phi_\lambda}(c(\lambda+h))| = L(\gamma)|h|$ and $|pr_{\phi_\lambda}(c(\lambda+h)), c(\lambda+h)| = |y_1, y_1'||h|$. Note that $c(\lambda)$ is a point of the straight line ϕ_λ. Fact 1.0 together with this imply that $|c(\lambda), c(\lambda+h)| = \sqrt{L(\gamma)^2|h|^2 + |y_1, y_1'|^2|h|^2} + o(|h|)$. Equivalently,

$$|\dot{c}|(\lambda) := \lim_{h\to 0} \frac{|c(\lambda), c(\lambda+h)|}{|h|} = \sqrt{L(\gamma)^2 + |y_1, y_1'|^2}$$

for all $\lambda \in (0,1)$. This implies that c is a Lipschitz curve. Therefore $L(c)$ is calculated as

$$L(c) = \int_0^1 |\dot{c}|(\lambda) d\lambda = \sqrt{L(\gamma)^2 + |y_1, y_1'|^2}.$$

A similar calculation shows that $L(pr_a \circ \tau) = \sqrt{L(\tau)^2 - |y_1, y_1'|^2}$, where pr_a is the same as in the proof of Corollary 4.2. Thus we obtain

$$L(c)^2 \geq L(\tau)^2 = L(pr_a \circ \tau)^2 + |y_1, y_1'|^2 \geq L(\gamma)^2 + |y_1, y_1'|^2 = L(c)^2.$$

This completes the proof. □

Acknowledgments. The author would like to express his thanks to Y. Otsu, K. Shiohama, T. Shioya and T. Yamaguchi for their valuable advice. The notion of functions of $D^{r,s}$ class is introduced by Y. Otsu.

References

[1] Yu.D. Burago, M. Gromov and G. Perelman, *A.D. Alexandrov's spaces with curvatures bounded below* I, Russian Math. Surveys, **47**(**2**) (1992), 1-58, MR 93m:53035.

[2] J. Cheeger and D. Gromoll, *The splitting theorem for manifolds of nonnegative Ricci curvature*, J. Diff. Geom., **61** (1971), 119-128, MR 46 #2597, Zbl 0223.53033.

[3] R.E. Greene and K. Shiohama, *The isometry groups of manifolds admitting non-constant convex functions*, J. Math. Soc. Japan, **39**(**1**) (1987), 1-16, MR 88a:53032, Zbl 0611.53039.

[4] K. Grove and S. Markvorsen, *New extremal problems for the Riemannian recognition program via Alexandrov geometry*, J. Amer. Math. Soc., **8** (1995), 1-28, MR 95j:53066, Zbl 0829.53033.

[5] K. Grove and F. Wilhelm, *Hard and soft packing radius theorems*, Ann. of Math., **142**(**2**) (1995), 213-237, MR 96h:53054, Zbl 0846.53042.

[6] N. Innami, *Splitting theorems of Riemannian manifolds*, Compositio Math., **47** (1982), 237-247, MR 84e:53053, Zbl 0514.53040.

[7] Y. Mashiko, *Affine functions on Alexandrov surfaces*, Osaka J. Math., **36** (1999), 853-859, MR 2001a:53057.

[8] A. Milka, *Metric structure of one class of spaces containing straight lines* (in Russian), Ukrain. Geom. Sbornik, **4** (1967), 43-48, MR 41 #983.

[9] Y. Otsu and T. Shioya, *The Riemannian structure of Alexandrov spaces*, J. Diff. Geom., **39** (1994), 629-658, MR 95e:53062, Zbl 0808.53061.

[10] G. Perelman and A. Petrunin, *Quasigeodesics and gradient curves in Alexandrov spaces*, preprint.

[11] T. Shioya, *Splitting theorems for nonnegatively curved open manifolds with large ideal boundary*, Math. Z., **212** (1993), 223-238, MR 93k:53036, Zbl 0791.53045.

[12] V.A. Toponogov, *Spaces with straight lines*, AMS Transl., **37** (1964), 278-280, MR 21 #7520, Zbl 0138.42902.

Received October 17, 2000.

Faculty of Science and Engineering
Saga University
Hojyoumachi 1
Saga 840-8502, Japan
E-mail address: mashiko@ms.saga-u.ac.jp

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

THE SECOND COHOMOLOGY OF SMALL IRREDUCIBLE MODULES FOR SIMPLE ALGEBRAIC GROUPS

George J. McNinch

Let G be a connected, simply connected, quasisimple algebraic group over an algebraically closed field of characteristic $p > 0$, and let V be a rational G-module such that $\dim V \leq p$. According to a result of Jantzen, V is completely reducible, and $H^1(G, V) = 0$. In this paper we show that $H^2(G, V) = 0$ unless some composition factor of V is a nontrivial Frobenius twist of the adjoint representation of G.

1. Introduction.

Let G be a quasisimple, connected, simply connected algebraic group over the algebraically closed field k of characteristic $p > 0$. By a G-module V, we always understand a rational G-module (one given by a morphism of algebraic groups $G \to \mathrm{GL}(V)$). In this paper, we study the cohomology of a G-module V such that $\dim V \leq p$. By results of Jantzen [**Jan96**] one knows that V is semisimple and that $H^1(G, V) = 0$.

Recall that the Lie algebra $\mathfrak{g}$ of G is a G-module via the adjoint action. Our main result is:

Theorem A. *Let V be a G-module with $\dim V \leq p$. Then $H^2(G, V) \neq 0$ if and only if V has a composition factor isomorphic with a* Frobenius twist $\mathfrak{g}^{[d]}$ *of $\mathfrak{g}$ for some $d \geq 1$.*

Differentiating the representation of G on V gives a representation for the Lie algebra $\mathfrak{g}$ on V. Assume that $V^{\mathfrak{g}} = 0$. Then the theorem says that $H^2(G, V) = 0$. For V of this sort, the vanishing of H^2 is a consequence of the linkage principle for G together with results in Section 2 which give estimates for the dimensions of Weyl modules whose high weights are simultaneously in the low alcove and in the orbit $W_p \bullet 0$. In fact, the same argument shows that $H^i(G, V)$ is 0 for all $i \geq 1$; see Proposition 5.2. It was pointed out to me that an earlier version of this manuscript contained an overly complicated proof of this observation.

The crucial case for Theorem A is when V is simple, nontrivial and $V^{\mathfrak{g}} = V$. There is a unique $d \geq 1$ such that the "Frobenius untwist" $V^{[-d]}$ is a G-module on which $\mathfrak{g}$ acts nontrivially. We have already seen that

$H^i(G, V^{[-d]}) = 0$ for $i = 1, 2$, so Theorem A follows from the following two results (see 5.4). [We denote by h the Coxeter number of the group G.]

Theorem B. *Suppose that $p \geq h$ and that W is a G-module for which $H^i(G, W) = 0$ for $i = 1, 2$. Then $H^2(G, W^{[d]}) \simeq \mathrm{Hom}_G(\mathfrak{g}, W)$ for all $d \geq 1$.*

Theorem C. *If $p > h$, $\dim H^2(G, \mathfrak{g}^{[d]}) = 1$ for all $d \geq 1$. For any p, there is a $d_0 \geq 1$ so that $H^2(G, \mathfrak{g}^{[d]}) \neq 0$ for all $d \geq d_0$.*

Theorem B is proved in 5.3; it immediately implies the first assertion of Theorem C (see 5.5). We give a proof of the second assertion of Theorem C in Section 5.6.

We end the paper by applying the results of Section 2 to calculations of cohomology groups $H^i(G_1, L)$, where G_1 is the Frobenius kernel, and L is a simple G_1 module with $\dim L \leq p$; see Proposition 6.

We make now the following remark concerning our hypothesis on G. Suppose that G is quasisimple, but not necessarily simply connected, and let $\pi : G_{\mathrm{sc}} \to G$ denote the isogeny from the corresponding simply connected covering group. Then any G representation V is also a G_{sc} representation, and the kernel of π is a diagonalizable group scheme. It follows that π induces an isomorphism $H^i(G, V) \simeq H^i(G_{\mathrm{sc}}, V)$ for each $i \geq 0$; see [**CPSvdK77**, Remark (2.7)]. I thank W. van der Kallen for pointing this out to me. One may check using Lemma 4.1(A) and Proposition 5.1 that $\mathrm{Lie}(G_{\mathrm{sc}})$ and $\mathrm{Lie}(G)$ are isomorphic simple G_{sc} representations whenever $\dim G \leq p$. Thus the conclusion of Theorem A remains true for G.

We conclude this introduction by remarking that the result of Jantzen [**Jan96**] cited above is one of several recent results studying the semisimplicity of low dimensional representations of groups in characteristic p. See [**Ser94**], [**McN98**], [**McN99**], [**Gur99**] and [**McN00**] for related work.

The author would like to acknowledge the hospitality of Bob Guralnick and the University of Southern California during a visit in June 1999; in particular, questions of Guralnick encouraged the author to consider the problems addressed in this paper, and several conversations inspired some useful ideas.

2. Root systems.

2.1. We denote by R an indecomposable root system in its weight lattice X with simple roots $S \subset R^+$. For each $\alpha \in S$, there is a fundamental dominant weight $\varpi_\alpha \in X$; the ϖ_α form a $\mathbb{Z}$ basis of X.

We write α_0 for the dominant short root, and $\widetilde{\alpha}$ for the dominant long root in R (these coincide in case there is only one root length).

The Coxeter number of R is given by

$$h - 1 = \sup_{\alpha \in R^+} \{\langle \rho, \alpha^\vee \rangle\} = \langle \rho, \alpha_0^\vee \rangle.$$

For $m \in \mathbb{Z}$ and $\alpha \in R$, let $s_{\alpha,m}$ denote the affine reflection of $X_{\mathbb{R}} = X \otimes_{\mathbb{Z}} \mathbb{R}$ in the hyperplane $H_{\alpha,m} = \{x \in X_{\mathbb{R}} : \langle x, \alpha^\vee \rangle = m\}$.

Let $l > h$ be an integer. The affine Weyl group W_l is the group of affine transformations of $X_{\mathbb{R}}$ generated by all $s_{\alpha,ln}$ for $n \in \mathbb{Z}$. According to [**Bou72**, Ch. VI, §2.1, Prop. 1] W_l is isomorphic to the semidirect product of W (the finite Weyl group) with $l\mathbb{Z}R$. The normalizer of W_l in the full affine transformation group of $X_{\mathbb{R}}$ contains all translations by lX, hence W_l is a normal subgroup of $\widehat{W}_l$, the semidirect product of W and lX. Moreover, $\widehat{W}_l/W_l \simeq lX/l\mathbb{Z}R \simeq X/\mathbb{Z}R$ is the fundamental group of R, which we will denote by π.

Let $\rho = \frac{1}{2}\sum_{\alpha \in S} \alpha$. We always consider the dot action of $\widehat{W}_l$ (also of W and W_l) on X: For $w \in \widehat{W}_l$ and $\lambda \in X$, this is given by $w \bullet \lambda = w(\lambda + \rho) - \rho$.

The closure of the subset C_l of $X_{\mathbb{R}}$ given by

$$C_l = \{\lambda \in X_{\mathbb{R}} \mid 0 < \langle \lambda + \rho, \alpha^\vee \rangle < l \quad \text{for each } \alpha \in R^+\}$$

is a fundamental domain for the dot action of W_l on X; the conjugates of C_l under W_l are known as alcoves, and C_l is the lowest alcove. Since $\widehat{W}_l$ normalizes W_l, [**Bou72**, Ch. VI, §2.1] shows that $\widehat{W}_l$ permutes the alcoves.

Let Ω be the stabilizer in $\widehat{W}_l$ of C. Since W_l permutes the alcoves simply transitively, one deduces that $\widehat{W}_l$ is the semidirect product of Ω and W_l. Thus $\Omega \simeq \widehat{W}_l/W_l \simeq \pi$.

Since $l > h$, the intersection $C_l \cap X^+$ is nonempty. [Note that if $l \leq h$ had been allowed, we would have $C_l \cap X^+ = \{0\}$ in case $l = h$, and $C_l \cap X^+ = \emptyset$ if $l < h$.] It is then clear that $\widehat{W}_l \bullet 0 \cap C_l = \{\omega \bullet 0 \mid \omega \in \Omega\}$.

2.2. Let I index the simple roots $S = \{\alpha_i\}$, write $\alpha_0^\vee = \sum_{i \in I} n_i \alpha_i^\vee$, and put $J = \{i \in I \mid n_i = 1\}$. A dominant weight $0 \neq \varpi \in X$ is *minuscule* if whenever $\lambda \leq \varpi$ and λ is a dominant weight, then $\varpi = \lambda$. According to [**Bou72**, Ch. VI, Exerc. 23, 24], ϖ is minuscule just in case $\varpi = \varpi_i$ for some $i \in J$.

For $i \in I \cup \{0\}$, let $S_i = S \setminus \{\alpha_i\}$ (so $S_0 = S$). Write $R_i \subset R$ for the root subsystem determined by S_i, and W_i for the parabolic subgroup of W associated with R_i. Let $w_i \in W_i$ be the unique element which makes all positive roots in R_i negative.

For $x \in X$, let $t(x)$ denote the affine translation by x; for $i \in J$, let $\gamma_i = t(l\varpi_i)w_0 w_i \in \widehat{W}_l$. Note that γ_i represents $\varpi_i \in X/\mathbb{Z}R \simeq lX/l\mathbb{Z}R \simeq \widehat{W}_l/W_l$.

Applying [**Bou72**, Ch. VI, §2.2 Prop. 6 and Cor.] one obtains:

Proposition.

(a) *Each non-*0 *coset of* $\mathbb{Z}R$ *in* X *is uniquely represented by a minuscule weight. In particular,* $|\pi| = |J| + 1$.

(b) *The nonidentity elements of* Ω *are precisely the* γ_i *for* $i \in J$. *We have*

$$\widehat{W}_l \bullet 0 \cap C_l = \{0\} \cup \{\gamma_i \bullet 0 = (l-h)\varpi_i \mid i \in J\}.$$

2.3. For a dominant weight λ, let

$$d(\lambda) = \prod_{\alpha>0} \frac{\langle \lambda + \rho, \alpha^\vee \rangle}{\langle \rho, \alpha^\vee \rangle} \tag{1}$$

be the value of Weyl's degree formula at λ.

Proposition. *Let* $\lambda = (l-h)\varpi_i$ *for some* $i \in J$.

(a) $d(\lambda) \geq \binom{l-1}{l-h}$, *with equality if and only if* $h - 1 = \ell(w_0 w_i)$.
(b) *If* $l - h \geq 2$ *and* $h \geq 3$, *then* $d(\lambda) > l$.

Proof. For $1 \leq k \leq h-1$, let $e(k)$ be the number of $\alpha \in R^+ \setminus R_i^+$ with $\langle \rho, \alpha^\vee \rangle = k$. The argument in the remark on p. 520-521 of [**Ser94**] (following Prop. 6) shows that $e(k) \geq 1$ for each $1 \leq k \leq h-1$. Thus, we have

$$d(\lambda) = \prod_{k=1}^{h-1} \left(\frac{l-h+k}{k} \right)^{e(k)} \geq \prod_{k=1}^{h-1} \frac{l-h+k}{k} = \binom{l-1}{l-h}.$$

If $\ell(w_0 w_i) = |R^+| - |R_i^+| = h-1$, then $e(k) = 1$ for each $1 \leq k \leq h-1$ and equality holds. This proves (a).

For (b), note that under the given hypothesis we have $l \geq 5$. Since $\binom{l-1}{l-h} \geq \binom{l-1}{2} > l$ for all such l, (b) follows immediately. □

Remark. Using the table in the proof of Proposition 2.4 below, it is straightforward to verify that equality holds in (a) if and only if either $R = A_r$ and $i \in \{1, r\}$ or $R = C_r$ and $i = 1$. (Since $B_2 = C_2$, the latter case includes B_2 and $i = 2$.)

2.4. In the following, let me emphasize the standing assumption $l > h$.

Proposition. *If* $0 \neq \lambda \in \widehat{W}_l \bullet 0 \cap C$ *and* $d(\lambda) < l$ *then* $d(\lambda) = \ell - 1$ *and* (R, λ) *is listed in the following table. If the rank of* R *is* ≥ 2, *then* $l = h+1$.

R	l	λ
A_1	any	$(l-2)\varpi_1$
A_{l-2}		ϖ_1, ϖ_{l-2}
B_2	$l = 5$	ϖ_2
$C_{(l-1)/2}$	l odd	ϖ_1

Proof. The rank 1 situation leads to the item listed in the table. When the rank is at least 2, one applies Proposition 2.3 to obtain $l = h+1$, whence $\lambda = \varpi_i$ for some $i \in J$; i.e., λ is minuscule.

We handle the minuscule cases by classification. For each indecomposable root system R for which $J \neq \emptyset$, we list in the following table the Coxeter number, the set J, and the value $d(\varpi_i)$ for each $i \in J$. The simple roots are indexed as in the tables in [**Bou72**, Planche I-X]; the data recorded here, with the exception of the values $d(\varpi_i)$, may be verified by inspecting those tables as well. The values $d(\varpi_i)$ are well-known (and can anyway be computed from the formula, or by representation theoretic arguments).

Type of R	h	J	$d(\varpi_i),\ i \in J$
A_r	$r+1$	$\{1, 2, \ldots, r\}$	$\binom{r+1}{i}$
$B_r, r \geq 2$	$2r$	$\{r\}$	2^r
$C_r, r \geq 2$	$2r$	$\{1\}$	$2r$
$D_r, r \geq 4$	$2r-2$	$\{1, r-1, r\}$	$2r, 2^{r-1}, 2^{r-1}$ respectively
E_6	12	$\{1, 6\}$	$27, 27$
E_7	18	$\{7\}$	56

From this table, one can list all pairs (R, λ) for which R has Coxeter number $l-1$ and λ is minuscule. It is a simple matter to see that $d(\lambda) < l$ only when (R, λ) is as claimed. □

3. The algebraic groups.

3.1. Let k be an algebraically closed field of characteristic $p > 0$, and let G be a connected, simply connected semisimple algebraic k-group. The non-0 weights of a maximal torus $T \leq G$ on $\mathfrak{g} = \mathrm{Lie}(G)$ form a root system R of rank $r = \dim T$ in the character group $X = X^*(T)$. Since G is simply connected, X identifies with the full weight lattice of R as in Section 2. We fix a choice of simple roots S and positive roots R^+. The dominant weights are denoted X^+. The group G is assumed to be *quasisimple*; i.e., the root system R is indecomposable.

3.2. For each dominant weight $\lambda \in X^+$, the space of global sections of the corresponding line bundle on the flag variety affords an indecomposable rational G-module $H^0(\lambda)$ with simple socle. The modules $L(\lambda) = \mathrm{soc}\, H^0(\lambda)$ comprise all of the simple rational modules for G (and are pairwise non-isomorphic).

The character of each $H^0(\lambda)$ is the same as in characteristic 0; hence in particular $\dim_k H^0(\lambda)$ is given by the Weyl degree formula, whose value at λ we denote $d(\lambda)$ as in 2.3.

3.3. Any dominant λ may be written as a finite sum $\sum_{i \geq 0} p^i \lambda_i$ with each λ_i a *restricted* weight. Recall that a dominant weight μ is restricted if $\langle \mu, \alpha^\vee \rangle < p$ for all simple roots α. Steinberg's tensor product theorem says:

$$L(\lambda) \simeq L(\lambda_0) \otimes L(\lambda_1)^{[1]} \otimes L(\lambda_2)^{[2]} \otimes \cdots$$

where for a G-module V, $V^{[m]}$ stands for the m-th Frobenius twist of V.

For $d \geq 1$, let G_d be the d-th Frobenius kernel of G. Let V be a rational G-module and $m \geq 1$. If there is a rational G module W with $W^{[m]} \simeq V$, we regard W as the Frobenius *untwist* $W = V^{[-m]}$ of V. Now regard V as a module for G_d. Since G_d is a normal subgroup scheme, G acts on V^{G_d}; since G_d acts trivially on this G-module, there is an untwisted rational G-module $(V^{G_d})^{[-d]}$. It follows that there is an untwist $H^i(G_d, V)^{[-d]}$ for all $i \geq 0$.

Consider now two G-modules V_1 and V_2, and form $W = V_1 \otimes V_2^{[d]}$. The Frobenius kernel G_d acts trivially on $V_2^{[d]}$, so that

$$H^i(G_d, W)^{[-d]} \simeq H^i(G_d, V_1)^{[-d]} \otimes V_2 \tag{1}$$

as G-modules for every $i \geq 0$.

3.4. Let $W_p \leq \widehat{W}_p$ be as in Section 2 (for $l = p$), let $C = C_p \cap X^+$ denote the dominant weights in the lowest alcove, and let $\overline{C} = \overline{C}_p \cap X^+$ ($\overline{C}_p$ is the closure in $X_{\mathbb{R}}$).

Proposition. *Let $\lambda \in X^+$.*

(a) *If $H^i(G, L(\lambda)) \neq 0$ for some $i \geq 0$, then $\lambda \in W_p \bullet 0$.*
(b) *If $H^i(G_1, L(\lambda)) \neq 0$ for some $i \geq 0$, then $\lambda \in \widehat{W}_p \bullet 0$.*
(c) *$H^i(G, H^0(\lambda)) = 0$ for all $i > 0$.*
(d) *If $\lambda \in \overline{C}$, then $L(\lambda) = H^0(\lambda)$; in particular, $\dim L(\lambda) = d(\lambda)$.*

Proof. (a) follows from the *linkage principle* for G [**Jan87**, Cor. II.6.17], and (b) from the linkage principle for G_1 [**Jan87**, Lemma II.9.16]. (c) follows from [**Jan87**, II.4.12]. (d) follows from [**Jan87**, II.6.13, II.5.10]. □

4. The Lie algebra and the cohomology of G_1.

We want to describe explicitly the cohomology $H^*(G_1, k)$ in degree ≤ 2. For this, we need some information on the Lie algebra $\mathfrak{g}$.

4.1. Recall that the prime p is *bad*[=not good] for the indecomposable root system R if one of the following holds: $p = 2$ and R is not of type A_r; $p = 3$ and R is of type G_2,F_4, or E_r; $p = 5$ and R is of type E_8.

The prime p is *very good* if it is not bad, and in case $R = A_r$, if also p does not divide $r + 1$. Notice that if $p > h$, then p is very good.

Application of the summary in [**Hum95**, 0.13] yields the following:

Lemma A. *Assume that p is* very good. *Then $\mathfrak{g}$ is a simple Lie algebra. The adjoint G-module is simple, self-dual, and isomorphic with $L(\widetilde{\alpha})$ where $\widetilde{\alpha}$ is the dominant long root.*

Lemma B. *Assume that $p \geq h$. If W is any G-module, then $\mathrm{Hom}_G(\mathfrak{g}, W^{[d]}) = 0$ for $d \geq 1$.*

Proof. When $p > h$ this follows since by the previous lemma $\mathfrak{g}$ is a simple $\mathfrak{g}$-module with restricted highest weight. When $p = h$, we have $R = A_{p-1}$. Since G is simply connected, we have $\mathfrak{g} = \mathfrak{sl}_p$. Thus $\mathfrak{g}$ is an indecomposable G-module with unique simple quotient $L(\widetilde{\alpha})$, and the lemma follows. □

4.2. Let B be a Borel subgroup of G, and let $\mathfrak{u}$ be the nilradical of $\mathrm{Lie}(B)$. Regarding $\mathfrak{u}^*$ as a B-module, we get a vector bundle on G/B which we also write as $\mathfrak{u}^*$. According to [**AJ84**, 3.8], the formal character of the G-module $H^0(G/B, \mathfrak{u}^*)$ is $\chi(\widetilde{\alpha}) = \mathrm{ch}(\mathfrak{g}^*)$.

Let $\mathcal{N} \subset \mathfrak{g}$ be the nilpotent cone. There is by [**AJ84**, 3.9] an injective homomorphism of graded algebras $k[\mathcal{N}] \to H^0(G/B, S\mathfrak{u}^*)$.

Lemma. *For simply connected, quasisimple algebraic groups G, $\mathfrak{g}^* \simeq k[\mathcal{N}]_1 \simeq H^0(G/B, \mathfrak{u}^*)$.*

Proof. Let $I(\mathcal{N}) \lhd k[\mathfrak{g}] = S\mathfrak{g}^*$ be the (homogeneous) defining ideal of the variety $\mathcal{N}$. We need to show that $I(\mathcal{N})_1 = 0$. If not, then $\mathcal{N} \subset V \subset \mathfrak{g}$ for some proper G-submodule V. A look at the summary in [**Hum95**, 0.13] shows that, since G is simply connected, the only G-submodules of $\mathfrak{g}$ have dimension 0 or 1. On the other hand, by [**Hum95**, Theorem 6.19], the variety $\mathcal{N}$ has codimension $\mathrm{rank}(G)$ in $\mathfrak{g}$ and so clearly can't be contained in a 1 dimensional linear subspace! □

Remarks.

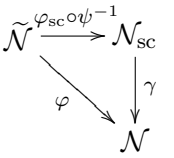
(1) Here is a fancier result which implies the lemma if we assume that the prime p is good for G. Since G is simply connected and p is good, the Springer resolution

$$\varphi : \widetilde{\mathcal{N}} = G \times^B \mathfrak{u} \to \mathcal{N}$$

given by $(g, X) \mapsto \mathrm{Ad}(g)(X)$ is a *desingularization*, hence in particular a birational map; see [**Hum95**, Theorem 6.3 and Theorem 6.20]. Again since G is simply connected and p is good, the variety $\mathcal{N}$ is normal ([**Hum95**, Theorem 4.24]). Standard arguments then yield an isomorphism of graded algebras $k[\mathcal{N}] \xrightarrow[\simeq]{\varphi^*} \Gamma(\widetilde{\mathcal{N}}, \mathcal{O}_{\widetilde{\mathcal{N}}})$. Finally, the projection $\widetilde{\mathcal{N}} \to G/B$ is an affine morphism, so that $\Gamma(\widetilde{\mathcal{N}}, \mathcal{O}_{\widetilde{\mathcal{N}}}) = H^0(G/B, S\mathfrak{u}^*)$ as a graded algebra.

(2) On the other hand, if $G = PGL_r$, and $p|r$, one can find a linear form on $\mathfrak{g}$ that vanishes on $\mathcal{N}$, hence there can be no isomorphism $k[\mathcal{N}]_1 \to H^0(G/B, \mathfrak{u}^*)$ (compare formal characters). So the lemma can fail when G is not simply connected. [Note that φ is not birational in this example. One can show that there is a G_{sc}-isomorphism $\psi : \widetilde{\mathcal{N}}_{\mathrm{sc}} \to \widetilde{\mathcal{N}}$ (using some obvious notations). We get therefore a commuting

diagram:



The map $\varphi_{\mathrm{sc}} \circ \psi^{-1}$ is birational. Since $\gamma^* k(\mathcal{N}) \subset k(\mathcal{N}_{sc})$ is a proper purely inseparable extension, so too is $\varphi^* k(\mathcal{N}) \subset k(\widetilde{\mathcal{N}})$.]

Proposition.

(1) *If $p \neq 2$ or if R is not of type C_r, then $H^1(G_1, k) = 0$.*
(2) *Assume that $p \geq h$. Then $H^2(G_1, k)^{[-1]} \simeq \mathfrak{g}^*$ as G-modules.*

Proof. For (1) see [**Jan87**, Lemma II.12.1]. For (2), first suppose that $p > h$. By [**AJ84**, 3.7, 3.9], there is a G-equivariant isomorphism of graded rings $k[\mathcal{N}]' \simeq H^*(G_1, k)^{[-1]}$ where $k[\mathcal{N}]'$ is again the graded coordinate ring of $\mathcal{N}$, but with the linear functions on $\mathfrak{g}$ given degree 2. The claim now follows from the lemma.

When $p = h$, apply [**AJ84**, Cor. 6.3] to see that $H^2(G_1, k)^{[-1]} \simeq H^0(G/B, \mathfrak{u}^*)$; the claim follows again from the lemma in this case. □

5. Low dimensional modules for G.

5.1. We recall first some facts about low dimensional modules established in [**Jan96**] and [**Ser94**].

Proposition. *Let L be a simple nontrivial restricted G module with highest weight λ. Suppose that* $\dim L \leq p$.

(a) $\lambda \in \overline{C}$.
(b) $\lambda \in C$ *if and only if* $\dim_k L < p$.
(c) $h \leq p$. *If moreover* $\dim L < p$, *then* $h < p$.
(d) *If R is not of type A and* $\dim L = p$, *then $h < p$. If $p = h$ and* $\dim L = p$, *then $R = A_{p-1}$ and $\lambda = \varpi_i$ with $i \in \{1, p-1\}$.*

Proof. (a) follows from [**Jan96**, Lemma 1.4], and (b) from [**Jan96**, 1.6], see also [**Ser94**]. For (c), note first that (a) implies $\dim L = d(\lambda)$ by Proposition 3.4(d). If $\lambda \in \overline{C} \setminus C$, then (a) and (b) imply that $\dim L = p$, whence $p = h$ follows from Weyl's degree formula. (c) now follows since C is empty if $p < h$ and $C = \{0\}$ if $p = h$.

In [**Jan96**, 1.6], Jantzen made a list of all simple restricted modules for G with dimension p. Inspecting that list yields (d). □

5.2. Vanishing results when $\mathfrak{g}$ acts nontrivially. Let L be a simple module for G.

Proposition. *If G_1 (equivalently: $\mathfrak{g}$) acts nontrivially on L and $\dim L \leq p$, then $H^i(G, L) = 0$ for all $i \geq 0$.*

Proof. Write the highest weight of L as $\lambda = \mu_1 + p\mu_2$ with μ_1 restricted. Since $L^{\mathfrak{g}} = 0$, we have $\mu_1 \neq 0$. Since $p \geq \dim L \geq \dim L(\mu_1)$, Proposition 5.1 implies that $\mu_1 \in \overline{C}$ and that $h \leq p$. We have in particular that $L(\mu_1) = H^0(\mu_1)$, hence the proposition will follow from Proposition 3.4 if we show that μ_2 is 0.

If $\dim L = p$, Steinberg's tensor product theorem gives $\mu_2 = 0$. If $\dim L < p$ then 5.1 shows that $p < h$ and $\mu_1 \in C$. If $H^i(G, L) \neq 0$ for some i, then $\lambda \in W_p \bullet 0$ by the linkage principle, whence $\mu_1 \in W \bullet 0 + pX = \widehat{W_p} \bullet 0$. Now Proposition 2.4 applies; it shows that $\dim L(\mu_1) = p - 1$ whence we have $\mu_2 = 0$ by another application of Steinberg's theorem. □

5.3. Second cohomology. Here we prove our main tool for describing second cohomology; first we require the following:

Lemma. *Let $E_2^{p,q} \implies H^{p+q}$ be a convergent, first quadrant spectral sequence.*

(1) *If $E_2^{0,1} = E_2^{1,1} = E_2^{0,2} = 0$, then $H^2 \simeq E_2^{2,0}$.*
(2) *If $E_2^{1,0} = E_2^{1,1} = E_2^{2,0} = 0$, then $H^2 \simeq E_2^{0,2}$.*

Proof. We verify (1), the argument for (2) is the same. We must show that $E_\infty^{2,0} \simeq E_2^{2,0}$; first note that $E_3^{2,0}$ is the cohomology of the sequence

$$E_2^{0,1} \to E_2^{2,0} \to E_2^{4,-1}$$

from which we get $E_3^{2,0} \simeq E_2^{2,0}$. For any first quadrant spectral sequence one has (by similar reasoning) that $E_a^{2,0} \simeq E_{a+1}^{2,0}$ for $a > 2$, so we get the desired isomorphism. □

Theorem. *Suppose that $p \geq h$. Let V be a G-module for which $H^i(G, V) = 0$ for $i = 1, 2$, and let $d \geq 1$.*

(1) *$H^1(G, V^{[d]}) = 0$, and*
(2) *$H^2(G, V^{[d]}) \simeq \mathrm{Hom}_G(\mathfrak{g}, V)$.*

Proof. The Frobenius kernel G_1 is a normal subgroup of G; thus there is a Lyndon-Hochschild-Serre spectral sequence computing $H^i(G, V^{[d]})$ which in view of 3.3 (1) has the form

$$E_2^{s,t} = H^s(G, H^t(G_1, V^{[d]})^{[-1]}) = H^s(G, H^t(G_1, k)^{[-1]} \otimes V^{[d-1]}).$$

If $t = 1$, $E_2^{s,t} = 0$ by Proposition 4.2(1).

There is an exact sequence of the form [**Jan87**, I.4.1(4)]

$$0 \to E_2^{1,0} \to H^1(G, V^{[d]}) \to E_2^{0,1} = 0.$$

Thus $H^1(G, V^{[d]}) \simeq E_2^{1,0} \simeq H^1(G, V^{[d-1]})$. We get now (1) by induction on d.

Proposition 4.2(2) shows now that $H^2(G_1, k) \simeq \mathfrak{g}^*$. Thus, the only possible non-0 E_2 terms of total degree 2 are

$$E_2^{0,2} = H^0(G, \mathfrak{g}^* \otimes V^{[d-1]}) = \mathrm{Hom}_G(\mathfrak{g}, V^{[d-1]})$$
$$E_2^{2,0} = H^2(G, V^{[d-1]}).$$

For $d > 1$, we apply Lemma 4.1(B) to see that $E_2^{0,2} = 0$ whence $H^2(G, V^{[d]}) \simeq E_2^{2,0} = H^2(G, V^{[d-1]})$ by part (1) of the lemma; thus (2) will follow provided it holds for $d = 1$. In that case, we have $E_2^{2,0} = 0$ by assumption, and the result just proved in part (1) shows that $E_2^{1,0} = 0$. Thus part (2) of the lemma applies; it shows that $H^2(G, V^{[1]}) \simeq E_2^{0,2} = \mathrm{Hom}_G(\mathfrak{g}, V)$ as desired. □

5.4. The second cohomology of small modules. Let $L = L(\lambda)$ be a simple G-module, and suppose that $\dim L \leq p$. Proposition 5.2 showed that the vanishing of cohomology for L is a consequence of the linkage principle when $\lambda \notin pX$. However, if $\lambda \in p\mathbb{Z}R$, λ is linked to 0, so the linkage principle does not yield vanishing. The following result shows that, despite the linkage of λ and 0 in this case, the second cohomology is usually 0.

Theorem. *Let L be a simple G-module with $\dim L \leq p$. If $H^2(G, L) \neq 0$, then $L \simeq \mathfrak{g}^{[d]}$ for some $d \geq 1$.*

Proof. Let L' be such that $L \simeq (L')^{[d]}$ for $d \geq 0$, and such that $\mathfrak{g}$ acts nontrivially on L'. We have by 5.1 that $p \geq h$. Also, we have by Proposition 5.2 that $H^i(G, L') = 0$ for $i \geq 1$. If $d = 0$, we are done. If $d > 1$, then Theorem 5.3 applies, and we get that

$$H^2(G, L) \simeq \mathrm{Hom}_G(\mathfrak{g}, L').$$

We get by Proposition 5.1 that $p > h$ unless $R = A_{p-1}$ and $L' = L(\varpi_i)$ with $i \in \{1, p-1\}$. If $p > h$, then $\mathfrak{g}$ is a simple G-module by Lemma 4.1(A). So if $\mathrm{Hom}_G(\mathfrak{g}, L') \neq 0$ then $L' \simeq \mathfrak{g}$ whence $L \simeq \mathfrak{g}^{[d]}$ as claimed.

In the remaining case, one must just note that weight considerations yield $\mathrm{Hom}_G(\mathfrak{g}, L(\varpi_i)) = 0$ for $i = 1, p-1$, whence $H^2(G, L) = 0$. □

5.5. The second cohomology of twists of the adjoint module. The first assertion of Theorem C of the introduction follows from the following:

Proposition. *Assume that $p > h$. Then $H^1(G, \mathfrak{g}^{[d]}) = 0$ and $H^2(G, \mathfrak{g}^{[d]}) \simeq \mathrm{End}_G(\mathfrak{g})$ has dimension 1 for $d \geq 1$.*

Proof. Since $p > h$, Lemma 4.1(A) shows that $\mathfrak{g}$ is the simple module with highest weight $\widetilde{\alpha}$. It follows that $\mathfrak{g} = H^0(\widetilde{\alpha})$, and thus that $H^i(G, \mathfrak{g}) = 0$ for $i \geq 1$ by Proposition 3.4. The proposition now follows from Theorem 5.3. □

Remark. Note that $\dim \mathfrak{g} > h$ (in fact, $\dim \mathfrak{g} = (h+1)r$ where r is the rank of G). So we get also: If $\dim \mathfrak{g} \leq p$, then $\dim H^2(G, \mathfrak{g}^{[d]}) = 1$ for $d \geq 1$.

5.6. A second proof. Here we give a second proof of the non-vanishing of H^2 for twists of the adjoint module; the result proved here verifies the remaining assertion of Theorem C of the introduction. We have included the argument since it offers some "explanation" for the non-vanishing.

The group G arises by base change from a split reductive group scheme $\mathbf{G}$ over $\mathbb{Z}$. Let $\mathbb{Z}_p$ be the complete ring of p-adic integers, and let $\mathbb{Q}_p$ be its field of quotients. For any finite field extension F of $\mathbb{Q}_p$, let $\mathfrak{o}$ denote the integers in F. The residue field $\mathfrak{o}/\mathfrak{m}$ may be identified with the extension $\mathbb{F}_q$ of $\mathbb{F}_p$.

Let K denote the group of points $\mathbf{G}(\mathfrak{o})$ regarded as a subgroup of $\mathbf{G}(F)$. Since $\mathbf{G}$ is smooth, the reduction homomorphism $K \to \mathbf{G}(\mathbb{F}_q)$ is surjective (see [**Tit79**, 3.4.4]).

For $n \geq 1$, let $K_n \subset K$ be the kernel of the map $K \to \mathbf{G}(\mathfrak{o}/\mathfrak{m}^n)$. Note that $K/K_1 = \mathbf{G}(\mathbb{F}_q)$ acts by conjugation on each quotient K_n/K_{n+1}.

Proposition. (a) *There is for each $m \geq 1$ a canonical isomorphism $K_m/K_{m+1} \simeq \mathfrak{g}_{\mathbb{F}_q}$ as representations for $\mathbf{G}(\mathbb{F}_q)$, where $\mathfrak{g}_{\mathbb{F}_q}$ is the Lie algebra of $\mathbf{G}_{\mathbb{F}_q}$.*

(b) *If $H^2(\mathbf{G}(\mathbb{F}_q), \mathfrak{g}_{\mathbb{F}_q}) = 0$, the exact sequence of groups*

$$1 \to K_1 \to K \to \mathbf{G}(\mathbb{F}_q) \to 1$$

splits.

(c) *There is a p-power q_0, depending only on the root system R of G, such that $H^2(\mathbf{G}(\mathbb{F}_q), \mathfrak{g}_{\mathbb{F}_q}) \neq 0$ whenever $q \geq q_0$.*

(d) *There is an integer $a_0 \geq 1$ such that $H^2(G, \mathfrak{g}^{[a]}) \neq 0$ whenever $a \geq a_0$.*

Proof. (a) Follows from [**DG70**, II.§4.3]. (b) Since K_1 is a pro-p group [**PR94**, Lemma 3.8], this follows from [**Ser67**, Lemma 3].

(c) Choose a $\mathbb{Q}_p$ vectorspace V and a nontrivial faithful $\mathbb{Q}_p$-rational representation $\mathbf{G}_{\mathbb{Q}_p} \to \mathrm{GL}(V)$. For each extension F of $\mathbb{Q}_p$ with integers $\mathfrak{o}$, the group $K = \mathbf{G}(\mathfrak{o})$ is a subgroup of (the group of F-points of) $\mathrm{GL}(V_F)$. If $H^2(\mathbf{G}(\mathbb{F}_q), \mathfrak{g}_{\mathbb{F}_q}) = 0$, the sequence in (b) is split and V_F is a nontrivial $F[\mathbf{G}(\mathbb{F}_q)]$-module.

Since F has characteristic 0, it is well-known that the minimal dimension of a nontrivial $F[\mathbf{G}(\mathbb{F}_q)]$ module is bounded below by the value $f(q)$ of a polynomial $f \in \mathbb{Q}[x]$, depending only on G, for which $f(q) \to \infty$ as $q \to \infty$. We may choose q_0 such that $f(q) > \dim_{\mathbb{Q}_p} V$ for each $q > q_0$, and (c) follows at once.

(d) now follows from (c) and [**CPSvdK77**, Cor. 6.9]. □

6. Small simple modules for G_1.

Combining results of [**KLT99**] with the results recorded in 2.4, we obtain some explicit results on G_1 cohomology of low dimensional simple modules:

Proposition. *Let L be a nontrivial simple G_1 module with $\dim \leq p$. Assume for some $i \geq 0$ that $H^i(G_1, L) \neq 0$. Then $\dim L = p - 1$. Moreover, there is a quadruple $(R, \lambda, i(0), V)$ in the following table for which R is the root system of G, λ the high weight of L, $i \geq i(0)$ and $H^{i(0)}(G_1, L)^{[-1]} \simeq V$ as G-modules.*

R		λ	$i(0)$	$H^{i(0)}(G_1, L)^{[-1]}$
A_1		$(p-2)\varpi_1$	1	$L(\varpi_1)$
A_{p-2}		ϖ_1, ϖ_{p-2}	$p-2$	$L(\lambda)$
$C_{(p-1)/2}$	*p odd*	ϖ_1	$p-2$	$L(\lambda)$

Proof. By [**Jan87**, Prop. II.3.14], $L = \mathrm{res}^G_{G_1} L(\lambda)$ for some restricted dominant weight $0 \neq \lambda$. Thus $L(\lambda)$ is a restricted, simple G module with dimension $\leq p$. It follows from Proposition 5.1 that $h \leq p$, that $\lambda \in \overline{C}$, and that $L = H^0(\lambda)$ as modules for G.

Suppose that $H^i(G_1, L) \neq 0$ for some i. By the linkage principle for G_1 (Proposition 3.4(b)), we must have $\lambda \in \widehat{W_p} \bullet 0$, hence $\lambda \in C$. This implies that $h < p$. Proposition 2.2 shows that $\lambda = (p-h)\varpi_i = w_0 w_i \bullet 0 + p\varpi_i$ for some $i \in J$, and Proposition 2.3 yields $\dim L = p - 1$ and lists the possible pairs (R, λ).

For $h < p$, Kumar, Lauritzen and Thomsen [**KLT99**, Theorem 8] have extended a result of Andersen and Jantzen [**AJ84**, 3.7]; this result implies in particular that the minimal degree for which $H^*(G_1, L)$ is non-0 is $\ell(w_0 w_i)$, and that

$$H^{\ell(w_0 w_i)}(G_1, L)^{[-1]} \simeq H^0(\varpi_i).$$

It is straightforward to compute for each pair (R, λ) the length $\ell(w_0 w_i)$; one gets in this way the result. □

Remark. The Theorem implies the fact (used by Jantzen in the proof of [**Jan96**, Lemma 1.7]) that $H^1(G_1, L) = 0$ for all simple G_1 modules L with $\dim L \leq p$. The argument used by Jantzen there relied on the calculations of H^1 carried out in [**Jan91**].

References

[AJ84] H.H. Andersen and J.C. Jantzen, *Cohomology of induced representations for algebraic groups*, Math. Ann., **269** (1984), 487-525, MR 86g:20057, Zbl 0529.20027.

[Bou72] N. Bourbaki, *Groupes et algèbres de Lie, Chapitres* 4, 5, 6, Hermann, Paris, 1972, MR 58 #28083a, Zbl 0249.22001.

[CPSvdK77] E. Cline, B. Parshall, L. Scott and W. van der Kallen, *Rational and generic cohomology*, Invent. Math., **39** (1977), 143-163, MR 55 #12737, Zbl 0346.20031.

[DG70] M. Demazure and P. Gabriel, *Groupes algébriques. Tome* I: *Géométrie algébrique, généralités, groupes commutatifs*, Masson & Cie, Éditeur, Paris, 1970, Avec un appendice 'Corps de classes local' par Michiel Hazewinkel, MR 46 #1800, Zbl 0203.23401.

[Gur99] R.M. Guralnick, *Small representations are completely reducible*, J. Algebra, **220**(**2**) (1999), 531-541, MR 2000m:20018, Zbl 0941.20001.

[Hum95] J.E. Humphreys, *Conjugacy classes in semisimple algebraic groups*, Math. Surveys and Monographs, **43**, Amer. Math. Soc., 1995, MR 97i:20057, Zbl 0834.20048.

[Jan87] J.C. Jantzen, *Representations of algebraic groups*, Pure and Applied Mathematics, **131**, Academic Press, Orlando, FL, 1987, MR 89c:20001, Zbl 0654.20039.

[Jan91] ———, *First cohomology groups for classical Lie algebras*, Representation Theory of Finite Groups and Finite Dimensional Algebras (Bielefeld) (G.O. Michler and C.M. Ringel, eds.), Progr. in Math., **95**, Birkhäuser, Boston, 1991, 289-315, MR 92e:17024, Zbl 0749.17020.

[Jan96] ———, *Low dimensional representations of reductive groups are semisimple*, Algebraic Groups and Related Subjects; a Volume in Honour of R. W. Richardson (G.I. Lehrer et al., ed.), Austral. Math. Soc. Lect. Ser., Cambridge Univ. Press, Cambridge, 1996, 255-266, MR 99g:20079, Zbl 0877.20029.

[KLT99] S. Kumar, N. Lauritzen and J.F. Thomsen, *Frobenius splitting of cotangent bundles of flag varieties*, Invent. Math., **136** (1999), 603–621, MR 2000g:20088, Zbl 0959.14031.

[McN98] G.J. McNinch, *Dimensional criteria for semisimplicity of representations*, Proc. London Math. Soc., **76**(**3**) (1998), 95-149, MR 99b:20076, Zbl 0891.20032.

[McN99] ———, *Semisimple modules for finite groups of Lie type*, J. London Math. Soc. **60** (1999), no. 2, 771–792, MR 2001k:20096, Zbl 0961.20014.

[McN00] ———, *Semisimplicity of exterior powers of simple representations of groups*, J. Alg., **225** (2000), 646-666, MR 2001c:20016, Zbl 0971.20005.

[PR94] V. Platonov and A. Rapinchuk, *Algebraic groups and number theory*, Pure and Applied Mathematics, **139**, Academic Press, 1994, English translation, MR 95b:11039, Zbl 0841.20046.

[Ser67] J.P. Serre, *Local class field theory*, Algebraic Number Theory (Proc. Instructional Conf., Brighton, 1965), Thompson, Washington, D.C., 1967, 128-161, MR 36 #3753.

[Ser94] ———, *Sur la semi-simplicité des produits tensoriels de représentations de groupes*, Invent. Math., **116** (1994), 513-530, MR 94m:20091, Zbl 0816.20014.

[Tit79] J. Tits, *Reductive groups over local fields*, **XXXIII** (**1**), Proc. Sympos. Pure Math., Amer. Math. Soc., (1979), 29-69, MR 80h:20064, Zbl 0415.20035.

Received October 26, 2000. This work was supported by a grant from the National Science Foundation.

Department of Mathematics
University of Notre Dame
Notre Dame, IN 46556
E-mail address: McNinch.1@nd.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

HOW LIKELY IS BUFFON'S NEEDLE TO FALL NEAR A PLANAR CANTOR SET?

Yuval Peres and Boris Solomyak

Dedicated to the memory of Thomas H. Wolff

Let Λ be a compact planar set of positive finite one-dimensional Hausdorff measure. Suppose that the intersection of Λ with any rectifiable curve has zero length. Then a theorem of Besicovitch (1939) states that the orthogonal projection of Λ on almost all lines has zero length. Consequently, the probability $p(\Lambda, \epsilon)$ that a needle dropped at random will fall within distance ϵ from Λ, tends to zero with ϵ. However, existing proofs do not yield any explicit upper bound tending to zero for $p(\Lambda, \epsilon)$, even in the simplest cases, e.g., when $\Lambda = K^2$ is the Cartesian square of the middle-half Cantor set K. In this paper we establish such a bound for a class of self-similar sets Λ that includes K^2. We also determine the order of magnitude of $p(\Lambda, \epsilon)$ for certain stochastically self-similar sets Λ. Determining the order of magnitude of $p(K^2, \epsilon)$ is an unsolved problem.

1. Introduction.

Consider $K = \{\sum_{n=1}^{\infty} a_n 4^{-n} : \ a_n \in \{0, 3\}\}$, the middle-half Cantor set, and the direct product $K^2 = K \times K \subset \mathbb{R}^2$. It is well-known that the one-dimensional Hausdorff measure of K^2 satisfies $0 < \mathcal{H}^1(K^2) < \infty$ and that K^2 is totally unrectifiable. Therefore, by Besicovitch's theorem (see [**4**, Theorem 6.13]), the projection of K^2 on almost every line through the origin, has zero length. This can be expressed by saying that the Favard length of K^2 equals zero. Recall (see [**2**, p. 357]) that the **Favard length** of a planar set E is defined by

$$\mathrm{Fav}(E) = \int_0^{\pi} |\mathrm{proj}_\theta E| \, d\theta,$$

where proj_θ denotes the orthogonal projection from $\mathbb{R}^2$ onto the line through the origin making angle θ with the horizontal axis, and $|A|$ denotes the Lebesgue measure of a measurable set $A \subset \mathbb{R}$. The Favard length of a set E in the unit square has a probabilistic interpretation: Up to a constant factor, it is the probability that "Buffon's needle," a long line segment dropped at

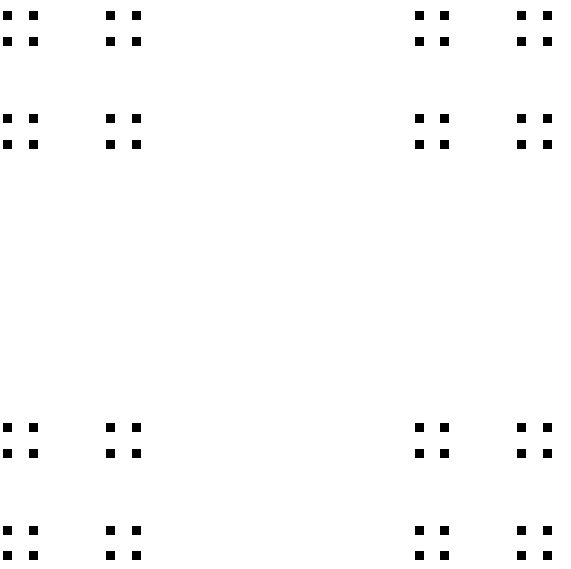


Figure 1. The Cantor set K^2, third stage of the construction.

random, hits E. (More precisely, suppose the needle's length is greater than $\sqrt{8}$, pick the distance r from the origin to the needle uniformly in $[0, \sqrt{2}]$, and locate the center of the needle at a uniformly chosen point on the circle $\{|z| = r\}$.)

Now consider the n-th stage of the Cantor set construction for K,

$$K_n = \Big\{ \sum_{k=1}^{\infty} a_k 4^{-k} : \ a_k \in \{0,3\} \quad \text{for } \ 1 \le k \le n \text{ and } \ a_k \in \{0,1,2,3\} \ \text{ for } k > n \Big\}.$$

Then K_n^2 is a union of 4^n squares of side 4^{-n} (see Figure 1 for a picture of K_3^2). Clearly, $\mathrm{Fav}(K^2) = 0$ implies $\lim_{n\to\infty} \mathrm{Fav}(K_n^2) = 0$. We are interested in the behavior of $\mathrm{Fav}(K_n^2)$ as $n \to \infty$. A lower bound $\mathrm{Fav}(K_n^2) \ge \frac{c}{n}$ for some $c > 0$ follows from Mattila [**14**, 1.4]. Our main result is a quantitative

upper bound. For $y \geq 1$ let

$$\log_* y = \min\left\{ n \geq 0 : \underbrace{\log\log\ldots\log}_{n} y \leq 1 \right\}. \tag{1.1}$$

Theorem 1.1. *There exist $C, a > 0$ such that*

$$\mathrm{Fav}(K_n^2) \leq C \exp[-a \log_* n] \quad \textit{for all} \ \ n \in \mathbb{N}.$$

Remarks.
1. The convergence of the upper bound to zero is extremely slow, but it is the best we could get. It is still much better than a purely qualitative convergence statement. The lower bound $\frac{c}{n}$ seems closer to the truth. In Theorem 2.2, proved in Section 6, we analyze a random analog of the Cantor set K^2. We show that, with high probability, the Favard length of the n-th stage in the construction has upper and lower bounds that are constant multiples of n^{-1}.

2. For $\rho \leq 4^{-n}$, the ρ-neighborhood $K(\rho) = \{\mathbf{x} : \ \mathrm{dist}(\mathbf{x}, K) \leq \rho\}$ of K can be covered by nine translates of K_n, so $\mathrm{Fav}(K(\rho)) \leq 9\mathrm{Fav}(K_n)$.

3. It follows from the results of Kenyon [**9**] and Lagarias and Wang [**10**] that $|\mathrm{proj}_\theta K^2| = 0$ for all θ such that $\tan\theta$ is irrational. However, this information does not seem to help obtain an upper bound for $\mathrm{Fav}(K_n^2)$.

4. The set K^2 was one of the first examples of sets of positive length and zero analytic capacity, see [**3**] for a survey. Recently Mateu, Tolsa and Verdera [**12**] proved that the analytic capacity of K_n^2 is bounded above and below by constant multiples of $n^{-1/2}$. The analytic capacity of certain related sets of non σ-finite length was determined by Mattila [**16**]. We consider the Favard length of such sets in Proposition 7.2.

In the next section we state our results for a class of planar self-similar Cantor sets. The method used for estimating the Favard length of the n-th stage of the construction also yields some information about gauges in which almost every projection of the Cantor set has zero Hausdorff measure. The proof of the main theorem for homogeneous self-similar sets (such as K^2) is presented in Sections 3 and 4. The non-homogeneous case, which is more involved, is dealt with in Section 5. Favard length of random Cantor sets is considered in Section 6. Section 7 contains some further extensions, remarks and unsolved problems.

2. Statement of results.

Consider a self-similar set $\Lambda \subset \mathbb{R}^2$, defined as the unique nonempty compact satisfying

(2.1)
$$\Lambda = \bigcup_{i=1}^{m} S_i\Lambda \quad \text{where} \quad S_i(\mathbf{x}) = r_i\mathbf{x} + b_i, \quad \text{with} \quad r_i \in (0,1) \quad \text{and} \quad b_i \in \mathbb{R}^2.$$

We assume that the **Strong Separation Condition** (SSC) holds, *i.e.*, that $S_i(\Lambda) \cap S_j(\Lambda) = \emptyset$ for $i \neq j$. The similarity dimension is defined as the unique solution s of the equation $\sum_{i=1}^{m} r_i^s = 1$. It is well-known that the Strong Separation Condition, and even the weaker Open Set Condition, imply that the Hausdorff dimension $\dim_H \Lambda$ equals the similarity dimension s, and the s-dimensional Hausdorff measure $\mathcal{H}^s(\Lambda)$ is positive and finite.

First suppose that $s = 1$. Then Λ is an irregular 1-set, and thus by Besicovitch's theorem (see [**4**, Theorem 6.13]) $\mathrm{Fav}(\Lambda) = 0$. Let $\Lambda(\rho) = \{\mathbf{x} : \ \mathrm{dist}(\mathbf{x}, \Lambda) \leq \rho\}$ denote the ρ-neighborhood of the set Λ. Clearly, $\lim_{\rho \to 0} \mathrm{Fav}(\Lambda(\rho)) = 0$. Mattila [**14**, 1.4] proved the lower bound

$$\mathrm{Fav}(\Lambda(\rho)) \geq c \left(\log \left(\frac{1}{\rho} \right) \right)^{-1} \quad \text{for all} \quad \rho > 0, \tag{2.2}$$

for some $c > 0$. (This lower bound follows from an energy estimate; it does not use self-similarity, but only positivity of $\mathcal{H}^1(\Lambda)$.) Our main result is the following upper bound.

Theorem 2.1. *Assuming that the* SSC *holds and $s = 1$, we have for some $C, a > 0$*

$$\mathrm{Fav}(\Lambda(\rho)) \leq C \exp\left[-a \log_* \left(\frac{1}{\rho} \right) \right] \quad \textit{for all} \quad \rho > 0. \tag{2.3}$$

Remark. The self-similar set is called **homogeneous** if $r_i = r$ for all $i \leq m$. The Cantor set K^2 in Section 1 is homogeneous. For a homogeneous set Λ, it is equivalent (up to uniform multiplicative constants) to consider the Favard length $\mathrm{Fav}(\Lambda_n)$ of the nth stage of the construction and $\mathrm{Fav}(\Lambda(\rho))$, with $\rho = r^n$.

We now consider random analogs of the sets K_n^2 from the introduction.

Partition the unit square into four dyadic subsquares of side $1/2$, and in each of these choose, uniformly at random, a dyadic subsquare of side $1/4$. Denote the union of four (closed) squares so obtained $\mathcal{R}_1$. Inductively, given $\mathcal{R}_k$ which is a union of 4^k dyadic squares of side 2^{-2k}, we partition each of them into four dyadic subsquares of side 2^{-2k-1}, and in each of these 4^{k+1} squares choose, uniformly at random, a dyadic subsquare of side 2^{-2k-2}, all these choices being independent. Call the union of 4^{k+1} (closed) squares so obtained $\mathcal{R}_{k+1}$. An example of $\mathcal{R}_3$ is given in Figure 2.

Figure 2. A random set $\mathcal{R}_3$.

Finally, write $\mathcal{R} = \bigcap_{k=1}^{\infty} \mathcal{R}_k$. Clearly $0 < \mathcal{H}^1(\mathcal{R}) < \infty$, and the arguments of Mattila [**14**, 1.4] still imply that $\mathrm{Fav}(\mathcal{R}_n) \geq \frac{c}{n}$. Denoting expectation by $\mathbf{E}$, we have:

Theorem 2.2.

$$\mathbf{E}\left[\mathrm{Fav}(\mathcal{R}_n)\right] \leq \frac{C}{n} \tag{2.4}$$

for some $C < \infty$. Consequently, with probability 1*,*

$$\liminf_{n\to\infty} n \cdot \mathrm{Fav}(\mathcal{R}_n) < \infty. \tag{2.5}$$

Next, we return to consider self-similar sets Λ as in (2.1), but only assume that their similarity dimension satisfies $s \leq 1$. Let

$$\mathcal{I}P(\Lambda) = \left\{\theta \in [0,\pi] : \ \mathrm{proj}_\theta|_\Lambda \ \text{ is not one-to-one}\right\}$$

(the letters "$\mathcal{I}P$" stand for "intersection parameters"). It is easy to see that if $s = 1$, then $\mathcal{I}P(\Lambda) = [0,\pi]$. (Indeed, if $\mathrm{proj}_\theta|_\Lambda$ is one-to-one, then $\mathrm{proj}_\theta(\Lambda)$ is a self-similar set on the real line satisfying the Strong Separation

Condition. Increasing the contraction rates r_i slightly (maintaining strong separation), we would get a subset of $\mathbb{R}$ with Hausdorff dimension greater than 1, a contradiction.)

It was proved in [**18**, Theorem 1.2] that if the set $\mathcal{I}P(\Lambda)$ contains a nonempty interval J, then $\mathcal{H}^s(\mathrm{proj}_\theta\Lambda) = 0$ for a.e. $\theta \in J$. Here we exhibit an explicit gauge function $\phi(t)$ such that $\lim_{t\to 0} \frac{\phi(t)}{t^s} = \infty$ but $\mathcal{H}^\phi(\mathrm{proj}_\theta\Lambda) = 0$ for a.e. $\theta \in J$.

Theorem 2.3. *If the* SSC *holds, $s \le 1$, and there is an interval $J \subset \mathcal{I}P(\Lambda)$, then $\mathcal{H}^\phi(\mathrm{proj}_\theta\Lambda) = 0$ for a.e. $\theta \in J$, where*

$$\phi(t) = t^s \exp[L \log_*(1/t)]$$

with $L \in (0, \log 2)$.

Sufficient conditions for the existence of an interval $J \subset \mathcal{I}P(\Lambda)$ were found in [**18**]. For instance, Theorem 2.3 applies to the planar Cantor set $K^{(r)} \times K^{(r)}$ where $K^{(r)} = \{\sum_{n=1}^\infty a_n r^n : \ a_n \in \{0,1\}\}$, with $r \in (\frac{1}{6}, \frac{1}{4})$. It is shown in [**18**, Example 6.1] that $J = [\arctan \frac{1-2r}{r}, \arctan \frac{2}{1-3r}] \subset \mathcal{I}P(K^{(r)})$.

3. Proof of Theorem 2.1 (the homogeneous case).

Here we prove Theorem 2.1 in the case when $r_i = r$; this includes Theorem 1.1. Note that $s = 1$ implies $r = m^{-1}$. Since some of the lemmas will also be used in the proof of Theorem 2.3, up to a point we allow any value of $s \le 1$. The more technical proofs of lemmas are postponed until the next section.

Let $m \ge 2$, $\mathcal{A} = \{1, \dots, m\}$ and $\mathcal{A}^* = \bigcup_{n\ge 0} \mathcal{A}^n$. Write $|u| = n$ for $u \in \mathcal{A}^n$ and let $\omega|_n = \omega_1 \dots \omega_n$ for $\omega \in \mathcal{A}^* \cup \mathcal{A}^{\mathbb{N}}$, with $|\omega| \ge n$. For $u \in \mathcal{A}^n$ we write $S_u = S_{u_1} \circ \dots \circ S_{u_n}$ and $\Lambda_u = S_u(\Lambda)$. In our homogeneous case we have $S_u(\mathbf{x}) = r^n \mathbf{x} + b_u$ for some $b_u \in \mathbb{R}^2$. It is convenient to identify the line through the origin with $\mathbb{R}$; formally we just let $\mathrm{proj}_\theta(x, y) = x\cos\theta + y \sin\theta$. For $\theta \in [0, \pi]$ and $u \in \mathcal{A}^n$ let

$$S_u^\theta(x) = r^n x + b_u^\theta, \ x \in \mathbb{R}, \quad \text{where} \ \ b_u^\theta = \mathrm{proj}_\theta b_u.$$

Observe that $\Lambda^\theta := \mathrm{proj}_\theta\Lambda$ is a self-similar set on the real line satisfying $\Lambda^\theta = \bigcup_{i=1}^m S_i^\theta(\Lambda^\theta)$. The sets $\Lambda_u^\theta := \mathrm{proj}_\theta\Lambda_u$ are called the **cylinders** of the self-similar set Λ^θ. The map $\Pi_\theta : \ \mathcal{A}^{\mathbb{N}} \to \Lambda^\theta$ defined by

$$\Pi_\theta(\omega) = \lim_{n\to\infty} S^\theta_{\omega|_n}(0) = \sum_{n=1}^\infty r^{n-1} b^\theta_{\omega_n},$$

is called the **natural projection map**. We equip the sequence space $\mathcal{A}^{\mathbb{N}}$ with the Bernoulli measure $(\frac{1}{m}, \dots, \frac{1}{m})^{\mathbb{N}}$. The projection of μ, that is, $\nu_\theta := \mu \circ \Pi_\theta^{-1}$ is called the **natural measure** on Λ^θ.

Definition 3.1. Let u, v be two words in $\mathcal{A}^*$ with $|u| = |v| = n$ and let $\theta \in [0, \pi]$. We say that S_u^θ and S_v^θ are ε-**relatively close** if

$$|S_u^\theta(x) - S_v^\theta(x)| \le \varepsilon r^n \quad \text{for all } \ x \in \Lambda^\theta. \tag{3.1}$$

This definition is motivated by the work of Bandt and Graf [**1**]; it was recently used in [**17**]. In order to develop the setting needed for the proof of both Theorem 2.1 and Theorem 2.3, we fix a nonempty interval $J \subset \mathcal{I}P(\Lambda)$; if $s = 1$ then we let $J = \mathcal{I}P(\Lambda) = [0, \pi]$.

Lemma 3.2. *There exists $C_1 \ge 1$ such that for all $\varepsilon \in (0, 1]$ and all $n \in \mathbb{N}$, for any interval $I \subset J$, with $|I| = C_1 r^n$, there is a subinterval $I' \subset I$ satisfying:*

(i) $|I'| \ge C_1^{-2} \varepsilon |I|$ *and*

(ii) *for every $\theta \in I'$ there exist $u \ne v$ in $\mathcal{A}^n$ such that S_u^θ and S_v^θ are ε-relatively close.*

This is a consequence of "transversality"; the proof is given in Section 4.

Notation. Let $\Psi(n, k, \varepsilon)$ be the set of $\theta \in J$ such that there is no collection of distinct words $u_1, \dots, u_k$, with $|u_1| = \dots = |u_k| \le n$, such that $S_{u_j}^\theta$, $j \le k$, are pairwise ε-relatively close.

Lemma 3.3. *There exist $c_2 > 0$ and $M > 0$ such that*

$$|\Psi(n, 2, \varepsilon)| \le M e^{-c_2 n \varepsilon} \quad \text{for all } \ n \in N, \ \varepsilon \in (0, 1]. \tag{3.2}$$

This follows from Lemma 3.2; see Section 4 for the proof.

Lemma 3.4. *If $n = \ell_0 + j_0$, with $\ell_0, j_0 \ge 1$, and $k \ge 2$, then*

$$\Psi(n, 2k, \varepsilon) \subset \Psi(\ell_0, 2, (\varepsilon/2) r^{j_0}) \cup \Psi(j_0, k, (\varepsilon/2)). \tag{3.3}$$

Proof. Suppose that θ is not in the right-hand side of (3.3). Then there exist distinct u_1, u_2, with $|u_i| \le \ell_0$, such that $S_{u_1}^\theta$ and $S_{u_2}^\theta$ are $\frac{\varepsilon}{2} r^{j_0}$-relatively close, and distinct $w_1, \dots, w_k$, with $|w_q| \le j_0$, such that $S_{w_1}^\theta, \dots, S_{w_k}^\theta$ are pairwise $\frac{\varepsilon}{2}$-relatively close. Let $1 \le p < q \le k$. By self-similarity, $S_{u_i w_p}^\theta$ and $S_{u_i w_q}^\theta$ are $\frac{\varepsilon}{2}$-relatively close. Further, $S_{u_1 w_q}^\theta$ and $S_{u_2 w_q}^\theta$ are $\frac{\varepsilon}{2}$-relatively close, since $r^{j_0 + |u_i|} \le r^{|u_i w_q|}$, for $i = 1, 2$. This implies that $S_{u_1 w_p}^\theta$ and $S_{u_2 w_q}^\theta$ are ε-relatively close. Thus, we have found $2k$ distinct words $u_i w_q$, with $i = 1, 2$ and $q \le k$, of length $\le n$, such that $S_{u_i w_q}^\theta$ are pairwise ε-relatively close, hence $\theta \notin \Psi(n, 2k, \varepsilon)$. □

Below we denote by $\log^i$ and $\exp^i$ the i-th iterate of log and exp respectively, assuming that $\log^0$ is the identity map.

Lemma 3.5. *There exists $c_3 > 0$ such that for all $i \ge 1$,*

$$|\Psi(n, 2^i, \varepsilon)| \le M 2^{i-1} \exp[-c_3 e^{-(i-1)} (\log^{i-1} n) \varepsilon] \ \text{ for all } \ n \in \mathbb{N}, \ \varepsilon \in (0, 1]. \tag{3.4}$$

This is proved by induction, using Lemmas 3.3 and 3.4. See Section 4 for details. Now let

$$n_k = \exp^{k-1}(c_3^{-1} k e^{k-1}), \tag{3.5}$$

so that, in view of (3.4),

$$|\Psi(n_k, 2^k, 1)| \le M 2^{k-1} e^{-k}. \tag{3.6}$$

For $v, w \in \mathcal{A}^*$ we write $v \sqsubset w$ if v is a **subword** of w, more precisely, if $w = v'vv''$ where v' and/or v'' may be empty. Let

$$N(k) := m^{n_k} \cdot n_k \cdot k. \tag{3.7}$$

For $u_1 \in \mathcal{A}^*$, with $|u_1| \le n_k$, we have

$$\begin{aligned} \#\{u \in \mathcal{A}^{N(k)} : \ u_1 \not\sqsubset u\} &\le (m^{n_k} - 1)^{N(k)/n_k} \\ &= m^{N(k)}(1 - m^{-n_k})^{m^{n_k} \cdot k} \le m^{N(k)} e^{-k}. \end{aligned} \tag{3.8}$$

Lemma 3.6. *For any $\xi > 0$ there exists $C_\xi > 0$ such that*

$$\log_*(r^{-N(k)}) \le C_\xi + (1 + \xi)k.$$

This is elementary; see Section 4 for a proof.

Proof of Theorem 2.1 (*homogeneous case*). Recall that now $s = 1$, so $r = m^{-1}$. We are going to show that, for some $c > 0$ and $\gamma \in (0, 1)$,

$$\mathrm{Fav}(\Lambda(\rho)) \le c\gamma^k, \quad \text{where } \ \rho = r^{N(k)}. \tag{3.9}$$

By Lemma 3.6, this will imply (2.3).

Turning to the proof of (3.9), we note that by (3.6),

$$\int_{\Psi(n_k, 2^k, 1)} |\mathrm{proj}_\theta \Lambda(\rho)| \, d\theta \le M 2^{k-1} e^{-k} (\mathrm{diam}(\Lambda) + 2). \tag{3.10}$$

Thus, it suffices to estimate $|\mathrm{proj}_\theta \Lambda(\rho)|$ from above for $\theta \notin \Psi(n_k, 2^k, 1)$.

Fix such a θ for the rest of the proof. By definition, this means that there exist words $u_1, \dots, u_{2^k}$, each of length not greater than n_k, such that $S^\theta_{u_j}$, $j \le 2^k$, are pairwise 1-relatively close. We have

$$\mathrm{proj}_\theta \Lambda(\rho) \subset \bigcup_{|u| = N(k)} \Lambda^\theta_u(\rho) = \bigcup_{\substack{|u| = N(k) \\ u_1 \not\sqsubset u}} \Lambda^\theta_u(\rho) \cup \bigcup_{\substack{|u| = N(k) \\ u_1 \sqsubset u}} \Lambda^\theta_u(\rho) =: F_1 \cup F_2.$$

Since $|u| = N(k)$ we have

$$\mathrm{diam}(\Lambda^\theta_u(\rho)) = \mathrm{diam}(\Lambda^\theta_u) + 2\rho \le (2 + \mathrm{diam}(\Lambda)) m^{-N(k)},$$

hence, in view of (3.8),

$$|F_1| \le m^{N(k)} e^{-k} (2 + \mathrm{diam}(\Lambda)) m^{-N(k)} = (2 + \mathrm{diam}(\Lambda)) e^{-k}. \tag{3.11}$$

It remains to estimate $|F_2|$. Suppose that $x \in F_2$. Then $x \in \Lambda^\theta_u(\rho)$ for some u containing u_1 as a subword. We have $u = v u_1 w$ for some (possibly empty)

words v and w. Then clearly $x \in \Lambda^{\theta}_{vu_1}(\rho)$. Recall that $S^{\theta}_{u_1}, \ldots, S^{\theta}_{u_{2^k}}$ are pairwise-1 close, hence $S^{\theta}_{vu_1}, \ldots, S^{\theta}_{vu_{2^k}}$ are pairwise-1 close as well. Let $q = |v| + |u_1|$. Of course, $q \leq N(k)$. It follows that the ball $B(x) := B(x, c_4 r^q)$, with $c_4 = 2 + \operatorname{diam}(\Lambda)$, contains all $\Lambda^{\theta}_{vu_j}$, for $j \leq 2^k$. Therefore, the natural measure ν_θ of the ball satisfies

$$\nu_\theta B(x) \geq 2^k m^{-q} = 2^{k-1} c_4^{-1} |B(x)|. \tag{3.12}$$

By a classical covering theorem (see [**15**, Theorem 2.1]), we can choose a disjoint family $\{B_j\}$ of the balls $\{B(x) : \ x \in F_2\}$ so that $F_2 \subset \bigcup_j 5B_j$. Thus,

$$|F_2| \leq 5 \sum_j |B_j| \leq 5c_4 2^{-(k-1)} \sum_j \nu_\theta B_j \leq 5c_4 2^{-(k-1)},$$

since ν_θ is a probability measure. Combining this estimate with (3.10) and (3.11) yields (3.9), with $\gamma = 2/e$, and the proof is complete. □

4. Proof of the lemmas.

Proof of Lemma 3.2. This is an easy "transversality argument", essentially contained in the proof of [**18**, Theorem 2.1(i)]. We provide a proof for the reader's convenience.

By increasing C_1 we can assume that n is sufficiently large. Let $\theta_0 \in \mathcal{I}P(\Lambda)$. This means, by definition, that $\operatorname{proj}_{\theta_0}|_\Lambda$ is not one-to-one, hence there exist $i \neq j$ such that $\Lambda^{\theta_0}_i \cap \Lambda^{\theta_0}_j \neq \emptyset$. Fix $\varepsilon > 0$ and $n \in \mathbb{N}$. There exist $u, v \in \mathcal{A}^n$ such that $u_1 = i, v_1 = j$, and $\Lambda^{\theta_0}_u \cap \Lambda^{\theta_0}_v \neq \emptyset$.

Recall that $S_u(\mathbf{x}) = r^n \mathbf{x} + b_u$, where $\mathbf{x}, b_u \in \mathbb{R}^2$, and $S^{\theta}_u(x) = r^n x + b^{\theta}_u$. Consider the function $f(\theta) = b^{\theta}_u - b^{\theta}_v = (b_u - b_v) \cdot (\cos\theta, \sin\theta)$. Observe that S^{θ}_u and S^{θ}_v are ε-relatively close if and only if $|f(\theta)| \leq \varepsilon r^n$. We have $|f(\theta)|^2 + |f'(\theta)|^2 = |b_u - b_v|^2$. Thus,

$$\eta^2 - |f(\theta)|^2 \leq |f'(\theta)|^2 \leq (\operatorname{diam}(\Lambda))^2 \quad \text{for} \ \ \theta \in [0, \pi], \tag{4.1}$$

where $\eta = \min\{\operatorname{dist}(\Lambda_p, \Lambda_q) : \ 1 \leq p < q \leq m\}$. Note that $\eta > 0$ by the Strong Separation Condition. Since $\Lambda^{\theta_0}_u \cap \Lambda^{\theta_0}_v \neq \emptyset$, we have $|f(\theta_0)| \leq \operatorname{diam}(\Lambda^{\theta_0}_u) \leq \operatorname{diam}(\Lambda) r^n$, which can be assumed less than $\frac{\eta}{2}$, since n is large. Then $|f'(\theta_0)| \geq \frac{\sqrt{3}\eta}{2} > \frac{\eta}{2}$ and it follows from (4.1) that there exists θ_1, with $|\theta_1 - \theta_0| \leq \frac{2}{\eta} \operatorname{diam}(\Lambda) r^n$, such that $f(\theta_1) = 0$. Then $S^{\theta_1}_u \equiv S^{\theta_1}_v$, and for all $\theta \in (\theta_1 - \frac{\varepsilon}{\operatorname{diam}(\Lambda)} r^n, \theta_1 + \frac{\varepsilon}{\operatorname{diam}(\Lambda)} r^n)$, by (4.1), the maps S^{θ}_u and S^{θ}_v are ε-relatively close. This implies that the interval $[\theta_0 - \frac{2}{\eta}\operatorname{diam}(\Lambda) r^n, \theta_0 + \frac{2}{\eta}\operatorname{diam}(\Lambda) r^n]$ contains a subinterval I' of length $\min\{\frac{\varepsilon}{\operatorname{diam}(\Lambda)}, \frac{4\operatorname{diam}(\Lambda)}{\eta}\} r^n$ which has the property (ii) from the statement of the lemma. The claim for an arbitrary interval $I \subset J$ now follows easily. □

Proof of Lemma 3.3. This proof is analogous to that of [**19**, Lemma 4.1].

Fix $\ell \in \mathbb{N}$ so that $r^\ell \le \frac{1}{2}(1-r)$ and ℓ_0 such that $C_1 r^{\ell_0} \le |J|$. We are going to construct inductively a family of compact sets $F_0 \supset F_1 \supset \cdots \supset F_n$, such that $|F_n| \le e^{-cn\varepsilon}$ for some $c > 0$ and F_n is a union of 2^n intervals, each of length at least $C_1 r^{\ell_0+\ell n}$. Most importantly, we will have that $F_n \supset \Psi(\ell_0 + \ell n, 2, \varepsilon)$. (Observe that $\Psi(k, 2, \varepsilon)$ are nested, decreasing with k, by the definition of these sets, so the desired estimate will follow.)

We can take $F_0 = J$. Suppose that we already have F_n, for some $n \ge 0$, and we need to construct F_{n+1}. Let I be any of the 2^n intervals of F_n and find $k \le n$ so that $C_1 r^k \le |I| < C_1 r^{k-1}$. By assumption, $k \le \ell_0 + \ell n$. Let I' be the subinterval of I of length $C_1 r^{k+1}$ with the same center. By Lemma 3.2, there is a subinterval $I'' \subset I'$ of length $\ge C_1^{-1}\varepsilon r^{k+1}$, which misses $\Psi(k+1, 2, \varepsilon) \supset \Psi(\ell_0 + \ell(n+1), 2, \varepsilon)$. Removing the interior of I'' makes two closed intervals out of I, each of length at least $\frac{1}{2}C_1(r^k - r^{k+1}) \ge C_1 r^{k+\ell} \ge C_1 r^{\ell_0+\ell(n+1)}$. In this way we construct F_{n+1}, a union of 2^{n+1} intervals. It remains to observe that

$$|I \setminus I''| \le |I| - C_1^{-1}\varepsilon r^{k+1} \le |I|(1 - C_1^{-2}\varepsilon).$$

Thus, $|F_{n+1}| \le (1 - C_1^{-2}\varepsilon)|F_n| \le e^{-C_1^2\varepsilon}|F_n|$, and the desired statement follows. □

Proof of Lemma 3.5. We are going to prove (3.4) by induction in i. We can assume that $c_3 \le c_2$; then the case $i = 1$ is just (3.2). Further, we can assume that $n \ge N_0$ and $\log^{i-1} n \ge M_0$ for any fixed constants N_0, M_0, since otherwise (3.4) holds trivially for $c_3 > 0$ sufficiently small.

Suppose that (3.4) holds for some $i \ge 1$. Then by (3.3) and (3.2),

$$\begin{aligned}(4.2)\quad &|\Psi(n, 2^{i+1}, \varepsilon)| \\ &\le |\Psi(\ell_0, 2, (\varepsilon/2)r^{j_0})| + |\Psi(j_0, 2^i, (\varepsilon/2))| \\ &\le M \exp[-c_2\ell_0(\varepsilon/2)r^{j_0}] + M2^{i-1}\exp[-c_3 e^{-(i-1)}(\log^{i-1} j_0)(\varepsilon/2)] \\ &=: M(A_1 + A_2),\end{aligned}$$

where $n = \ell_0 + j_0$. Let j_0 be the smallest integer $\ge \frac{1}{2}\frac{\log n}{|\log r|}$. Then we have for n sufficiently large:

$$\ell_0 = n - j_0 \ge n - \frac{1}{2}\frac{\log n}{|\log r|} - 1 \ge \frac{n}{2},$$

$$r^{j_0} \ge r^{\frac{1}{2}\frac{\log n}{|\log r|}-1} = r^{-1}n^{-1/2},$$

hence

$$(4.3)\qquad A_1 \le \exp[-c_2(n/2)r^{-1}n^{-1/2}(\varepsilon/2)] = \exp[-c_2(4r)^{-1}n^{1/2}\varepsilon].$$

Let $\alpha := 2|\log r|$. Note that $\alpha > 1$ since $r \le m^{-1} \le \frac{1}{2}$. Turning to A_2 in (4.2), we obtain from our choice of j_0:

$$A_2 \le 2^{i-1} \exp[-c_3 e^{-(i-1)} \log^{i-1}(\alpha^{-1} \log n) \cdot (\varepsilon/2)]. \tag{4.4}$$

If $i = 1$, then in (4.2) we could use (3.2) for the second summand as well, in which case

$$A_2 \le 2^{i-1} \exp[-c_2 \alpha^{-1} \log n \cdot (\varepsilon/2)] \le 2^{i-1} \exp[-c_3 \log n \cdot (\varepsilon/2)], \tag{4.5}$$

assuming $c_3 \le c_2 \alpha^{-1}$. For $i \ge 2$ we use the elementary inequality

$$\log(x + y) \le \log x + y \quad \text{for all } \ x \ge 1,\ y \ge 0. \tag{4.6}$$

We can assume that $\log^{i-1} n \ge \log \alpha + 1$, since otherwise (3.4) holds trivially for $c_3 > 0$ sufficiently small. Then applying (4.6) $i - 2$ times to $\log^2 n = \log(\alpha^{-1} \log n) + \log \alpha$ we obtain

$$\log^i n \le \log^{i-1}(\alpha^{-1} \log n) + \log \alpha.$$

Combining this with (4.4) and (4.5) yields

$$A_2 \le 2^{i-1} \exp[-c_3 e^{-(i-1)}(\log^i n - \log \alpha)(\varepsilon/2)] \quad \text{for all } \ i \ge 1.$$

In view of (4.3), the induction step will be finished once we check the inequality

$$\begin{aligned} &\exp[-c_2 (4r)^{-1} n^{1/2} \varepsilon] + 2^{i-1} \exp[-c_3 e^{-(i-1)}(\log^i n - \log \alpha)(\varepsilon/2)] \\ &\quad \le 2^i \exp[-c_3 e^{-i}(\log^i n)\varepsilon]. \end{aligned} \tag{4.7}$$

This is equivalent to

$$\begin{aligned} 1 &\ge 2^{-i} \exp[(c_3 e^{-i} \log^i n - c_2 (4r)^{-1} n^{1/2})\varepsilon] \\ &\quad + 2^{-1} \exp[c_3 \varepsilon e^{-(i-1)}(\log^i n \cdot (e^{-1} - 2^{-1}) + 2^{-1} \log \alpha)] =: B_1 + B_2. \end{aligned}$$

We have

$$c_3 e^{-i} \log^i n - c_2 (4r)^{-1} n^{1/2} \le c_2 \log n - c_2 (4r)^{-1} n^{1/2} < 0$$

for n sufficiently large, hence $B_1 \le 2^{-i}$. Further, we can assume that

$$\log^i n > \log \alpha \cdot (1/2 - e^{-1})^{-1};$$

then $B_2 \le \frac{1}{2}$. This implies (4.7), and the proof of the lemma is complete. □

Proof of Lemma 3.6. It follows from (4.6) and (1.1) that

$$\log_*(x + y) \le \log_* x + \log_*(1 + y) \quad \text{for all } \ x \ge 1,\ y \ge 0.$$

Using this inequality, (3.7), and (3.5), we obtain

$$\begin{aligned}\log_*(r^{-N(k)}) &\le 1+\log_*(\log(r^{-1})+\log N(k))\\ &\le \text{const}+\log_*(n_k\log m+\log n_k+\log k)\\ &\le \text{const}+\log_* n_k\\ &= \text{const}+(k-1)+\log_*(c_3^{-1}ke^{k-1})\\ &\le \text{const}+k+\log_* k.\end{aligned}$$

Now the desired statement is immediate. □

5. Non-homogeneous case.

Here we prove Theorem 2.1 in full generality and Theorem 2.3. The proofs follow the same path most of the way. We use the same notation as in Section 3, as much as possible, so the same letters often represent different but analogous objects here and there.

We have $S_u(\mathbf{x}) = r_u\mathbf{x} + b_u$ for some $b_u \in \mathbb{R}^2$, where $r_u = r_{u_1}\cdot\ldots\cdot r_{u_n}$. The natural projection map onto Λ^θ is defined by $\Pi_\theta(\omega) = \lim_{n\to\infty} S^\theta_{\omega|_n}(0)$. The natural measure on Λ^θ is $\nu_\theta = \mu\circ\Pi_\theta^{-1}$, where $\mu = (r_1^s,\ldots,r_m^s)^{\mathbb{N}}$ and $s \le 1$ is the similarity dimension of Λ. For $\delta > 0$ consider the "cut-set" $\mathcal{W}(\delta) = \{u\in\mathcal{A}^* :\ r_u \le \delta,\ r_{u'} > \delta\}$ where u' is obtained from u by dropping the last symbol. Let $r_{\min} = \min\{r_i :\ i\le m\}$ and $r_{\max} = \max\{r_i :\ i\le m\}$. For $u, v \in \mathcal{W}(\delta)$ we have $r_{\min} \le r_u/r_v \le r_{\min}^{-1}$. Throughout this section we fix a nonempty interval $J \subset \mathcal{I}P(\Lambda)$ (assuming that it exists). Let $X_\Lambda = [-d_\Lambda, d_\Lambda]$ where $d_\Lambda = \max\{|\mathbf{x}| :\ \mathbf{x}\in\Lambda\}$. Observe that $S_i^\theta(X_\Lambda) \subset X_\Lambda$ for all $i \le m$ and all $\theta \in [0,\pi]$.

Definition 5.1. Let $\theta\in[0,\pi]$ and $u, v \in \mathcal{A}^*$. We say that S_u^θ and S_v^θ are **ε-relatively close at** x if

$$|S_u^\theta(x) - S_v^\theta(x)| \le \varepsilon\min\{r_u, r_v\}.$$

Lemma 5.2. *There exists $C_1 \ge 1$ such that for all $x\in X_\Lambda$, for all $\varepsilon\in(0,1]$ and all $\delta > 0$, for any interval $I\subset J$, with $|I| = C_1\delta$, there is a subinterval $I'\subset I$ such that $|I'| \ge C_1^{-2}\varepsilon|I|$ and for every $\theta\in I'$ there exist $u\ne v$ in $\mathcal{W}(\delta)$ such that S_u^θ and S_v^θ are ε-relatively close at x.*

Proof. The proof is analogous to the proof of Lemma 3.2. Let $g(\theta) = S_u^\theta(x) - S_v^\theta(x) = (r_u - r_v)x + f(\theta)$ where $f(\theta) = b_u^\theta - b_v^\theta$. If $\theta_0 \in \mathcal{I}P(K)$, then $|g(\theta_0)| \le 4\delta d_\Lambda$ which can be assumed small, increasing C_1 if necessary. Since $g'(\theta) = f'(\theta)$, the rest of the proof of Lemma 3.2 transfers. □

Notation. For $x\in X_\Lambda$, $k\ge 2$, $\varepsilon\in(0,1]$ and $n\ge 1$ denote by $\Phi(n,k,x,\varepsilon)$ the set of $\theta\in J$ such that there is no collection of distinct words $u_1,\ldots,u_k$, with $r_{u_j} \ge r_{\max}^n$, such that $S^\theta_{u_j}$, $j\le k$, are pairwise ε-relatively close at x. Denote by $\Phi'(n,k,x,\varepsilon)$ the analogous set where it is required, in addition,

that $r_{\min} \le r_{u_i}/r_{u_j} \le r_{\min}^{-1}$ for $i, j \le k$. By the definition, $\Phi'(n,k,x,\varepsilon) \supset \Phi(n,k,x,\varepsilon)$. Further, let

$$\Psi(n,k,\varepsilon) = \bigcup_{x\in X_\Lambda} \Phi(n,k,x,\varepsilon) \quad \text{and} \quad \Psi'(n,k,\varepsilon) = \bigcup_{x\in X_\Lambda} \Phi'(n,k,x,\varepsilon).$$

Lemma 5.3. *There exist $c_2 > 0$ and $M > 0$ such that*

$$|\Psi'(n,2,\varepsilon)| \le M\varepsilon^{-1}e^{-c_2 n\varepsilon} \quad \textit{for all } \ n \in N,\ \varepsilon \in (0,1]. \tag{5.1}$$

Proof. Using Lemma 5.2 and repeating the proof of Lemma 3.3, we obtain for any fixed $x \in X_\Lambda$ that

$$|\Phi'(n,2,x,\varepsilon)| \le \widetilde{M}e^{-\widetilde{c}_1 n\varepsilon} \tag{5.2}$$

for some constants $\widetilde{M}, \widetilde{c}_1 > 0$. Observe that if S_u^θ and S_v^θ are $\frac{\varepsilon}{2}$-relatively close at x and $C^{-1} \le r_u/r_v \le C$, then S_u^θ and S_v^θ are ε-relatively close at x', provided that $|x'-x| \le \frac{\varepsilon}{2C}$. Taking $C = r_{\min}^{-1}$ and choosing an $\frac{\varepsilon}{2C}$-net $\mathcal{N}$ of X_Λ, we obtain

$$\Psi'(n,2,\varepsilon) \subset \bigcup_{x\in\mathcal{N}} \Phi'(n,2,x,\varepsilon/2).$$

Since $\#(\mathcal{N}) \le \text{const} \cdot \varepsilon^{-1}$, this and (5.2) yield (5.1). □

Lemma 5.4. *There exists $C \ge 1$ such that for $n = \ell_0 + j_0$, with $\ell_0, j_0 \ge 1$, and $k \ge 2$, we have*

$$\Psi(n,2k,\varepsilon) \subset \Psi(j_0,k,C^{-1}\varepsilon) \cup \Psi'(\ell_0,2,C^{-1}r_{\max}^{j_0}\varepsilon). \tag{5.3}$$

Proof. Fix $x_0 \in X_\Lambda$. We want to show that $\Phi(n,2k,x_0,\varepsilon)$ lies in the right-hand side of (5.3). Suppose that θ is not in the right-hand of (5.3). Then there exist distinct words $w_1, \dots, w_k$, with $r_{w_i} \ge r_{\max}^{j_0}$, such that $S_{w_i}^\theta$ are pairwise $C^{-1}\varepsilon$-relatively close at x_0. Without loss of generality, suppose that $r_{w_1} = \min_{i\le k} r_{w_i}$. Further, $\theta \notin \Psi'(\ell_0,2,C^{-1}r_{\max}^{j_0}\varepsilon) = \bigcup_{x\in X} \Phi'(\ell_0,2,x, C^{-1}r_{\max}^{j_0}\varepsilon)$, so there exist distinct u_1, u_2 such that $r_{u_i} \ge r_{\max}^{\ell_0}$, $r_{\min} \le r_{u_1}/r_{u_2} \le r_{\min}^{-1}$, and $S_{u_1}^\theta$ and $S_{u_2}^\theta$ are $C^{-1}r_{\max}^{j_0}\varepsilon$-relatively close at $S_{w_1}^\theta(x_0) \in X_\Lambda$. Then $u_i w_j$, for $i = 1,2$ and $j \le k$, are all distinct and satisfy $r_{u_i w_j} \ge r_{\max}^{\ell_0+j_0} = r_{\max}^n$. We claim that $S_{u_i w_j}^\theta$ are pairwise ε-close at x_0 if C is sufficiently large. This will imply that $\theta \notin \Phi(n,2k,x_0,\varepsilon)$, and since x_0 is arbitrary, the lemma will be proved.

We have for $i = 1,2$ and for all $j \le k$,

$$|S_{u_i w_1}^\theta(x_0) - S_{u_i w_j}^\theta(x_0)| \le r_{u_i} C^{-1}\varepsilon \min\{r_{w_1}, r_{w_j}\} = C^{-1}\varepsilon r_{u_i w_1}.$$

Further,

$$|S_{u_1 w_1}^\theta(x_0) - S_{u_2 w_1}^\theta(x_0)| \le C^{-1} r_{\max}^{j_0}\varepsilon \min\{r_{u_1}, r_{u_2}\} \le C^{-1}\varepsilon \min\{r_{u_1 w_1}, r_{u_2 w_1}\}.$$

Therefore, for $1 \le p < q \le k$,

$$\begin{aligned} |S^\theta_{u_1 w_p}(x_0) - S^\theta_{u_2 w_q}(x_0)| &\le C^{-1}\varepsilon r_{w_1}(r_{u_1} + r_{u_2} + \min\{r_{u_1}, r_{u_2}\}) \\ &\le C^{-1}\varepsilon r_{w_1}(2 + r_{\min}^{-1}) \min\{r_{u_1}, r_{u_2}\} \\ &\le C^{-1}\varepsilon(2 + r_{\min}^{-1}) \min\{r_{u_1 w_p}, r_{u_2 w_q}\}, \end{aligned}$$

and the claim follows with $C = 2 + r_{\min}^{-1}$. The lemma is proved. □

Lemma 5.5. *There exist $a > 0$ and $b > 1$ such that for all $\varepsilon \in (0, 1]$, $n \in \mathbb{N}$ and $i \ge 1$,*

$$|\Psi(n, 2^i, \varepsilon)| \le M\varepsilon^{-1} b^i \exp[-ae^{-(i-1)}(\log^{i-1} n)\varepsilon]. \tag{5.4}$$

Proof is analogous to the proof of Lemma 3.5, based on Lemmas 5.3 and 5.4. We leave the details to the reader. □

Let

$$n_k = \exp^{k-1}((\log b + 1)a^{-1}ke^{k-1}), \tag{5.5}$$

so that, in view of (5.4),

$$|\Psi(n_k, 2^k, 1)| \le Me^{-k}. \tag{5.6}$$

Let

$$N(k) = n_k \cdot \lceil r_{\min}^{-s} \rceil^{n_k} \cdot k. \tag{5.7}$$

Similarly to the proof of Lemma 3.6, we deduce from (5.5) and (5.7) that

$$\log_*(r_{\min}^{-N(k)}) \le C_\xi + (1 + \xi)k, \tag{5.8}$$

for any $\xi > 0$. For any $u_1 \in \mathcal{A}^*$, with $|u_1| \le n_k$, we have

$$\sum_{\substack{|u|=N(k) \\ u_1 \not\sqsubset u}} r_u^s \le (1 - r_{\min}^{n_k s})^{N(k)/n_k} \le e^{-k}. \tag{5.9}$$

Now suppose that $\theta \notin \Psi(n_k, 2^k, 1)$ and $x_0 \in \Lambda^\theta$. Then $\theta \notin \Phi(n_k, 2^k, x_0, 1)$, so we can find distinct words $u_1, \dots, u_{2^k}$, with $r_{u_i} \ge r_{\max}^{n_k}$, such that

$$|S_{u_i}(x_0) - S_{u_j}(x_0)| \le \min\{r_{u_i}, r_{u_j}\} \quad \text{for all } \ i, j \le 2^k. \tag{5.10}$$

Without loss of generality, assume that $r_{u_1} = \min\{r_{u_i} : \ i \le 2^k\}$. We have

$$\Lambda^\theta = \bigcup_{|u|=N(k)} \Lambda^\theta_u = \bigcup_{\substack{|u|=N(k) \\ u_1 \not\sqsubset u}} \Lambda^\theta_u \cup \bigcup_{\substack{|u|=N(k) \\ u_1 \sqsubset u}} \Lambda^\theta_u =: Y_{u_1} \cup Z_{u_1}. \tag{5.11}$$

We claim that for some $C \ge 1$,

$$\forall\, x \in Z_{u_1},\ \exists\, t \in [C^{-1} r_{\min}^{N(k)}, C r_{u_1}] :\ \nu_\theta B(x, t) \ge C^{-1} 2^k t^s. \tag{5.12}$$

Indeed, suppose that $x \in \Lambda^\theta_u$ for some $u \in \mathcal{A}^{N(k)}$ such that $u_1 \sqsubset u$. Then $u = vu_1 w$ for some (possibly empty) words v and w. Let $\omega \in \mathcal{A}^{\mathbb{N}}$ be such that $x_0 = \Pi_\theta(\omega)$. For each u_j, with $2 \le j \le 2^k$, there exists a unique

$q = q_j \in \mathbb{N} \cup \{0\}$ such that $\widetilde{u}_j := u_j\omega|_q \in \mathcal{W}(r_{u_1})$. Notice that $S^\theta_{u_j}(x_0) \in \Lambda^\theta_{\widetilde{u}_j}$ whence $S^\theta_{vu_j}(x_0) \in \Lambda^\theta_{v\widetilde{u}_j}$. By (5.10), we have $|S^\theta_{vu_j}(x_0) - S^\theta_{vu_1}(x_0)| \le r_{vu_1}$. Finally, $x \in \Lambda^\theta_{vu_1}$, which implies that the distance from x to $\Lambda_{v\widetilde{u}_j}$ is at most $\operatorname{diam}(\Lambda^\theta_{vu_1}) + r_{vu_1}$. Since $r_{v\widetilde{u}_j} = r_v r_{\widetilde{u}_j} \le r_{vu_1}$, we obtain that

$$B(x, C' r_{vu_1}) \supset \Lambda^\theta_{vu_1} \cup \bigcup_{j=2}^{2^k} \Lambda^\theta_{v\widetilde{u}_j}, \quad \text{where } C' = 1 + 2\operatorname{diam}(\Lambda). \tag{5.13}$$

Therefore,

$$\nu_\theta B(x, C' r_{vu_1}) \ge r^s_{vu_1} + \sum_{j=2}^{2^k} r^s_{v\widetilde{u}_j} \ge 2^k r^{-s}_{\min} r^s_{vu_1}. \tag{5.14}$$

This implies (5.12) since $r^{N(k)}_{\min} \le r_u \le r_{vu_1} \le r_{u_1}$.

Proof of Theorem 2.1. Recall that now $s = 1$. By (5.8), it suffices to show that for some $c > 0$ and $\gamma \in (0,1)$, we have

$$\operatorname{Fav}(\Lambda(\rho)) \le c\gamma^k, \quad \text{where } \rho := r^{N(k)}_{\min}.$$

In view of (5.6), it is sufficient to estimate $|\Lambda^\theta(\rho)|$ from above for $\theta \notin \Psi(n_k, 2^k, 1)$. Fix such a θ, $x_0 \in \Lambda^\theta$, and the words $u_1, \dots, u_{2^k}$ as before, satisfying (5.10), and let u_1 be the word with the minimal r_{u_i}. By (5.11) we have $\Lambda^\theta(\rho) = Y_{u_1}(\rho) \cup Z_{u_1}(\rho)$. Clearly, $\operatorname{diam}(\Lambda^\theta_u(\rho)) \le (2 + \operatorname{diam}(\Lambda)) r_u$ for any $u \in \mathcal{A}^{N(k)}$, so (5.9), with $s = 1$, implies that $|Y_{u_1}(\rho)| \le \text{const} \cdot e^{-k}$. Since $t \ge C^{-1}\rho$ in (5.12), the balls $B(x, (1+C)t)$, for $x \in Z_{u_1}$, cover the ρ-neighborhood $Z_{u_1}(\rho)$. Now (5.12) implies $|Z_{u_1}(\rho)| \le \text{const} \cdot 2^{-k}$, by repeating the argument at the end of Section 3, and the proof is finished. □

Proof of Theorem 2.3. We use the same setting as in the proof of Theorem 2.1, except that now $s \le 1$ and $J \subset \mathcal{I}P(\Lambda)$ is a nonempty interval. In view of (5.6), the Borel-Cantelli Lemma implies that the set

$$E := \bigcup_{n=1}^{\infty} \bigcap_{k \ge n} (J \setminus \Psi(n_k, 2^k, 1))$$

has full Lebesgue measure in J. Thus, it is enough to show that $\mathcal{H}^\phi(\Lambda^\theta) = 0$ for all $\theta \in E$.

Suppose that $\theta \in E$; then $\theta \in J \setminus \Psi(n_k, 2^k, 1)$ for all k sufficiently large. We fix $x_0 \in \Lambda^\theta$ and find $u_1 = u_1(k)$ as above (now we have to make the dependence on k explicit). For $\rho_k = r^{N(k)}_{\min}$ we have the decomposition (5.11) $\Lambda^\theta = Y_{u_1(k)} \cup Z_{u_1(k)}$. We can write $\Lambda^\theta = \Omega_1 \cup \Omega_2$ where Ω_1 is the set of x which belong to infinitely many $Y_{u_1(k)}$ and Ω_2 is the set of x which

belong to all $Z_{u_1(k)}$ for k sufficiently large. Recall that $Y_{u_1(k)} = \bigcup_{\substack{|u|=N(k)\\ u_1(k)\not\subset u}} \Lambda_u^\theta$ and $\operatorname{diam}(\Lambda_u^\theta) \le \operatorname{diam}(\Lambda)\cdot r_u$. Thus,

$$\begin{aligned}
\mathcal{H}^\phi(\Omega_1) &\le \text{const}\cdot \lim_{k\to\infty} \sum_{\substack{|u|=N(k)\\ u_1(k)\not\subset u}} \phi(r_u)\\
&= \text{const}\cdot \lim_{k\to\infty} \sum_{\substack{|u|=N(k)\\ u_1(k)\not\subset u}} r_u^s \exp[L\log_*(r_u^{-1})]\\
&\le \text{const}\cdot \lim_{k\to\infty} \sum_{\substack{|u|=N(k)\\ u_1(k)\not\subset u}} r_u^s \exp[L\log_*(\rho_k^{-1})]\\
&\le \text{const}\cdot \lim_{k\to\infty} e^{-k} e^{L(1+\xi)k} = 0,
\end{aligned}$$

using (5.9) and (5.8), with $0 < \xi < L^{-1} - 1$, in the last estimate. Recall that $L < \log 2 < 1$.

It remains to prove that $\mathcal{H}^\phi(\Omega_2) = 0$. For any $x \in Z_{u_1(k)}$ we have by (5.12), with $t = t_k \ge C^{-1}\rho_k$,

$$\begin{aligned}
\frac{\nu_\theta B(x,t_k)}{\phi(t_k)} &\ge \frac{\text{const}\cdot 2^k t_k^s}{t_k^s \exp[L\log_*(2t_k^{-1})]}\\
&\ge \text{const}\cdot 2^k \exp[-L\log_*(2C\rho_k^{-1})]\\
&\ge \text{const}\cdot 2^k e^{-L(1+\xi)k} \to \infty, \quad \text{as } k\to\infty.
\end{aligned} \tag{5.15}$$

In the last line we used (5.8) with $0 < \xi < L^{-1}\log 2 - 1$. Notice that $t_k \le r_{u_1(k)} \to 0$, as $k\to\infty$ (since $r_{u_1(k)}$ is the smallest among $r_{u_i(k)}$, $i \le 2^k$, and all $u_i(k)$ are distinct). Thus, (5.15) implies

$$\overline{D}_\phi(\nu_\theta, x) := \limsup_{t\to 0} \frac{\nu_\theta B(x,t)}{\phi(2t)} = \infty \quad \text{for all } \ x \in \Omega_2,$$

and hence $\mathcal{H}^\phi(\Omega_2) = 0$ by the Rogers-Taylor Density Theorem, see [**20**]. The proof of Theorem 2.3 is complete. □

6. Random Cantor sets.

The proof of Theorem 2.2 is inspired by an argument of Lyons [**11**] involving percolation on trees; the negative dependence in the construction of $\mathcal{R}_k$ that arises from choosing exactly one of the four dyadic subsquares in the inductive step of the construction, makes the proof here a little more delicate.

Denote by $\mathcal{G}_k$ the collection of 4^k (closed) dyadic subsquares of the unit square $[0,1]^2$ having side length 2^{-k}. We consider all dyadic subsquares as a rooted tree, with $[0,1]^2$ being the root and $\mathcal{G}_k$ being the set of nodes at the

kth level. For each node there are four edges leading to nodes at the next level, (its "children").

Let ℓ be a line intersecting $[0,1]^2$, that does not go through any of the vertices of the squares in $\mathcal{G}_{2n}$. Further, let

$$A_n(\ell) = \#\{B \in \mathcal{G}_{2n} : \ B \cap \ell \neq \emptyset\}.$$

Observe that

$$A_n(\ell) \leq 2^{2n+1}. \tag{6.1}$$

To verify this we may assume, using symmetry, that ℓ forms an angle $\alpha \in [0, \pi/4]$ with the horizontal. Then ℓ intersects at most two squares in each of the 2^{2n} columns of $\mathcal{G}_{2n}$, and (6.1) follows.

Below $\mathbf{P}(E)$ denotes the probability of an event E.

Lemma 6.1. *Suppose that the line ℓ does not hit any vertices of the squares in $\mathcal{G}_{2n}$. Then*

$$\mathbf{P}(\mathcal{R}_n \cap \ell \neq \emptyset) \leq \frac{C_1}{n} \tag{6.2}$$

for some constant $C_1 > 0$ independent of ℓ and n.

Proof of Theorem 2.2 *assuming Lemma* 6.1. Let $\theta \in [0, \pi]$ be such that the line $y \cos\theta = x \sin\theta$ is orthogonal to ℓ, and let $\mathbf{n}$ be the unit normal vector for ℓ. Then by Fubini's Theorem and Lemma 6.1,

$$\mathbf{E}\left[|\mathrm{proj}_\theta \mathcal{R}_n|\right] = \int_{\mathbb{R}} \mathbf{P}(\mathcal{R}_n \cap (\ell + t\mathbf{n}) \neq \emptyset)\, dt \leq \sqrt{2}\frac{C_1}{n}\,, \tag{6.3}$$

and (2.4) follows by integrating over θ.

Finally, (2.5) follows directly from (2.4) by Fatou's lemma. □

Proof of Lemma 6.1. We label the four dyadic subsquares of a square as in Figure 3.

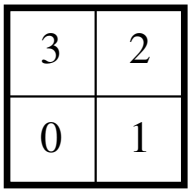


Figure 3. Labeling subsquares.

This labeling induces a natural addressing scheme for each dyadic square $B \in \mathcal{G}_k$. The address has length k and the symbols are from $\{0,1,2,3\}$; we write it as $\omega(B) = \{\omega_i(B)\}_{i=1}^k$. Recall that we arrange all dyadic squares in a tree. The construction of the random set is such that at even levels we take all children, but at odd levels we choose for each remaining square

one child, uniformly at random and independently of the choices in other squares. This yields a subtree of the full 4-ary tree, where the nodes at level $2n$ correspond to the random set $\mathcal{R}_n$.

By symmetry, we may assume that the slope of ℓ is positive.

Fix a small positive constant δ, to be chosen later. We subdivide $\mathcal{G}_{2n}$ into three types as follows:

(i) Say that $B \in \mathcal{G}_{2n}$ is **Type 1** if

$$\#\{i \le n-1 :\ \omega_{2i+1}(B) = 0\} \ge \delta n.$$

(ii) Say that $B \in \mathcal{G}_{2n}$ is **Type 2** if it is not Type 1, and

$$\#\{i \le n-1 :\ \omega_{2i+1}(B) = 2\} \ge \delta n.$$

(iii) All remaining $B \in \mathcal{G}_{2n}$ are said to be **Type 3**.

Consider the events

$$Z_i = \Big\{\exists\, B \subset \mathcal{R}_n :\ B \in \text{Type } i \ \&\ B \cap \ell \ne \emptyset\Big\} \quad \text{for} \ \ i = 1, 2, 3.$$

We have

$$\mathbf{P}(\mathcal{R}_n \cap \ell \ne \emptyset) \le \sum_{i=1}^{3} \mathbf{P}(Z_i).$$

First we estimate $\mathbf{P}(Z_1)$. We have

$$\mathbf{E}\left[\#\{B \subset \mathcal{R}_n :\ B \cap \ell \ne \emptyset\} \,|\, Z_1\right] \le \frac{\mathbf{E}\left[\#\{B \subset \mathcal{R}_n :\ B \cap \ell \ne \emptyset\}\right]}{\mathbf{P}(Z_1)}. \tag{6.4}$$

Writing

$$\#\{B \subset \mathcal{R}_n :\ B \cap \ell \ne \emptyset\} = \sum_{B \in \mathcal{G}_{2n}} \mathbf{1}_{\{B \subset \mathcal{R}_n :\ B \cap \ell \ne \emptyset\}}$$

and using that $\mathbf{P}(B \subset \mathcal{R}_n) = 4^{-n}$ for any $B \in \mathcal{G}_{2n}$, we obtain by (6.1) that

$$\mathbf{E}\Big[\#\{B \subset \mathcal{R}_n :\ B \cap \ell \ne \emptyset\}\Big] = A_n(\ell) \cdot 4^{-n} \le 2. \tag{6.5}$$

Thus, it remains to estimate the left-hand side of (6.4) from below. Let

$$\Psi_1 := \{Q \in \mathcal{G}_{2n} :\ Q \in \text{Type } 1 \ \&\ Q \cap \ell \ne \emptyset\}.$$

Order the squares in $\mathcal{G}_{2n}$ hit by ℓ from left to right and from bottom to top. This is a total order by the assumption on slope of the line ℓ. For $Q \in \Psi_1$ consider the event

$$Y_Q = \Big\{Q \ \text{ is the first square in } \Psi_1 \text{ hit by } \ell\Big\}.$$

Then $Z_1 = \bigcup_{Q \in \Psi_1} Y_Q$ is a disjoint union, and so, for any random variable f,

$$\mathbf{E}\left[f \,|\, Z_1\right] = \sum_{Q \in \Psi_1} \frac{\mathbf{P}(Y_Q)}{\mathbf{P}(Z_1)} \mathbf{E}\left[f \,|\, Y_Q\right] \ge \min_{Q \in \Psi_1} \mathbf{E}\left[f \,|\, Y_Q\right]. \tag{6.6}$$

Fix $Q \in \Psi_1$. We have

$$\mathbf{E}\left[\#\{B \subset \mathcal{R}_n : \ B \cap \ell \neq \emptyset\} \,|\, Y_Q\right] = \sum_{B \in \mathcal{G}_{2n} :\ B \cap \ell \neq \emptyset} \mathbf{P}(B \subset \mathcal{R}_n \,|\, Y_Q). \tag{6.7}$$

By the definition of Type 1 squares,

$$\#\{i \leq n-1 : \ \omega_{2i+1}(Q) = 0\} \geq \delta n\,.$$

Fix i such that $\omega_{2i+1}(Q) = 0$, and denote by $\widetilde{Q}$ the dyadic square in $\mathcal{G}_{2i}$ having the address $\omega(\widetilde{Q}) = \omega_1(Q)\ldots w_{2i}(Q)$. The fact that $Q \subset \mathcal{R}_n$ implies that $\widetilde{Q}$ was chosen at the ith stage of the random construction, *i.e.*, $\widetilde{Q} \subset \mathcal{R}_i$. (Note that by definition, $[0,1]^2 \supset \mathcal{R}_1 \supset \ldots \supset \mathcal{R}_n$.) Since the slope of ℓ is positive, ℓ intersects at least $\frac{1}{2}4^{n-i}$ squares $B \in \mathcal{G}_{2n}$ whose addresses start with $\omega(\widetilde{Q})k$, for $k \in \{1,2,3\}$ (see Figure 3). For each of these squares we have (using the independence of Y_Q from the random choices involving the descendants of $\omega(\widetilde{Q})k$ with $k \in \{1,2,3\}$), that

$$\mathbf{P}(B \subset \mathcal{R}_n \,|\, Y_Q) = \mathbf{P}(B \subset \mathcal{R}_n \,|\, \widetilde{Q} \subset \mathcal{R}_i) = 4^{i-n}.$$

Therefore, the sum of $\mathbf{P}(B \subset \mathcal{R}_n \,|\, Y_Q)$ over the set of squares

$$\mathcal{B}_i = \Big\{B \in \mathcal{G}_{2n}, : \ B \cap \ell \neq \emptyset, \ \ \{\omega_j(B)\}_1^{2i} = \{\omega_j(Q)\}_1^{2i}, \ \ \omega_{2i+1}(B) \in \{1,2,3\}\Big\},$$

is at least $\frac{1}{2}4^{n-i} \cdot 4^{i-n} = \frac{1}{2}$. Notice that the sets $\mathcal{B}_i$ are disjoint for distinct i with $\omega_{2i+1}(Q) = 0$. Thus, the right-hand side of (6.6) is at least $\frac{1}{2}\delta n$, which, together with (6.7), (6.6), (6.5) and (6.4), implies

$$\mathbf{P}(Z_1) \leq \frac{4}{\delta n}. \tag{6.8}$$

By symmetry, we obtain

$$\mathbf{P}(Z_2) \leq \frac{4}{\delta n}. \tag{6.9}$$

It remains to estimate $\mathbf{P}(Z_3)$. We have

$$\begin{aligned} \mathbf{P}(Z_3) &\leq \mathbf{E}\left[\#\{B \subset \mathcal{R}_n : \ B \in \text{Type 3 } \&\ B \cap \ell \neq \emptyset\}\right] \\ &= \sum_{B \in \text{Type 3} \,:\, B \cap \ell \neq \emptyset} \mathbf{P}(B \subset \mathcal{R}_n) \\ &= 4^{-n}\#\{B \in \text{Type 3} \ : \ B \cap \ell \neq \emptyset\}. \end{aligned} \tag{6.10}$$

Thus, it suffices to bound the number of Type 3 squares hit by ℓ. Consider the subtree of all dyadic squares that are hit by ℓ. Since we assumed that ℓ does not hit any vertices, it can hit at most three children of a dyadic square that it intersects. For a Type 3 square, at least $n - 2\delta n$ of the digits at odd levels are either 1 or 3, and our assumption that the slope of ℓ is positive

guarantees that it cannot intersect both of the children labeled by 1 and 3 of any dyadic square (see Figure 3). Therefore, summing over the number

$$j = \#\Big\{i \le n-1 :\ \omega_{2i+1}(B) \in \{0,2\}\Big\},$$

we obtain

$$\#\{B \in \text{Type 3} :\ B \cap \ell \ne \emptyset\} \le \sum_{j \le 2\delta n} \left(\begin{array}{c} n \\ j \end{array} \right) 3^n 2^j \le C_2 \cdot (1+\varepsilon(\delta))^n 3^{n+2\delta n},$$

where $\varepsilon(\delta) \to 0$, as $\delta \to 0$. Now we can choose δ so that $(1+\varepsilon(\delta))\cdot 3^{1+2\delta} < 3.5$, and, in view of (6.10),

$$\mathbf{P}(Z_3) \le \text{const} \cdot (7/8)^n.$$

Combining this with (6.8) and (6.9) yields (6.2), and the proof is complete. □

7. Concluding remarks and problems.

7.1. More general families of self-similar sets. Theorems 2.1 and 2.3 extend to parametrized families of self-similar sets satisfying the "transversality condition." The following set-up is taken from [**18**].

Let $J \subset \mathbb{R}$ be a closed interval. Consider a one-parameter family of iterated function systems $\{S_1^\lambda, \dots, S_m^\lambda\}_{\lambda \in J}$ where $S_i^\lambda(x) = r_i x + a_i(\lambda)$, with $r_i \in (0,1)$ and $a_i(\lambda) \in C^1(J)$. Let $\Pi(\lambda, \cdot) :\ \mathcal{A}^{\mathbb{N}} \to \mathbb{R}$ be the natural projection map associated with the system and let $\Lambda^\lambda = \Pi(\lambda, \mathcal{A}^{\mathbb{N}})$. Then $\{\Lambda^\lambda\}_{\lambda \in J}$ is a family of self-similar sets on the real line. Note that the similarity dimension s does not depend on λ. We denote $f_{\omega,\tau}(\lambda) = \Pi(\lambda,\omega) - \Pi(\lambda,\tau)$ and say that the **transversality condition** holds on J if for any $\omega, \tau \in \mathcal{A}^{\mathbb{N}}$,

$$\text{if } \ \exists\, \lambda \in J :\ f_{\omega,\tau}(\lambda) = f'_{\omega,\tau}(\lambda) = 0 \quad \text{then} \quad \omega = \tau.$$

Define

$$\mathcal{I}P = \{\lambda \in J :\ \exists\, \omega, \tau \in \mathcal{A}^{\mathbb{N}} :\ f_{\omega,\tau}(\lambda) = 0 \ \text{ but } \ \omega \ne \tau\}.$$

Theorem 7.1. *Suppose that the one-parameter family of iterated function systems defined above satisfies the transversality condition and $\mathcal{I}P = J$.*

(i) *Assume that $s = 1$. Then there exist $C, a > 0$ such that*

$$\int_J |\Lambda^\lambda(\rho)|\, d\lambda \le C \exp[-a \log_*(\rho^{-1})] \quad \textit{for all} \ \ \rho > 0.$$

(ii) *Assume that $s \le 1$. Then $\mathcal{H}^\phi(\Lambda^\lambda) = 0$ for Lebesgue-a.e. $\lambda \in J$ where $\phi(t) = t^s \exp[L \log_*(t^{-1})]$, with $L \in (0, \log 2)$.*

The proof of this theorem is very similar to the proofs of Theorems 2.1 and 2.3. The only change is in Lemma 3.2, where one needs to use the general form of transversality rather than the special form (4.1) valid for projection families. In [**18**] it is proved, under the assumptions of Theorem 7.1(ii), that $\mathcal{H}^s(\Lambda^\lambda) = 0$ for a.e. $\lambda \in J$.

Example. Let $\Lambda^\lambda = \{\sum_{n=0}^{\infty} a_n 4^{-n} : \ a_n \in \{0, 1, 2, \lambda\}\}$. Then all the assumptions of Theorem 7.1 hold for $\lambda \in [0, 3]$.

7.2. Cantor sets with varying contraction ratios. Let $D = \{b_1, \dots, b_m\} \subset \mathbb{R}^2$ be a digit set. Suppose that $\delta_n \geq 0$ and let

$$r^{(n)} = m^{-n} \prod_{i=1}^{n} (1 + \delta_i) \quad \text{for} \ \ n \geq 1.$$

Define $\Pi : \mathcal{A}^{\mathbb{N}} \to \mathbb{R}^2$ by $\Pi(\omega) = \sum_{n=1}^{\infty} r^{(n-1)} b_{\omega_n}$ and consider the set $\Lambda = \Pi(\mathcal{A}^{\mathbb{N}})$. If $\delta_n = 0$ for all n, then Λ is self-similar, but we now assume that $\delta_n > 0$ and $\delta_n \downarrow 0$. Further, suppose that

$$\min_{i \neq j} |b_i - b_j| \cdot r^{(1)} > \max_{i,j} |b_i - b_j| \cdot \sum_{n=2}^{\infty} r^{(n)}.$$

Then it is easy to see that Π is one-to-one and so Λ is a planar Cantor set. One can show that if the product $\prod_{i=1}^{\infty}(1 + \delta_i)$ diverges, then the one-dimensional Hausdorff measure of Λ is not σ-finite (this follows, e.g., from applying the results of [**20**] to the natural measure on Λ). It turns out that if this product diverges sufficiently slowly, then $\mathrm{Fav}(\Lambda) = 0$.

(Other deterministic sets of non-σ-finite $\mathcal{H}^1$ measure but zero Favard length can be found in [**13**, **7**, **8**].)

Proposition 7.2. *There exists $c > 0$ such that if*

$$\prod_{i=1}^{n} (1 + \delta_i) \leq \exp[c \log_* n],$$

then $\mathrm{Fav}(\Lambda) = 0$.

Sketch of the proof. The argument closely follows the proof of Theorem 2.3 (in the homogeneous case), so we only give a brief sketch.

Let $\Pi_\theta = \mathrm{proj}_\theta \circ \Pi$ and $\Lambda_u^\theta = \Pi_\theta([u])$ where $[u]$ is the cylinder set corresponding to $u \in \mathcal{A}^*$. For $u, v \in \mathcal{A}^*$, with $|u| = |v| = n$, we say that Λ_u^θ and Λ_v^θ are ε-relatively close if the Hausdorff distance between these sets is not greater than $\varepsilon r^{(n)}$. Define $\Psi(n, k, \varepsilon)$ as the set of $\theta \in [0, \pi]$ such that there is no collection of distinct words $u_1, \dots, u_k$, with $|u_j| \leq n$ for $j \leq k$, such that $\Lambda_{u_j}^\theta$, $j \leq k$, are pairwise ε-relatively close. The four lemmas in Section 3 and the proof of Theorem 2.3 (specialized to the homogeneous case $r_i = r$) go through essentially unchanged, replacing only r^n, r^q, *etc.*, with $r^{(n)}, r^{(q)}$, *etc.*

We use $\phi(t) = t$, so that $\mathcal{H}^\phi(\Lambda^\theta) = 0$ for a.e. θ is equivalent to $\mathrm{Fav}(\Lambda) = 0$. Further details are left to the reader. □

7.3. Unsolved problems.

Question 7.3. For a one-dimensional self-similar set in the plane, which satisfies strong separation, can the bound (2.3) be strengthened to

$$\mathrm{Fav}(\Lambda(\rho)) \le C\left(\log\left(\frac{1}{\rho}\right)\right)^{-1} \quad \text{for all} \quad \rho > 0, \tag{7.1}$$

for some $C < \infty$?

Perhaps a more accessible goal is to improve our estimates for random Cantor sets.

Question 7.4. For the random sets $\mathcal{R}_n$ considered in Theorem 2.2, can the upper bound (2.5) be improved to

$$\limsup_{n\to\infty} n \cdot \mathrm{Fav}(\mathcal{R}_n) < \infty \quad \text{a.s. ?} \tag{7.2}$$

A more ambitious program would be to relate the decay rate of Favard length of neighborhoods, to other quantitative measures of nonrectifiability. The following question is motivated by Jones' Traveling Salesman Theorem [**6**]. Given a compact planar set Λ, and $\epsilon > 0$, let

$$\ell(\Lambda, \epsilon) = \sup \mathcal{H}^1_\infty\Big(\Gamma(\epsilon) \cap \Lambda\Big),$$

where Γ runs over recifiable curves of length 1, and $\mathcal{H}^1_\infty$ denotes one-dimensional Hausdorff content. We can show that the four-corner set K^2 considered in the introduction satisfies $\ell(K^2, \epsilon) = O(|\log\epsilon|^{-1})$ as $\epsilon \to 0$.

Question 7.5. Is there a quantitative estimate of $\mathrm{Fav}(\Lambda(\epsilon))$ in terms of $\ell(\Lambda, \epsilon)$?

In particular, is $\mathrm{Fav}(\Lambda(\epsilon)) = O(\ell(\Lambda, \epsilon))$ as $\epsilon \to 0$?

Acknowledgments. The impetus for this work came from our joint paper with M. Rams and K. Simon [**17**] on the Hausdorff measure of self-conformal sets. We are grateful to them, and to P. Mattila, for useful discussions. Part of this work was done while the first author was visiting Microsoft Research.

References

[1] C. Bandt and S. Graf, *Self-similar sets* 7. *A characterization of self-similar fractals with positive Hausdorff measure*, Proc. Amer. Math. Soc., **114** (1992), 995-1001, MR 93d:28014, Zbl 0823.28003.

[2] A.S. Besicovitch, *Tangential properties of sets and arcs of infinite linear measure*, Bull. Amer. Math. Soc., **66** (1960), 353-359, MR 22 #11090, Zbl 0093.35801.

[3] G. David, *Analytic capacity, Calderón-Zygmund operators, and rectifiability*, Publ. Mat., **43** (1999), 3-25, MR 2000e:30044.

[4] K.J. Falconer, *The geometry of fractal sets*, Cambridge Tracts in Mathematics, **85**, C.U.P., Cambridge-New York, 1986, MR 88d:28001, Zbl 0587.28004.

[5] J.E. Hutchinson, *Fractals and self-similarity*, Indiana Univ. Math. J., **30** (1981), 713-747, MR 82h:49026, Zbl 0598.28011.

[6] P.W. Jones, *Rectifiable sets and the travelling salesman problem*, Invent. Math., **102** (1990), 1-15, MR 91i:26016, Zbl 0731.30018.

[7] P.W. Jones and T. Murai, *Positive analytic capacity but zero Buffon needle probability*, Pacific J. Math., **133** (1988), 99-114, MR 89m:30050, Zbl 0653.30016.

[8] H. Joyce and P. Mörters, *A set with finite zero curvature and projections of zero length*, J. Math. Anal. Appl., **247** (2000), 126-135, MR 2001j:28006, Zbl 0973.30022.

[9] R. Kenyon, *Projecting the one-dimensional Sierpinski gasket*, Israel J. Math., **97** (1997), 221-238, MR 98i:28002, Zbl 0871.28006.

[10] J.C. Lagarias and Y. Wang, *Tiling the line with translates of one tile*, Invent. Math., **124** (1996), 341-365, MR 96i:05040, Zbl 0847.05037.

[11] R. Lyons, *Random walks, capacity and percolation on trees*, Ann. Probab., **20** (1992), 2043-2088, MR 93k:6017, Zbl 0766.60091.

[12] J. Mateu, X. Tolsa and J. Verdera, *On the semiadditivity of analytic capacity and planar Cantor sets*, preprint, 2002.

[13] P. Mattila, *Smooth maps, null-sets for integralgeometric measures and analytic capacity*, Ann. Math., **123** (1986), 303-309, MR 87d:28010, Zbl 0589.28006.

[14] ———, *Orthogonal projections, Riesz capacities and Minkowski content*, Indiana Univ. Math. J., **39** (1990), 185-198, MR 91d:28018, Zbl 0682.28003.

[15] ———, *Geometry of Sets and Measures in Euclidean Spaces*, Cambridge University Press, 1995, MR 96h:28006, Zbl 0819.28004.

[16] ———, *On the analytic capacity and curvature of some Cantor sets with non-σ-finite length*, Publ. Mat., **40**(**1**) (1996), 195-204, MR 97d:30052, Zbl 0888.30026.

[17] Y. Peres, M. Rams, K. Simon and B. Solomyak, *Equivalence of positive Hausdorff measure and the open set condition for self-conformal sets*, Proc. Amer. Math. Soc., **129**(**9**) (2001), 2689-2699 (electronic), MR 2002d:28004.

[18] Y. Peres, K. Simon and B. Solomyak, *Self-similar sets of zero Hausdorff measure and positive packing measure*, Israel J. Math., **117** (2000), 353-379, MR 2001g:28017, Zbl 0963.28008.

[19] Y. Peres and B. Solomyak, *Approximation by polynomials with coefficients* ± 1, J. Number Theory, **84**(**2**) (2000), 185-198, CMP 1 795 789.

[20] C.A. Rogers and S.J. Taylor, *Functions continuous and singular with respect to a Hausdorff measure*, Mathematika, **8** (1962), 1-31, MR 24 #A200, Zbl 0145.28701.

Received October 31, 2000 and revised May 9, 2001. Research of Peres was partially supported by NSF grant #DMS-9803597, and by the Landau center for Mathematical Analysis at the Hebrew University. Research of Solomyak was supported in part by NSF grant #DMS 9800786 and the Institute of Mathematics at the Hebrew University.

Department of Mathematics
Hebrew University
Jerusalem

Department of Statistics
University of California
Berkeley, CA 94720
E-mail address: peres@stat.berkeley.edu

Department of Mathematics, Box 354350
University of Washington
Seattle, WA 98195
E-mail address: solomyak@math.washington.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 204, No. 2, 2002

DETERMINING THE POTENTIAL OF A STURM–LIOUVILLE OPERATOR FROM ITS DIRICHLET AND NEUMANN SPECTRA

Virgil Pierce

In this paper we consider the inverse spectral problem for the Sturm–Liouville Operator on the interval $[0, 1]$. We show that given the Dirichlet and Neumann spectra of such an operator we find a generically uncountable family of potentials with these spectra.

1. Introduction.

We will consider this problem: Given the Dirichlet and Neumann spectra of the Sturm–Liouville Operator

$$-\frac{d^2}{dx^2} + q(x) \tag{1}$$

for a potential q in $C^3([0,1])$, determine q. Instead of finding a unique q we get a generically uncountable family of potentials that will have the given joint spectra. Borg [**1**] showed that if the gaps (see Figure 1) are all trivial then the potential $q(x)$ is 0. Levinson [**11**] showed that if given the spectra of (1) corresponding to the two sets of boundary conditions,

$$y(0)\cos\alpha + y'(0)\sin\alpha = 0, \quad y(1)\cos\beta + y'(1)\sin\beta = 0 \tag{2}$$

$$y(0)\cos\alpha + y'(0)\sin\alpha = 0, \quad y(1)\cos\gamma + y'(1)\sin\gamma = 0 \tag{3}$$

with $\sin(\gamma - \beta) \neq 0$, then $q(x)$ is uniquely determined. Notice that this theorem does not include the case of Dirichlet (boundary conditions $y(0) = y(1) = 0$) and Neumann (boundary conditions of $y'(0) = y'(1) = 0$) spectra. Borg [**1**], Levinson [**11**], Isaacson, McKean and Trubowitz [**8**] among others demonstrated that the spectrum given by one boundary condition does not determine the operator.

The dynamical behavior of solutions to Hill's Operator (the 1-D Schrödinger or Sturm-Liouville Operator with periodic potential) is determined by the properties of the associated Floquet discriminant function [**12**]. Its and Matveev [**10**], Gelfand [**5**], Gelfand and Levitan [**6**], McKean [**15**], Garnett [**4**], Trubowitz [**17**], and Buslaev and Faddeev [**2**] illustrate that for periodic potentials the periodic, anti-periodic, and Dirichlet spectra determine the potential.

We address our stated problem by applying the well understood periodic theory to a periodic extension of q. This approach to the problem was originally suggested by H. McKean [private communication]. To state the theorem we must first introduce some terminology. By $\{\mu_n\}$ (resp. $\{\nu_n\}$) we denote the Dirichlet (resp. Neumann) spectrum of the operator (1). Define an even periodic potential $Q(x)$ for which $\{4\mu_n, 4\nu_n\}$ comprise the periodic spectrum of the operator $-\frac{d^2}{dx^2} + Q(x)$. Let $\{\lambda_j\}$ denote the joint periodic and anti-periodic spectra of this operator. Note that the periodic spectrum determines the anti-periodic spectrum (see Proposition (6)). The advantage of an even potential is that its periodic and anti-periodic spectra are also its Dirichlet and Neumann spectra.

Theorem 1 (Main Theorem)**.** *We are given the Dirichlet* $\{\mu_n\}$*, and Neumann* $\{\nu_n\}$*, eigenvalues of* (1) *which satisfy the asymptotics*

$$\mu_n,\, \nu_n = n^2\pi^2 + \mathcal{O}\left(\frac{1}{n^2}\right).$$

The family of potentials, $q(x)$*, having the same Dirichlet and Neumann eigenvalues is of the form* $q(x) = \frac{1}{4}Q(\frac{1}{2}x) \quad x \in [0,1]$ (6) *where* $Q(x)$ *is an even potential of the form*

$$Q(x) = \lambda_0 + \sum_{n\geq 1} \lambda_{2n-1} + \lambda_{2n} - 2c_n(x), \tag{4}$$

with $c_n(x)$ *the* $W^{1,2}_{\mathrm{per}}([0,1])$ (*the Sobolev space of differentiable functions with* $L^2([0,1])$ *first derivative*) *solution of*

$$\begin{aligned} c_{2n}(0) &= 4\mu_n \\ c_{2n-1}(0) &= \lambda_{4n-3} \quad \textit{or} \quad \lambda_{4n-2} \\ \frac{dc_n}{dx} &= \sqrt{(c_n - \lambda_{2n-1})(c_n - \lambda_{2n}) \prod_{k\neq n} \frac{(c_n - \lambda_{2k-1})(c_n - \lambda_{2k})}{(c_n - c_k)^2}}. \end{aligned} \tag{5}$$

In the above theorem we utilize the Trace Formula (4) for potentials q which are C^3 on all but a finite number of points. This formula says that such a q is determined by the periodic, anti-periodic and shifted Dirichlet eigenvalues of the operator $-\frac{d^2}{dx^2} + q(x)$ [**17**]. The shifted Dirichlet eigenvalues satisfy a first order ODE (5) and so are themselves determined by the Dirichlet eigenvalues of q.

Therefore the periodic and Dirichlet spectra of the Sturm-Liouville operator with periodic Q uniquely determine Q. It is at the step of passing from knowing the periodic, and only the half of the Dirichlet spectrum corresponding to periodic eigenvalues of the operator $-\frac{d^2}{dx^2} + Q(x)$ that we reach an ambiguity when we are given a choice as to the anti-periodic half of the Dirichlet spectrum.

2. Even potentials.

For an arbitrary potential $q(x) \in L^2_{\mathbb{R}}([0,1])$ we form an even periodic potential

$$Q(x) = \begin{cases} 4q(2x) & : \quad x \in [0, 1/2] \\ 4q(2(1-x)) & : \quad x \in (1/2, 1] \end{cases}. \tag{6}$$

Notice that for $q(x) \in C^3([0,1])$, $Q(x)$ is C^3 everywhere except at the point $\frac{1}{2}$ where it is only continuous.

Let $u(x)$ be a Dirichlet eigenfunction of the operator $-\frac{d^2}{dx^2} + q(x)$ corresponding to eigenvalue μ_j. Then

$$U(x) = \begin{cases} u(2x) & : \quad x \in [0, 1/2] \\ -u(2(1-x)) & : \quad x \in (1/2, 1] \end{cases} \tag{7}$$

is a Dirichlet eigenfunction of the operator $-\frac{d^2}{dx^2} + Q(x)$ with eigenvalue $4\mu_j$. Likewise if $v(x)$ is a Neumann eigenfunction with eigenvalue ν_j then

$$V(x) = \begin{cases} v(2x) & : \quad x \in [0, 1/2] \\ v(2(1-x)) & : \quad x \in (1/2, 1] \end{cases} \tag{8}$$

is a Neumann eigenfunction for the operator $-\frac{d^2}{dx^2} + Q(x)$ with eigenvalue $4\nu_j$.

Both $U(x)$ and $V(x)$ are also periodic eigenfunctions of the operator $-\frac{d^2}{dx^2} + Q(x)$. We conclude that the Dirichlet and Neumann spectra of $-\frac{d^2}{dx^2} + q(x)$ determine the periodic spectrum of $-\frac{d^2}{dx^2} + Q(x)$. Using the Counting Lemma (2) and Proposition (4) we see that the Dirichlet and Neumann eigenvalues paired with their respective $U(x)$ or $V(x)$ account for only the gaps which are given by the periodic spectrum.

Therefore we reduce the inverse problem to the case of a periodic potential. We use the monodromy matrix,

$$\begin{pmatrix} y_1(1,\lambda) & y_2(1,\lambda) \\ y_1'(1,\lambda) & y_2'(1,\lambda) \end{pmatrix} = M(\lambda), \tag{9}$$

where y_1 and y_2 are the two linearly independent fundamental solutions given by $y_1(0,\lambda) = y_2'(0,\lambda) = 1$ and $y_1'(0,\lambda) = y_2(0,\lambda) = 0$. This matrix describes the behavior of the solutions to Hill's operator on $\mathbb{R}$. For example periodic solutions to the differential equation correspond to unit eigenvalues of this matrix. Because of the initial values of y_1 and y_2 a Dirichlet eigenvalue μ corresponds to $y_2(1,\mu) = 0$ and a Neumann eigenvalue η corresponds to $y_1'(1,\eta) = 0$.

The periodic and anti-periodic eigenvalues are values of λ for which $M(\lambda)$ has respectively eigenvalues ± 1. In either case Equation (1) has a solution with period 2. $\Delta(\lambda)$ denotes the trace of $M(\lambda)$. Periodic (resp. anti-periodic) eigenvalues of $q(x)$ are roots of $\Delta - 2$ (resp. $\Delta + 2$).

Lemma 1. *If $Q(x)$ is an even potential and λ_j is both a Neumann and Dirichlet eigenvalue then $\Delta(\lambda_j) = \pm 2$ and $\Delta'(\lambda_j) = 0$.*

Proof. From the results above, $y_2(1, \lambda_j) = 0$ and $y_1'(1, \lambda_j) = 0$. Therefore $M(\lambda_j)$ is diagonal with determinant 1, so $\Delta(\lambda_j) = \pm 2$. To prove the statement about the derivative of Δ we will use a formula from [**12**],

$$\Delta'(\lambda) = \big(y_1(1,\lambda) - y_2'(1,\lambda)\big) \int_0^1 y_1(x,\lambda) y_2(x,\lambda) dx - y_2(1,\lambda) \int_0^1 y_1^2(x,\lambda) dx + y_1'(1,\lambda) \int_0^1 y_2^2(x,\lambda) dx. \tag{10}$$

Now notice that if $y_1(x, \lambda_j)$ is a solution of (1) then so is

$$y_1(1-x,\lambda_j) = y_1(1,\lambda_j) y_1(x,\lambda_j) + y_1'(1,\lambda_j) y_2(x,\lambda_j) \tag{11}$$

as $y_1(1 - x, \lambda_j)$ will satisfy Equation (1) with $Q(1 - x) = Q(x)$. Therefore, since λ_j is a Neumann eigenvalue we see that

$$y_1(1-x,\lambda_j) = y_1(1,\lambda_j) y_1(x,\lambda_j). \tag{12}$$

Setting $x = 1$ in the above equation we get $y_1(1, \lambda_j) = \pm 1$. The determinant of the monodromy matrix is 1, and because λ_j is a Dirichlet eigenvalue $y_2(1, \lambda_j) = 0$ so

$$y_2'(1,\lambda_j) = \frac{1}{y_1(1,\lambda_j)} = \pm 1. \tag{13}$$

We then substitute into (10)

$$y_2(1,\lambda_j) = y_1'(1,\lambda_j) = 0$$

and

$$y_1(1,\lambda_j) = y_2'(1,\lambda_j) = \pm 1$$

to get $\Delta'(\lambda_j) = 0$. □

Proposition 1. *λ_j is a periodic or anti-periodic eigenvalue of an even potential Q if and only if λ_j is a Neumann or Dirichlet eigenvalue of Q.*

Proof. Suppose λ_j is a periodic eigenvalue so $\Delta(\lambda_j) = 2$. We must show that $y_2(1, \lambda_j) = 0$ or $y_1'(1, \lambda_j) = 0$. Suppose λ_j is not a Neumann eigenvalue for Q; that is $y_1'(1, \lambda_j) \neq 0$.

From $\Delta(\lambda_j) = 2$ we see that the monodromy matrix is of the form

$$M(\lambda_j) = \begin{pmatrix} y_1(1,\lambda_j) & y_2(1,\lambda_j) \\ y_1'(1,\lambda_j) & 2 - y_1(1,\lambda_j) \end{pmatrix}, \tag{14}$$

as λ_j is a periodic eigenvalue.

If $Q(x)$ is even then $y_1(1 - x, \lambda_j)$ is also a solution of

$$-\frac{d^2y}{dx^2} + (Q(x) - \lambda_j) y = 0. \tag{15}$$

Because y_1 and y_2 form a basis for the solutions to this equation we may write $y_1(1-x)$ as

$$y_1(1-x,\lambda_j) = y_1(1,\lambda_j)y_1(x,\lambda_j) - y_1'(1,\lambda_j)y_2(x,\lambda_j). \tag{16}$$

By setting $x = 1$ in this equation and its derivative we get the following two equations:

$$1 = y_1(1,\lambda_j)^2 - y_1'(1,\lambda_j)y_2(1,\lambda_j) \tag{17}$$

$$0 = y_1(1,\lambda_j)y_1'(1,\lambda_j) - y_1'(1,\lambda_j)y_2'(1,\lambda_j). \tag{18}$$

From (14) we have $y_2'(1,\lambda_j) = 2 - y_1(1,\lambda_j)$ and using this relation in (18) we get the equation

$$0 = 2y_1'(1,\lambda_j)(y_1(1,\lambda_j) - 1). \tag{19}$$

By the assumption that λ_j is not a Neumann eigenvalue we conclude that $y_1(1,\lambda_j) = 1$. Substituting this into Equation (17) we conclude that $y_2(1,\lambda_j) = 0$ and so λ_j is a Dirichlet eigenvalue.

Conversely, suppose that λ_j is a Dirichlet eigenvalue, $y_2(1,\lambda_j) = 0$. We must show that $\Delta(\lambda_j) = \pm 2$. Since $\det M(\lambda) = 1$ we have $y_2'(1,\lambda_j) = 1/y_1(1,\lambda_j)$. Substituting this into (18) we conclude that either $y_1'(1,\lambda_j) = 0$ in which case Lemma (1) completes the proof; otherwise, we get $y_1(1,\lambda_j) = \pm 1$, which implies that $\Delta(\lambda_j) = \pm 2$.

A similar argument may be made for the Neumann case with the additional feature that, when Q is an even potential, the lowest Neuman eigenvalue, ν_0, is equal to the lowest periodic eigenvalue, λ_0. □

We introduce the picture of gaps and bands associated to $\Delta(Q,\lambda)$. The bands are the ranges of eigenvalues whose eigenfunctions are bounded (stable) on $\mathbb{R}$. That is the range of λ's for which the eigenvalues of the monodromy matrix are complex valued with modulus less than 1. These bands are clearly the intervals over which $|\Delta(\lambda)| < 2$. Correspondingly the intervals for which $|\Delta(\lambda)| > 2$ are called the gaps. These are intervals for which there exist unbounded (unstable) solutions to (1). Gap intervals may be trivial; i.e., they may collapse to a single point.

Finally we need Theorem 2 from [**17**].

Theorem 2 (Trace Formula)**.** *Let $q \in C^3[0,1]$ be a potential with Dirichlet eigenvalues μ_n and periodic, anti-periodic eigenvalues λ_j. Let $\mu_n(t)$, $n \geq 1$, be the unique periodic solution of the system*

$$\frac{d\mu_n}{dt} = \sqrt{(\mu_n - \lambda_{2n-1})(\mu_n - \lambda_{2n}) \prod_{k\neq n} \frac{(\mu_n - \lambda_{2k-1})(\mu_n - \lambda_{2k})}{(\mu_n - \mu_k)^2}}, \tag{20}$$

on the Riemann surface given by the equation

$$y_n = \sqrt{(\mu_n - \lambda_{2n-1})(\mu_n - \lambda_{2n})},$$

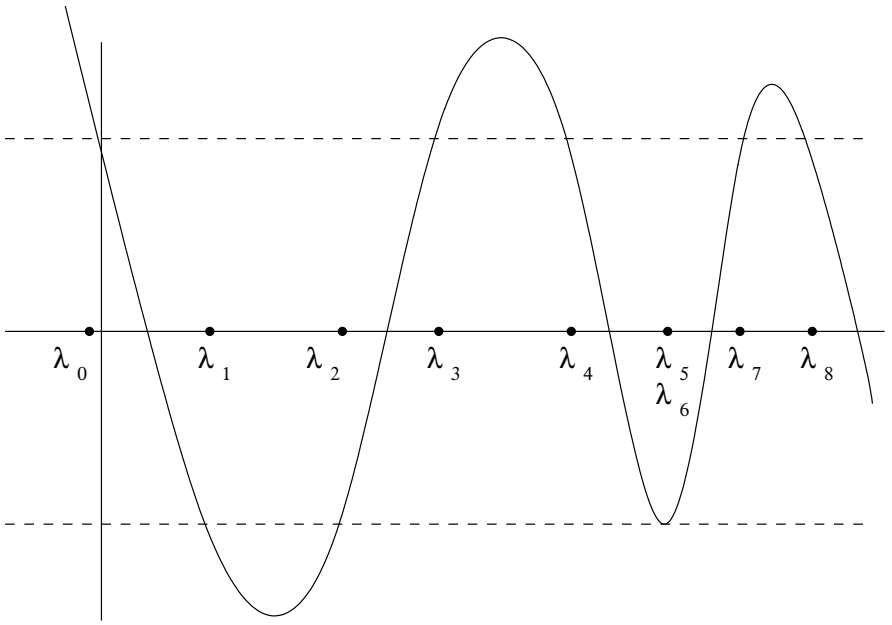


Figure 1. Gaps and Bands.

whose initial values $\mu_n(0) = \mu_n$ *the* n^{th} *Dirichlet eigenvalue and for which the initial velocities are prescribed by choosing the signature of the radical* $\sqrt{\Delta^2(\mu_n) - 4}$ *such that*

$$\sqrt{\Delta(\mu_n)^2 - 4} = 2y_2'(1, \mu_n) - \Delta(\mu_n). \tag{21}$$

Then,

$$q(t) = \lambda_0 + \sum_{n \geq 1} \lambda_{2n-1} + \lambda_{2n} - 2\mu_n(t). \tag{22}$$

The proof in [**17**] is given for potentials in $C^3_{\text{per}}([0,1])$. For the purposes of this paper we wish to apply this theorem to the potential $Q(x)$ which is C^3 for every point except 0, $\frac{1}{2}$ and 1 where Q is not continuously differentiable. To show that the theorem still holds in this case we will demonstrate that the trace formula is still well-defined.

We have from [**16**] the estimate

$$\mu_n = n^2\pi^2 - \int_0^1 \cos(2\pi nx) q(x) dx + \mathcal{O}\left(\frac{1}{n^2}\right) \tag{23}$$

for $q(x) \in W^{2,2}([0,1])$. In fact our $q(x)$ is twice differentiable for all but one point. Below we give an argument for a bound on the $\cos(2\pi nx)$ inner product above. This same technique shows that the $\mathcal{O}\left(\frac{1}{n^2}\right)$ term above will

remain of the same order. Compute an estimate of the integral above

$$\begin{aligned}(24)\qquad &\int_0^1 \cos(2\pi nx)q(x)dx\\ &= q(x)\ \frac{\sin(2\pi nx)}{2\pi n}\Big|_0^1\\ &\quad - \int_0^{\frac{1}{2}} q'(x)\frac{\sin(2\pi nx)}{2\pi n}dx - \int_{\frac{1}{2}}^1 q'(x)\frac{\sin(2\pi nx)}{2\pi n}dx.\end{aligned}$$

Integrating by parts a second time we get

$$\begin{aligned}(25)\qquad &\int_0^1 \cos(2\pi nx)q(x)dx\\ &= q'(x)\ \frac{\cos(2\pi nx)}{4\pi^2n^2}\Big|_0^{\frac{1}{2}} + q'(x)\ \frac{\cos(2\pi nx)}{4\pi^2n^2}\Big|_{\frac{1}{2}}^1\\ &\quad - \int_0^{\frac{1}{2}} q''(x)\frac{\cos(2\pi nx)}{4\pi^2n^2}dx - \int_{\frac{1}{2}}^1 q''(x)\frac{\cos(2\pi nx)}{4\pi^2n^2}dx.\end{aligned}$$

So we see that $\mu_n = n^2\pi^2 + \mathcal{O}(1/n^2)$ for the $Q(x)$ we are concerned with. There was nothing special about the points where differentiability failed so the shifted Dirichlet eigenvalues will have the same asymptotics as well. The periodic and anti-periodic spectra satisfy the same asymptotics as they are the Dirichlet and Neumann spectra of the even potential. The sum we are concerned with is

$$(26)\qquad \sum_{n\geq 1} \left|\lambda_{2n-1} + \lambda_{2n} - 2\mu_n(t)\right|.$$

By the analysis above each term satisfies the asymptotics $n^2\pi^2 + \mathcal{O}(1/n^2)$ and so the sum converges absolutely.

Proposition 2. *If $\mu_n(0) = \lambda_{2n-1}$ or λ_{2n} for all n then the function $q(x)$ determined by* (22) *is even.*

Proof. As a consequence of the trace formula it will suffice to show that $\mu_n(x) = \mu_n(1-x)$. Differentiating this function with respect to x we get

$$\begin{aligned}(27)\qquad &\frac{d}{dx}\mu_n(1-x)\\ &= -\sqrt{(\mu_n - \lambda_{2n-1})(\mu_n - \lambda_{2n})\prod_{k\neq n}\frac{(\mu_n - \lambda_{2k-1})(\mu_n - \lambda_{2k})}{(\mu_n - \mu_k)^2}}.\end{aligned}$$

From the periodicity of the original solutions the initial conditions which determine $\mu_n(1-x)$ are the same as the ones for $\mu_n(x)$ specifically $\mu_n(0) = \mu_n(1) = \lambda_{2n-1}$ or λ_{2n}. There exists a solution of Equation (27) which is

periodic and does not pause at the endpoints of the interval $[\lambda_{2n-1}, \lambda_{2n}]$. However from the ambiguity of the choice of sign this solution must be identical to that of the original equation which is also periodic and does not pause at the endpoints of the interval. □

3. Eigenvalues.

The following proposition examines further the relationship between the Dirichlet and Neumann spectra, and the gaps.

Proposition 3. *Suppose q is a potential on $[0,1]$ and $[\lambda_{2j-1}, \lambda_{2j}]$ is a gap. In other words $|\Delta(\lambda)| \geq 2$ for all λ in $[\lambda_{2j-1}, \lambda_{2j}]$, then there is a μ and η in $[\lambda_{2j-1}, \lambda_{2j}]$ such that μ is a Dirichlet eigenvalue and η is a Neumann eigenvalue.*

Proof. We will prove this by showing that $y_1'(1,\lambda)$ and $y_2(1,\lambda)$ switch sign from the left of λ_{2j-1} to the right of λ_{2j}.

We follow Magnus and Winkler for this proof [**12**].

Combining Formula (10) into one integral and shortening notation via $\eta_1 = y_1(1,\lambda_{2j-1})$, $\eta_1' = y_1'(1,\lambda_{2j-1})$ etc., we get the equation

$$\Delta'(\lambda_{2j-1}) = \int_0^1 \left(\left(\eta_1 - \eta_2'\right) y_1 y_2 - \eta_2 y_1^2 + \eta_1' y_2^2 \right) dx. \tag{28}$$

We also need the formula

$$\begin{aligned} \Delta^2 - 4 &= \left(\eta_1 + \eta_2'\right)^2 - 4\left(\eta_1\eta_2' - \eta_1'\eta_2\right) \\ &= \left(\eta_1 - \eta_2'\right)^2 + 4\eta_1'\eta_2. \end{aligned} \tag{29}$$

Note that $\eta_1' \neq 0$ to the left of λ_{2j-1} and to the right of λ_{2j}, where $|\Delta(\lambda)| < 2$. So by adding and subtracting $(\Delta^2 - 4)\int_0^1 y_1^2 dx/4\eta_1'$ from (28) we get

$$\begin{aligned} &\int_0^1 \Bigg(\left(\eta_1 - \eta_2'\right) y_1 y_2 - \eta_2 y_1^2 + \eta_1' y_2^2 \\ &\qquad + \frac{\left(\eta_1 - \eta_2'\right)^2 y_1^2}{4\eta_1'} + \frac{4\eta_1'\eta_2 y_1^2}{4\eta_1'} - \frac{\Delta^2 - 4}{4\eta_1'} y_1^2 \Bigg) dx \\ &= \operatorname{sign}(\eta_1') \int_0^1 \left(\left(\sqrt{|\eta_1'|}\, y_2 + \frac{\eta_1 - \eta_2'}{2\sqrt{|\eta_1'|}} \operatorname{sign}(\eta_1') y_1 \right)^2 - \frac{\Delta^2 - 4}{4|\eta_1'|} y_1^2 \right). \end{aligned} \tag{30}$$

As $|\Delta(\lambda)| < 2$ in the regions being considered, the integrand is a positive number. Yet $\Delta'(\lambda)$ switches sign once in $[\lambda_{2j-1}, \lambda_{2j}]$. This implies that η_1' switches sign as needed. A similar proof will show that η_2 also switches sign. □

A corollary of Formula (30) is that if $\Delta'(\lambda) = 0$ then $|\Delta(\lambda)| \geq 2$.

4. Counting Lemma.

How many Dirichlet and Neumann eigenvalues are in each gap? To answer this question we use the Counting Lemma from Pöschel and Trubowitz [**16**].

Lemma 2 (Counting Lemma: Dirichlet Eigenvalues)**.** *Let* $q \in L^2_{\mathbb{R}}([0,1])$ *and let* $N > 2e^{\|q\|}$ *be an integer. Then* $y_2(1,\lambda)$ *has exactly* N *roots, counted with multiplicities, in the open half plane*

$$Re\,(\lambda) < \left(N + \frac{1}{2}\right)^2 \pi^2 \tag{31}$$

and for each $n > N$*, exactly one simple root in the egg shaped region*

$$|\sqrt{\lambda} - n\pi| < \frac{\pi}{2}. \tag{32}$$

There are no other roots.

An analogous result is true for Neumann eigenvalues. The necessary tools are available in [**16**]. For completeness we will state the lemma here:

Lemma 3 (Counting Lemma: Neumann Eigenvalues)**.** *Let* $q \in L^2_{\mathbb{R}}([0,1])$ *and let* $N > 2e^{\|q\|}$ *be an integer. Then* $y_1'(1,\lambda)$ *has exactly* $N+1$ *roots, counted with multiplicities, in the open half plane*

$$Re\,(\lambda) < \left(N + \frac{1}{2}\right)^2 \pi^2 \tag{33}$$

and for each $n > N$*, exactly one simple root in the egg shaped region*

$$|\sqrt{\lambda} - n\pi| < \frac{\pi}{2}. \tag{34}$$

There are no other roots.

The "extra" Neumann eigenvalue in the half plane corresponds to the "ground state" of the Neumann problem. For general potentials, this eigenvalue is less than or equal to λ_0, the first periodic eigenvalue; but, when q is even it is pinned at λ_0.

For the periodic and anti-periodic spectra we shift the potential until it is an even potential, then the Dirichlet and Neumann spectra form the periodic and anti-periodic spectra. Therefore we get the analogous result for the periodic and anti-periodic spectra (the periodic and anti-periodic spectra are invariant under translation of the potential q and the average value of q is invariant under translation).

Proposition 4. *For periodic* $q \in L^2_{\mathbb{R}}([0,1])$ *there is one and only one Dirichlet eigenvalue within each gap.*

Proof. Choose N satisfying the hypothesis of the Counting Lemma. By Proposition 3 there is at least one Dirichlet eigenvalue in each gap. There are N Dirichlet eigenvalues in the region $\mathrm{Re}(\lambda) < (N+\frac{1}{2})^2\pi^2$. Therefore in the same region there is one and only one Dirichlet eigenvalue within each gap.

With $n > N$ for the intervals $|\sqrt{\lambda} - n\pi| < \frac{\pi}{2}$ there is one gap. Within this same region there is one Dirichlet eigenvalue. In these zones there is one and only one Dirichlet eigenvalue within each gap. □

Proposition 5. *For periodic $q \in L^2_{\mathbb{R}}([0,1])$ there is one and only one Neumann eigenvalue within each gap. There is one and only one Neumann eigenvalue within the interval $(-\infty, \lambda_0]$.*

The proof of this proposition follows the one above.

Proposition 6. *Δ is determined by the periodic eigenvalues of q.*

Proof. The periodic eigenvalues are the roots of $\Delta - 2$, they are real since they are eigenvalues of a self-adjoint operator . Therefore we have

$$\Delta - 2 = C\left(\prod_{n=1}^{\infty} \frac{(\lambda_{2n-1} - \lambda)(\lambda_{2n} - \lambda)}{n^4\pi^4}\right)(\lambda - \lambda_0), \tag{35}$$

(see [**13**]) where $\{\lambda_i\}$ are the periodic eigenvalues and C is a constant, provided that this product converges. C is determined by the asymptotic condition on the roots of $\Delta^2 - 4$,

$$\lambda_{2n-1}, \lambda_{2n} = n^2\pi^2 + \int_0^1 q(x)dx + \mathcal{O}(n^{-2}) \tag{36}$$

for $q \in C^3([0,1])$. Without loss of generality we may take $\int_0^1 q(x)dx = 0$.

We first show that the product converges uniformly. From Markushevich ([**13**]) we have that $\prod_{i=1}^{\infty}(1 - \frac{\lambda}{\lambda_i})$ converges uniformly if and only if $\sum_{i=1}^{\infty} \frac{\lambda}{\lambda_i}$ is uniformly convergent.

Choose N such that $|\lambda_{2n-1} - n^2\pi^2| < \delta$ and $|\lambda_{2n} - n^2\pi^2| < \delta$ for all $n > N$ and that

$$\sum_{i=N+1}^{\infty} \frac{1}{i^2\pi^2} < \frac{\delta}{4}. \tag{37}$$

Consider

$$\sum_{i=1}^{\infty}\left|\frac{\lambda}{\lambda_i}\right| = |\lambda|\sum_{i=1}^{\infty}\left|\frac{1}{\lambda_i}\right| = |\lambda|\left(\sum_{j=1}^{\infty}\left|\frac{1}{\lambda_{2j-1}}\right| + \sum_{j=1}^{\infty}\left|\frac{1}{\lambda_{2j}}\right|\right). \tag{38}$$

We examine the tail of this series,

$$\sum_{i=N+1}^{\infty}\left|\frac{1}{\lambda_{2i-1}}\right|+\sum_{i=N+1}^{\infty}\left|\frac{1}{\lambda_{2i}}\right|\leq\sum_{i=N+1}^{\infty}\frac{2}{i^2\pi^2-\delta}\leq 4\sum_{i=N+1}^{\infty}\frac{1}{i^2\pi^2}. \tag{39}$$

Which is the estimate we need.

Finally to combine this with our problem we have the infinite product

$$\prod_{i=1}^{\infty}\frac{\lambda_{2n-1}\lambda_{2n}}{n^4\pi^4}\left(1-\frac{\lambda}{\lambda_{2n}}\right)\left(1-\frac{\lambda}{\lambda_{2n-1}}\right). \tag{40}$$

The term we have factored out of each part of the product is a constant in λ and therefore our conclusion is that the original product converges uniformly. □

This proposition says that Δ is determined by the periodic spectrum.

5. Proof of the main theorem.

If we are given the Dirichlet and Neumann spectra with appropriate asymptotic conditions for an $C^3([0,1])$ potential q on $[0,1]$ we begin the solution of the inverse problem by first extending q to an even potential $Q(x)$ on $[0,1]$. $Q(x)$ is $C^3([0,1])$ at all but one point. As described in Section 1 the Dirichlet and Neumann spectra, $\{\mu_n,\nu_n\}$ give the periodic spectrum, $\{4\mu_n,4\nu_n,4\nu_0\}$, of the operator with potential $Q(x)$. The eigenvalue $4\nu_0$ is the first periodic eigenvalue of the operator with potential $Q(x)$. The corresponding eigenfunctions remain Dirichlet and Neumann eigenfunctions.

For an even potential we have shown that for each pair of endpoints of a gap one is Dirichlet and the other is Neumann. We have also shown that these are all of the Dirichlet and Neumann spectra. The endpoints of a gap are either a pair of periodic or of anti-periodic eigenvalues of Q. Therefore the Dirichlet and Neumann eigenvalues of $-\frac{d^2}{dx^2}+q(x)$ give the periodic half of the Dirichlet and Neumann spectra.

This periodic spectrum, $\{4\mu_n,4\nu_n,4\nu_0\}$ determines $\Delta(\lambda)$ by Proposition (6). From $\Delta(\lambda)$ we find the anti-periodic spectrum as the roots of $\Delta(\lambda)+2$. For each pair of anti-periodic eigenvalues one must be Dirichlet and the other Neumann by Propositions (4) and (1). The choice we make as to which anti-periodic eigenvalue of a given pair are to be a Dirichlet eigenvalue is where the ambiguity in the problem arises. That is we do not get a determined Dirichlet spectrum; potentially, one half of the spectrum is known only up to a sequence of pairs from which it may be chosen.

The Dirichlet spectrum once chosen specifies the initial conditions for the ODEs found previously (20). The solutions to these ODEs and the periodic and anti-periodic spectrum are inserted into the trace formula (22) giving an expression for $Q(x)$. An admissible $q(x)$ for the stated inverse problem is

the first half of $Q(x)$ appropriately scaled to be a function on $[0,1]$, explicitly $q(x) = \frac{1}{4}Q(\frac{1}{2}x)$ $x \in [0,1]$.

What Dirichlet and Neumann spectra would lead to only a finite number of possibilities for $q(x)$ determined by the method discussed above? One interesting method used to address this question utilizes theta functions and other tools from algebraic geometry, constructing $q(x)$ as a ratio of theta functions ([**10**] and [**7**]). Hochstadt [**7**] showed that if the gaps (see Figure 1) are all trivial then the potential $q(x)$ is 0. Hochstadt went on to show that if only one of the gaps does not vanish then $q(x)$ is an elliptic function. He finished with a proof that if only a finite number of the instability intervals were nontrivial then $q(x)$ is a C^∞ function. In these cases there are only finitely many $q(x)$ solving the inverse problem.

This work was supported by a scholarship from the ARCS Foundation and by an NSF VIGRE Grant #DMS9977116. I would like to thank Nick Ercolani and Leonid Friedlander for introducing me to this problem and for many helpful conversations. I would also like to thank Doug Pickrell for his comments.

References

[1] G. Borg, *Eine unkehrung der Sturm-Liouvilleschen eigenwertaufgabe*, Acta Math., **78** (1945), 1-96, MR 7,382d, Zbl 0063.00523.

[2] V. Buslaev and L. Faddeev, *On trace formulas for singular Sturm-Liouville operators*, Dokl. Akad. Nauk, **132** (1960), 13-16.

[3] N. Ercolani and H.P. McKean, *A quick proof of Fay's secant identities*, Analysis, et cetera, (1990), 301-307, MR 91b:14031, Zbl 0719.14029.

[4] J. Garnett and E. Trubowitz, *Gaps and bands of one dimensional periodic Schödinger operators*, II, Comment. Math. Helvetici, **62** (1987), 18-37, MR 88g:34028, Zbl 0649.34034.

[5] I.M. Gelfand, *On identities for the eigenvalues of a second-order differential operator*, Uspehi. Mat. Nauk, **11** (1956), 191-198.

[6] I.M. Gelfand and B.M. Levitan, *On a simple identity for the eigenvalues of a second-order differential operator*, Dokl. Akad. Nauk, **88** (1953), 953-956, MR 15,33a.

[7] H. Hochstadt, *On the determination of a Hill's Equation from its spectrum*, Arch. Rational Mech. Anal., **19** (1965), 353-362, MR 31 #6019, Zbl 0128.31201.

[8] E.L. Isaacson and E. Trubowitz, *The inverse Sturm-Liouville Problem*, I, Comm. Pure Appl. Math., **36**(**6**) (1983), 767-783, MR 85d:34024, Zbl 0517.58031.

[9] E.L. Isaacson, H.P. McKean and E. Trubowitz, *The inverse Sturm-Liouville problem*, II, Comm. Pure Appl. Math., **37**(**1**) (1984), 1-11, MR 86f:34051, Zbl 0517.58031.

[10] A.R. Its and V.B. Matveev, *Hill's operator with finitely many gaps*, Funktsional'nyi Analiz i ego Prilozheniya, **9**(**1**) (1975), 69-70.

[11] N. Levinson, *The inverse Sturm-Liouville problem*, Mat. Tidsskr., **B** (1949), 25-30, MR 11,248e, Zbl 0045.36402.

[12] W. Magnus and S. Winkler, *Hill's Equation*, Dover Publications, Inc., New York, 1966, MR 33 #5991, Zbl 0158.09604.

[13] A.I. Markushevich, *Theory of Functions of a Complex Variable*, Chelsea Publishing Company, NY, 1965, MR 30 #2125, Zbl 0135.12002.

[14] H.P. McKean, *Integrable Systems and Algebraic Curves*, Global Analysis: Lecture Notes in Mathematics, **755**, Springer-Verlag, 1979, MR 81g:58017, Zbl 0449.35080.

[15] H.P. McKean and E. Trubowitz, *Hill's operator and hyperelliptic function theory in the presence of infinitely many branch points*, Comm. Pure Appl. Math., **29** (1976), 143-226, MR 55 #761, Zbl 0339.34024.

[16] J. Pöschel and E. Trubowitz, *Inverse Spectral Theory*, Academic Press, Inc., Orlando, 1987, MR 89b:34061, Zbl 0623.34001.

[17] E. Trubowitz, *The inverse problem for periodic potentials*, Comm. Pure Appl. Math., **30** (1977), 321-337, MR 55 #3408, Zbl 0403.34022.

Received October 20, 2000 and revised December 15, 2000.

Department of Mathematics
University of Arizona
Tucson, AZ 85721
E-mail address: vpierce@math.arizona.edu

CONTENTS

Volume 204, no. 1 and no. 2

Guidelines for Authors

Authors may submit manuscripts at pjm.math.berkeley.edu/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use LaTeX, but papers in other varieties of TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PACIFIC JOURNAL OF MATHEMATICS

Volume 204 No. 2 June 2002

0030-8730(200206)204:2;1-Q