THE DISCRIMINANT OF A SYMPLECTIC INVOLUTION

Grégory Berhuy, Marina Monsurrò, and Jean-Pierre Tignol

An invariant for symplectic involutions on central simple algebras of degree divisible by 4 over fields of characteristic different from 2 is defined on the basis of Rost’s cohomological invariant of degree 3 for torsors under symplectic groups. We relate this invariant to trace forms and show how its triviality yields a decomposability criterion for algebras of degree 8 with symplectic involution.

1. Introduction and statement of results.

In contrast with orthogonal involutions, for which invariants corresponding to the discriminant and Clifford algebras of quadratic forms are defined, no “classical” invariant is known for symplectic involutions on central simple algebras, besides the signature (see [6, (11.10)]). Using the cohomological invariant of degree 3 defined by Rost for torsors under simply connected absolutely simple linear algebraic groups, we introduce an invariant of symplectic involutions on central simple algebras of degree a multiple of 4 with values in the third Galois cohomology group of the center with coefficients \( \{\pm 1\} \) and give an alternative description in terms of trace forms. We call this invariant the discriminant since it is the first nontrivial invariant, and because it is directly linked to the discriminant of Hermitian forms, see Example 2. Even though its definition is elementary, Rost’s computation of the invariants of torsors under symplectic groups is needed to prove that there is no other cohomological invariant of degree 3 and to establish the relationship with trace forms. In the final section, we prove that symplectic involutions with trivial discriminant on central simple algebras of degree 8 and index 4 afford a special type of decomposition. In a sequel to this paper, the discriminant is used to give examples of non \( R \)-trivial adjoint symplectic groups of even index.

1.1. Definition of the discriminant. Throughout this paper, \( F \) denotes a field of characteristic different from 2. Let \( A \) be a finite-dimensional central simple \( F \)-algebra, and \( \theta: A \to A \) be an anti-automorphism of order 2. We recall that \( \theta \) is called a symplectic involution on \( A \) if, after scalar extension to a splitting field, \( \theta \) is adjoint to an alternating form, see [6, (2.5)]. From now on, we suppose the involution \( \theta \) is of this type. In this case, the degree \( \deg A \) is necessarily an even integer \( n = 2m \).
The symplectic group \( \text{Sp}(A, \theta) \) is the group scheme over \( F \) defined by
\[
\text{Sp}(A, \theta)(E) = \{ x \in A \otimes_F E \mid \theta(x)x = 1 \}
\]
for any commutative \( F \)-algebra \( E \).

Let \( \text{Sym}(A, \theta) \) be the \( F \)-vector space of elements in \( A \) fixed by \( \theta \),
\[
\text{Sym}(A, \theta) = \{ x \in A \mid \theta(x) = x \}.
\]
We denote by \( \text{Sym}(A, \theta)^\times \) the set of units in \( \text{Sym}(A, \theta) \),
\[
\text{Sym}(A, \theta)^\times = \text{Sym}(A, \theta) \cap A^\times.
\]

We recall that the \textit{pfaffian reduced norm} is the homogeneous polynomial function of degree \( m \)
\[
\text{Nrp}_\theta : \text{Sym}(A, \theta) \to F
\]
uniquely determined by the following conditions:
\[
\text{Nrp}_\theta(1) = 1 \quad \text{and} \quad \text{Nrp}_\theta(x)^2 = \text{Nrd}_A(x) \quad \text{for } x \in \text{Sym}(A, \theta),
\]
see \cite[p. 19]{6}.

The cohomology set \( H^1(F, \text{Sp}(A, \theta)) \) can be represented as
\[
H^1(F, \text{Sp}(A, \theta)) \simeq \text{Sym}(A, \theta)^\times / \sim
\]
where \( \sim \) is the equivalence relation defined by \( x \sim y \) if and only if there exists \( u \in A^\times \) such that \( y = ux\theta(u) \), see \cite[(29.24)]{6}.

Let \( G_m \) be the multiplicative group. The Kummer exact sequence
\[
1 \to \mu_2 \to G_m \xrightarrow{2} G_m \to 1
\]
allows us to identify the cohomology sets \( H^1(F, \mu_2) \) and \( H^2(F, \mu_2) \) respectively with the quotient \( F^\times / F^\times 2 \) and with the 2-torsion subgroup of the Brauer group. For all \( x \in F^\times \), we denote by \( (x)_2 \in H^1(F, \mu_2) \) the cohomology class associated to \( xF^\times 2 \). Similarly, we denote by \( [A] \in H^2(F, \mu_2) \) the cohomology class associated to the Brauer class of \( A \). We define
\[
\Delta_\theta : \text{Sym}(A, \theta)^\times \to H^3(F, \mu_2)
\]
as the map given by the cup-product
\[
\Delta_\theta(s) = ([\text{Nrp}_\theta(s)]_2 \cup [A]).
\]

It follows from the properties of \( \text{Nrp}_\theta \) (see the proof of Proposition 1 below) that \( \Delta_\theta \) is well-defined on the set of equivalence classes under the relation \( \sim \). The induced map on the quotient can be interpreted under the bijection (1) as the Rost invariant of \( H^1(F, \text{Sp}(A, \theta)) \), see \cite[p. 440]{6}.

Since \( \text{Nrp}_\theta \) is homogeneous of degree \( m \), we obtain, for \( \alpha \in F^\times \) and \( s \in \text{Sym}(A, \theta)^\times \), the following relation:
\[
\Delta_\theta(\alpha s) = ([\alpha^m \text{Nrp}_\theta(s)]_2 \cup [A]) = \begin{cases} 
\Delta_\theta(s) & \text{if } m \text{ is even}, \\
\Delta_\theta(s) + (\alpha)_2 \cup [A] & \text{if } m \text{ is odd}.
\end{cases}
\]
Therefore, if \( m \) is even one can define a relative invariant for symplectic involutions on \( A \) as follows:

**Definition.** Let \( A \) be a central simple algebra over \( F \) of degree \( n = 2m \equiv 0 \mod 4 \). Let \( \theta \) and \( \sigma \) be symplectic involutions on \( A \). There exists (see [6, (2.7)]) \( s \in \text{Sym}(A, \theta)^{\times} \) such that

\[
\sigma = \text{Int}(s) \circ \theta
\]

where \( \text{Int}(s) \) denotes the inner automorphism associated with \( s \),

\[
\text{Int}(s)(x) = sx^{-1} \quad \text{for} \quad x \in A.
\]

The element \( s \) is uniquely determined up to multiplication by an element of \( F^{\times} \). By (2), it follows that \( \Delta_{\theta}(s) \in H^3(F, \mu_2) \) only depends on \( \sigma \), since \( m \) is even. We call this element the **discriminant** of \( \sigma \) with respect to \( \theta \) and denote it by \( \Delta_{\theta}(\sigma) \). Thus,

\[
\Delta_{\theta}(\sigma) = \left(\text{Nrp}_{\theta}(s)\right)^2 \cup [A] \in H^3(F, \mu_2).
\]

In the case \( m = 2 \), an analogue of this invariant has been studied in [6, §16.B], where it is denoted by \( j_{\theta}(\sigma) \). Theorem (16.19) of [6] shows that this invariant classifies, up to conjugation, symplectic involutions on a central simple algebra of degree 4.

In Section 2 we establish the following elementary result:

**Proposition 1.**

(a) The discriminant \( \Delta_{\theta}(\sigma) \) only depends on the conjugacy classes of \( \theta \) and \( \sigma \), namely, if \( u, v \in A^{\times} \) and

\[
\theta' = \text{Int}(u) \circ \theta \circ \text{Int}(u)^{-1}, \quad \sigma' = \text{Int}(v) \circ \sigma \circ \text{Int}(v)^{-1},
\]

then

\[
\Delta_{\theta}(\sigma') = \Delta_{\theta}(\sigma).
\]

In particular, if \( \sigma \) and \( \theta \) are conjugate, then \( \Delta_{\theta}(\sigma) = 0 \).

(b) Let \( \rho, \sigma \) and \( \theta \) be symplectic involutions on \( A \); then

\[
\Delta_{\rho}(\sigma) = \Delta_{\rho}(\theta) + \Delta_{\theta}(\sigma) \quad \text{and} \quad \Delta_{\theta}(\sigma) = \Delta_{\sigma}(\theta).
\]

If the Schur index \( \text{ind} A \) divides \( \frac{1}{2} \deg A \), i.e., if \( A \simeq M_2(A_0) \) for some central simple \( F \)-algebra \( A_0 \), then \( A \) carries hyperbolic symplectic involutions, such as \( \gamma \otimes \theta_0 \), where \( \gamma \) is the (unique) symplectic involution on \( M_2(F) \) and \( \theta_0 \) is an arbitrary orthogonal involution on \( A_0 \). Since all hyperbolic involutions are pairwise conjugate, we may set \( \Delta = \Delta_{\theta} \) for any hyperbolic symplectic involution \( \theta \).

**Example 2.** Consider the algebra \( A = \text{End}_Q V \), where \( Q \) is a quaternion division \( F \)-algebra and \( V \) is an \( m \)-dimensional \( Q \)-vector space. Symplectic involutions on \( A \) are then adjoint to Hermitian forms on \( V \) with respect to
the conjugation involution on $Q$. Suppose that $m$ is even and let $\sigma$ be the involution adjoint to a fixed Hermitian form $h$ on $V$. Let

$$h = \langle \alpha_1, \ldots, \alpha_m \rangle$$

be the diagonalization of $h$ relative to some orthogonal basis $e$ of $V$ ($\alpha_1, \ldots, \alpha_m \in F^\times$), then

$$\Delta(\sigma) = ((-1)^{m/2} \alpha_1 \ldots \alpha_m)^2 \cup [Q].$$

Indeed, let $\theta$ be the hyperbolic involution adjoint to the Hermitian form with diagonalization $\langle 1, -1, \ldots, 1, -1 \rangle$ relative to the basis $e$. Then, identifying $A$ with $M_m(Q)$ by $e$, we get

$$\sigma = \mathrm{Int} \: \mathrm{diag}(\alpha_1, -\alpha_2, \ldots, \alpha_{m-1}, -\alpha_m) \circ \theta,$$

and we can compute $\Delta(\sigma) = \Delta(\theta) \circ (\sigma)$ by Lemma 9(e) below.

Note that if $V_0 \subset V$ is the $F$-subspace spanned by $e$, then $A = \mathrm{End}_F V_0 \otimes Q$, and we obtain a decomposition $\sigma = \sigma_0 \otimes \gamma$ where $\sigma_0$ is the involution adjoint to the bilinear form on $V_0$ with diagonalization $\langle \alpha_1, \ldots, \alpha_m \rangle$ relative to $e$, and $\gamma$ is the canonical (conjugation) involution on $Q$.

This example can be slightly generalized:

**Example 3.** Consider the algebra $A = A_0 \otimes_F Q$ where $Q$ is a quaternion $F$-algebra and $A_0$ is a central simple $F$-algebra. Let $\sigma_0$ be an orthogonal involution on $A_0$, $\gamma$ the canonical involution on $Q$, and

$$\sigma = \sigma_0 \otimes \gamma.$$ 

Suppose that $\text{ind} \: A_0$ divides $\frac{1}{2} \deg A_0$. Then

$$\Delta(\sigma) = (\text{disc} \: \sigma_0)_2 \cup [Q],$$

where $\text{disc} \: \sigma_0 \in F^\times/F^\times 2$ is the discriminant of the orthogonal involution $\sigma_0$ (see [6, §7]). Indeed, let $\theta_0$ be a hyperbolic orthogonal involution on $A_0$ and let $x_0 \in \text{Sym}(A_0, \theta_0)^\times$ be such that $\sigma_0 = \text{Int}(x_0) \circ \theta_0$. The involution $\theta = \theta_0 \otimes \gamma$ is hyperbolic, and we have $\sigma = \text{Int}(x_0 \otimes 1) \circ \theta$, so that

$$\Delta(\sigma) = (\text{Nrd}_{A_0}(x_0 \otimes 1))^2 \cup [A].$$

Now, by Lemma 9(d), $\text{Nrd}_{A_0}(x_0 \otimes 1) = \text{Nrd}_{A_0}(x_0)$. Equation (3) follows, since $\text{disc} \: \sigma_0$ is represented by $\text{Nrd}_{A_0}(x_0)$, and $(\text{Nrd}_{A_0}(x_0))^2 \cup [A_0] = 0$.

### 1.2. Trace forms.

Let $A$ be an arbitrary central simple $F$-algebra. For every involution $\sigma : A \to A$, the associated trace form $T_\sigma : A \to F$ is defined as follows:

$$T_\sigma(x) = \text{Trd}_A(\sigma(x)x)$$
where Trd$_A$ denotes the reduced trace. Denote by T$^+_\sigma$ the restriction of T$_\sigma$ to Sym(A, $\sigma$); this form can also be seen as the restriction to Sym(A, $\sigma$) of the form T$_A$: A $\rightarrow$ F defined by

$$T_A(x) = \text{Trd}_A(x^2).$$

As is the case with involutions of the other types (see [6, §11]), the discriminant of symplectic involutions can be expressed in terms of trace forms; indeed we have the following result:

**Theorem 4.** Let A be a central simple algebra over F and let $\theta$ and $\sigma$ be symplectic involutions on A. The class in the Witt ring WF of the difference T$^+_\sigma$ − T$^+_\theta$ lies in the third power of the fundamental ideal, namely

$$T^+_\sigma - T^+_\theta \in I^3F.$$

Moreover, if $e_3: I^3F \rightarrow H^3(F, \mu_2)$ denotes the Arason invariant, we obtain

$$e_3(T^+_\sigma - T^+_\theta) = \begin{cases} \Delta_\theta(\sigma) & \text{if } \deg A \equiv 0 \mod 4, \\ 0 & \text{if } \deg A \equiv 2 \mod 4. \end{cases}$$

A proof of this result is given in Section 3 below. For the trace forms T$^+_{\sigma}$, we have the following result:

**Corollary 5.** Keeping the notation of the previous theorem, we have T$^+_{\sigma}$ − T$^+_{\theta}$ ∈ I$^4F$ and

$$e_4(T^+_{\sigma} - T^+_{\theta}) = \begin{cases} (-1)^2 \cup \Delta_\theta(\sigma) & \text{if } \deg A \equiv 0 \mod 4, \\ 0 & \text{if } \deg A \equiv 2 \mod 4, \end{cases}$$

where $e_4: I^4F \rightarrow H^4(F, \mu_2)$ denotes the degree 4 invariant.

**Proof.** Let $T^-_{\sigma}$ be the restriction of T$^-_{\sigma}$ (or of $-T_A$) to the space of skew-symmetric elements in A. We have

$$T^-_{\sigma} = T^+_\sigma + T^+_\sigma$$

and

$$T_A = T^+_\sigma - T^+_\sigma,$$

so that $T^+_\sigma = 2T^+_\sigma - T_A$. Similarly, $T^+_{\theta} = 2T^+_{\theta} - T_A$, so that

$$T^+_{\sigma} - T^+_{\theta} = 2(T^+_{\sigma} - T^+_{\theta}),$$

hence the corollary is a direct consequence of the previous theorem. \[\square\]

In the special case where $\theta$ is hyperbolic we get:

**Proposition 6.** Suppose $A = M_2(A_0)$ for some central simple F-algebra $A_0$, and let $\theta$ be a hyperbolic symplectic involution on A. Then $T^+_{\theta}$ is Witt-equivalent to $\langle 2 \rangle \cdot T_{A_0}$, and $T_{\theta}$ is hyperbolic. If $\deg A \equiv 2 \mod 4$, then A is split, hence every symplectic involution on A is hyperbolic. If $\deg A \equiv 0 \mod 4$, then, for any symplectic involution $\sigma$ on A, we have $T^+_{\sigma} \in I^4F$ and

$$e_4(T^+_{\sigma}) = (-1)^2 \cup \Delta(\sigma).$$

The proof is at the end of Section 3.
1.3. Decomposability of symplectic involutions. Section 4 below will be devoted to the relations between the discriminant and the decomposability of symplectic involutions as tensor products of involutions defined on subalgebras. Our main result is concerned with degree 8 algebras with index dividing 4. Such algebras can be written in the form \( A = M_2(A_0) \), where \( A_0 \) is a central simple algebra of degree 4, hence they carry hyperbolic symplectic involutions. The case \( \text{ind} A = 1 \) is trivial, since every symplectic involution on a split algebra is hyperbolic, and is omitted in the following theorem:

**Theorem 7.** Let \( A \) be a central simple \( F \)-algebra of degree 8 having index 2 or 4. For any symplectic involution \( \sigma \) on \( A \), there is a decomposition

\[
(A, \sigma) = (Q, \gamma) \otimes_F (A_0, \sigma_0)
\]

where \( Q \) is a quaternion subalgebra, \( A_0 \) is its centralizer (which is a central simple \( F \)-subalgebra of degree 4 in \( A \)), \( \gamma \) is the conjugation involution on \( Q \) and \( \sigma_0 \) is an orthogonal involution on \( A_0 \).

When \( \text{ind} A = 2 \), this theorem is easily proved and can be readily generalized to any degree, see Example 2 or [1, Proposition 3.4]. The case \( \text{ind} A = 4 \) is treated in Section 4.

Theorem 7 shows that the discriminant of a symplectic involution \( \sigma \) on a central simple \( F \)-algebra of degree 8 and index 2 or 4 can be computed as in Example 3 above. The following theorem gives a necessary and sufficient condition for the discriminant to be trivial.

**Theorem 8.** Let \( A \) be a central simple \( F \)-algebra of degree 8 with index dividing 4. For any symplectic involution \( \sigma \) on \( A \), \( \Delta(\sigma) = 0 \) if and only there is a decomposition

\[
(A, \sigma) = (A_1, \sigma_1) \otimes_F (A_2, \sigma_2) \otimes_F (A_3, \gamma_3)
\]

where \( A_1, A_2, A_3 \) are quaternion subalgebras of \( A \), \( \sigma_1, \sigma_2 \) are orthogonal involutions on \( A_1 \) and \( A_2 \) respectively, \( \gamma_3 \) is the conjugation involution on \( A_3 \), and \( A_1 \) is split,

\[
A_1 \simeq M_2(F).
\]

A proof is given in Section 4.

2. Discriminants and Pfaffian norms.

The goal of this section is to prove Proposition 1. Throughout the section, \( A \) denotes a central simple \( F \)-algebra of degree \( n = 2m \).

**Lemma 9.** Let \( \sigma \) and \( \theta \) be symplectic involutions on \( A \) and let \( s \) be an element in \( \text{Sym}(A, \theta)^\times \) such that \( \sigma = \text{Int}(s) \circ \theta \). Then:
(a) For every \(x \in \text{Sym}(A, \sigma) \cap \text{Sym}(A, \theta)\),
\[ \text{Nrp}_\sigma(x) = \text{Nrp}_\theta(x). \]

(b) For every \(x \in \text{Sym}(A, \theta)\), the product \(sx\) lies in \(\text{Sym}(A, \sigma)\) and
\[ \text{Nrp}_\sigma(sx) = \text{Nrp}_\theta(s) \text{Nrp}_\theta(x). \]

(c) For every \(x \in \text{Sym}(A, \theta)\),
\[ \text{Nrp}_\theta(x^{-1}) = \text{Nrp}_\theta(x)^{-1}. \]

(d) Suppose \(A = A_1 \otimes_F A_2\) for some central simple \(F\)-algebras \(A_1, A_2 \subset A\) of degree \(n_1 = 2m_1\) and \(n_2 = 2m_2\) respectively; if \(x_1 \in A_1\) and \(x_2 \in A_2\) are such that \(x_1 \otimes x_2 \in \text{Sym}(A, \theta)\), then
\[ \text{Nrp}_\theta(x_1 \otimes x_2) = \text{Nrd}_{A_1}(x_1)^{m_2} \text{Nrd}_{A_2}(x_2)^{m_1}. \]

(e) Suppose \(A = M_r(A_0)\) and \(\theta((a_{ij})_{1 \leq i,j \leq r}) = (\theta_0(a_{ij}))_{1 \leq i,j \leq r}\) for some symplectic involution \(\theta_0\) on the central simple \(F\)-algebra \(A_0\). For the diagonal matrix \(x = \text{diag}(x_1, \ldots, x_r)\) with \(x_i \in \text{Sym}(A_0, \theta_0)\) for \(i = 1, \ldots, r\), we have
\[ \text{Nrp}_\theta(x) = \text{Nrp}_{\theta_0}(x_1) \cdots \text{Nrp}_{\theta_0}(x_r). \]

Proof. (a) Let \(t\) be an indeterminate over \(F\). Define
\[ \text{Prp}_\sigma(x, t) = \text{Nrp}_\sigma(t - x) \in F[t], \quad \text{Prp}_\theta(x, t) = \text{Nrp}_\theta(t - x) \in F[t]. \]
Those polynomials, called pfaffian characteristic polynomials in [6, p. 19], are monic and satisfy
\[ \text{Prp}_\sigma(x, t)^2 = \text{Pcrd}_{A,x} \text{Prp}_\theta(x, t)^2, \]
where \(\text{Pcrd}_{A,x}(t) = \text{Nrd}_{A(t)}(t - x)\) is the reduced characteristic polynomial of \(x\). Therefore, \(\text{Prp}_\sigma(x, t) = \text{Prp}_\theta(x, t)\), and evaluation at \(t = 0\) yields \(\text{Nrp}_\sigma(x) = \text{Nrp}_\theta(x)\).

(b) Straightforward calculations show that \(\sigma(sx) = sx\) if \(\theta(x) = x\). Let us consider the two sides of the equality we aim to prove as polynomial functions of \(x\). The squares of the two sides are equal since the reduced norm is multiplicative, hence they are equal up to sign. Moreover, they are equal and nonzero for \(x = 1\) in view of Part (a). Hence, they are equal for all \(x\).

(c) We apply (b) with \(x = s^{-1}\) and use the relation \(\text{Nrp}_\sigma(1) = 1\).

(d) By taking the square root on both sides of the equation
\[ \text{Pcrd}_{A,x_1 \otimes x_2} = (\text{Pcrd}_{A_1,x_1})^{n_2}(\text{Pcrd}_{A_2,x_2})^{n_1}, \]
we obtain
\[ \text{Prp}_{\theta,x_1 \otimes x_2} = (\text{Pcrd}_{A_1,x_1})^{m_2}(\text{Pcrd}_{A_2,x_2})^{m_1}. \]
The property follows by considering the constant terms.
(e) As in the preceding case, the property follows by extracting the monic square root of each side of the equation

$$P_{crd} A, x = P_{crd} A_0, x_1 \ldots P_{crd} A_0, x_r .$$

\[\square\]

Proposition 1 easily follows from the lemma above. Indeed, if \( \sigma = \text{Int}(s) \circ \theta \), so that \( \Delta_\theta(\sigma) = (\text{Nrp}_\theta(s))^2 \cup [A] \), then \( \theta = \text{Int}(s^{-1}) \circ \sigma \) and hence \( \Delta_\sigma( \theta ) = (\text{Nrp}_\sigma(s^{-1}))^2 \cup [A] \). Now Lemma 9 shows that

\[\text{Nrp}_\sigma(s^{-1}) = \text{Nrp}_\theta(s)^{-1},\]

hence \((\text{Nrp}_\sigma(s^{-1}))^2 = (\text{Nrp}_\theta(s))^2\), and so \( \Delta_\sigma( \theta ) = \Delta_\theta( \sigma ) \). If \( \rho \) is another symplectic involution, and if \( t \in \text{Sym}(A, \rho) \) is such that \( \theta = \text{Int}(t) \circ \rho \), then \( \sigma = \text{Int}(st) \circ \rho \). Part (b) of Lemma 9 yields

\[\text{Nrp}_\sigma(st) = \text{Nrp}_\theta(st) \text{Nrp}_\rho(t).\]

Moreover, Part (a) shows that \( \text{Nrp}_\rho(st) = \text{Nrp}_\rho(st) \) and \( \text{Nrp}_\rho(t) = \text{Nrp}_\rho(t) \). Therefore, the preceding equality can be written as

\[\text{Nrp}_\rho(st) = \text{Nrp}_\theta(st) \text{Nrp}_\rho(t).\]

It follows that

\[\Delta_\rho(\sigma) = (\text{Nrp}_\rho(st))^2 \cup [A]
= (\text{Nrp}_\theta(s))^2 \cup [A] + (\text{Nrp}_\rho(t))^2 \cup [A] = \Delta_\theta(\sigma) + \Delta_\rho(\theta),\]

which completes the proof of Part (b) of Proposition 1.

Now let \( v \in \mathbb{A}^\times \) and \( \sigma' = \text{Int}(v) \circ \sigma \circ \text{Int}(v)^{-1} \), so that \( \sigma' = \text{Int}(v s \theta(v)) \circ \sigma \). Then,

\[\Delta_\theta(\sigma') = (\text{Nrp}_\theta(v s \theta(v)))^2 \cup [A].\]

By [6, (2.13)], \( \text{Nrp}_\theta(v s \theta(v)) = \text{Nrd}_A(v) \text{Nrp}_\theta(s) \). Since \((\text{Nrd}_A(v))^2 \cup [A] = 0\), it follows that

\[\Delta_\theta(\sigma') = \Delta_\theta(\sigma).\]

Similarly, if \( \theta' \) is a symplectic involution conjugate to \( \theta \), then \( \Delta_{\theta'}(\theta') = \Delta_{\sigma'}(\theta) \). Now, Part (b) of Proposition 9 shows that \( \Delta_{\theta'}(\sigma') = \Delta_{\sigma'}(\theta') \) and \( \Delta_{\theta}(\sigma') = \Delta_{\sigma'}(\theta) \). Therefore,

\[\Delta_{\theta'}(\sigma') = \Delta_{\theta}(\sigma').\]

We already observed that \( \Delta_{\theta}(\sigma') = \Delta_{\theta}(\sigma) \), hence

\[\Delta_{\theta'}(\sigma') = \Delta_{\theta}(\sigma)\]

and the proof of Proposition 1 is complete.
3. Discriminant and trace form.

In this section we prove Theorem 4 and Proposition 6.

Let $\text{Fields}_F$ be the category of fields containing $F$ and let $G$ be any algebraic group over $F$. We consider the functor

$$H^1(G): \text{Fields}_F \to \text{Sets}^*$$

where $\text{Sets}^*$ denotes the category of pointed sets, associating to every $L \in \text{Fields}_F$ the Galois cohomology set $H^1(L, G)$.

For any integer $d \geq 0$, we let $\mathbb{Q}/\mathbb{Z}(d-1) = \varprojlim \mu_{n}^{\otimes (d-1)}$, where $\mu_n$ is the group of $n$-th roots of unity in a separable closure of $F$. We may then consider the functor

$$H^d(\mathbb{Q}/\mathbb{Z}(d-1)): \text{Fields}_F \to \text{Sets}^*$$

which carries $L \in \text{Fields}_F$ to the Galois cohomology group $H^d(L, \mathbb{Q}/\mathbb{Z}(d-1))$ (and forgets the group structure). The natural transformations $H^1(G) \to H^d(\mathbb{Q}/\mathbb{Z}(d-1))$ are called cohomological invariants of dimension (or degree) $d$ in [6, §31.B]. Since $H^d(L, \mathbb{Q}/\mathbb{Z}(d-1))$ is a group for $L \in \text{Fields}_F$, these invariants form a group. In reference to Rost’s “cohomological cycle module” $M = \bigoplus_{d \geq 0} H^d(\mathbb{Q}/\mathbb{Z}(d-1))$ (see [8]), we denote it simply by $\text{Inv}^d(H^1(G), M)$. (This group is denoted by $\text{Inv}^d(G, \mathbb{Q}/\mathbb{Z}(d-1))$ in [6, §31.B].)

Now let $A$ be a central simple $F$-algebra of degree $n = 2m$ and let $\theta$ be a symplectic involution on $A$. We take for $G$ the group $\text{GSp}(A, \theta)$ of symplectic similitudes; this is the algebraic group scheme defined by

$$\text{GSp}(A, \theta)(E) = \{ g \in A \otimes_F E \mid \theta(g)g \in E^x \}$$

for any commutative $F$-algebra $E$. The set $H^1(L, \text{GSp}(A, \theta))$ is in one-to-one correspondence with the set of conjugacy classes of symplectic involutions defined on $A_L = A \otimes_F L$, the class of $\theta$ being the distinguished one (see [6, (29.23)])]. The following proposition shows that for symplectic involutions there is no (nontrivial) cohomological invariant of degree 1 or 2.

**Proposition 10.** If the algebra $A$ is split, we have $H^1(L, \text{GSp}(A, \theta)) = 1$ for every $L \in \text{Fields}_F$, so that

$$\text{Inv}^d(H^1(\text{GSp}(A, \theta)), M) = 0 \quad \text{for all } d.$$  

If $A$ is not split, we have

$$\text{Inv}^d(H^1(\text{GSp}(A, \theta)), M) = 0 \quad \text{for } d = 1, 2$$

and

$$\text{Inv}^3(H^1(\text{GSp}(A, \theta)), M) = \begin{cases} 0 & \text{if } \deg A \equiv 2 \mod 4, \\ \mathbb{Z}/2\mathbb{Z} & \text{if } \deg A \equiv 0 \mod 4. \end{cases}$$
Proof. Every symplectic involution on a split algebra is hyperbolic. Therefore, when \( A \) is split, \( H^1(L, \text{GSp}(A, \theta)) = 1 \) for all \( L \in \text{Fields}_F \).

For the rest of the proof, we may thus assume that \( A \) is not split. Let \( \mu : \text{GSp}(A, \theta) \to G_m \) be the homomorphism which associates to each similitude \( g \) its multiplier \( \mu(g) = \theta(g)g \). The cohomology sequence induced by the exact sequence

\[
1 \to \text{Sp}(A, \theta) \to \text{GSp}(A, \theta) \overset{\mu}{\to} G_m \to 1
\]

yields for every \( L \in \text{Fields}_F \) the exact sequence

\[
L^\times \to H^1(L, \text{Sp}(A, \theta)) \to H^1(L, \text{GSp}(A, \theta)) \to 1
\]

since \( H^1(L, G_m) = 1 \) by Hilbert’s Theorem 90. Therefore, for every \( d \), we have an exact sequence

\[
0 \to \text{Inv}^d(H^1(\text{Sp}(A, \theta)), M) \to \text{Inv}^d(H^1(\text{Sp}(A, \theta)), M) \to \text{Inv}^d(G_m, M).
\]

For \( d = 1 \) or \( 2 \), we obtain, by [6, (31.15)], \( \text{Inv}^d(H^1(\text{Sp}(A, \theta)), M) = 0 \) and hence

\[
\text{Inv}^d(H^1(\text{GSp}(A, \theta)), M) = 0.
\]

The group \( \text{Inv}^3(H^1(\text{Sp}(A, \theta)), M) \) is of order 2, the nontrivial element being the Rost invariant \( \Delta_\theta \) defined in the introduction. Equation (2) shows that this invariant is zero in \( \text{Inv}^3(G_m, M) \) if and only if \( \deg A \equiv 2 \mod 4 \). \( \square \)

When \( \deg A \equiv 0 \mod 4 \), the unique nontrivial invariant of degree 3 is the discriminant. Our next goal is to give an explicit description of this invariant in terms of trace forms.

Let \( T_\theta^+ : \text{Sym}(A, \theta) \to F \) be the quadratic form

\[
T_\theta^+(x) = \text{Trd}_A(\theta(x)x) = \text{Trd}_A(x^2).
\]

This forms only depends, up to isometry, on the conjugacy class of \( \theta \) since, if \( \theta' = \text{Int}(v) \circ \theta \circ \text{Int}(v)^{-1} \) for some \( v \in A^\times \), then \( \text{Int}(v) \) defines an isometry between \( T_\theta^+ \) and \( T_{\theta'}^+ \). Consider \( L \in \text{Fields}_F \). The map sending every symplectic involution \( \sigma : A_L \to A_L \) to the discriminant

\[
\text{disc}(T_\sigma^+ - T_\theta^+) \in L^\times / L^\times 2 = H^1(L, \mu_2)
\]

defines a cohomological invariant \( H^1(\text{GSp}(A, \theta)) \to H^1(\mu_2) \). By Proposition 10, this invariant is trivial, hence \( T_\sigma^+ - T_\theta^+ \in I^2 L \). Similarly, the map sending every symplectic involution \( \sigma \) to the Witt (-Clifford) invariant

\[
e_2(T_\sigma^+ - T_\theta^+) \in H^2(L, \mu_2)
\]

defines a cohomological invariant of degree 2. Again, by Proposition 10, we get \( e_2(T_\sigma^+ - T_\theta^+) = 0 \), and hence \( T_\sigma^+ - T_\theta^+ \in I^3 L \) using Merkurjev’s theorem. This proves the first part of Theorem 4. Note that the equality
$e_2(T^+_\sigma - T^+_\theta) = 0$ can also be derived from Quéguiner’s explicit calculation of the Hasse invariant of trace forms in [7, p. 307].

Consider the map associating to every symplectic involution $\sigma : A_L \to A_L$ the Arason invariant

$$e_3(T^+_\sigma - T^+_\theta) \in H^3(L, \mu_2).$$

Using Proposition 10, we see that this invariant is trivial if $\deg A \equiv 2 \mod 4$ or if $A$ is split. We claim that it coincides with the discriminant $\Delta_{\theta}(\sigma)$ if $A$ is nonsplit and $\deg A \equiv 0 \mod 4$. To prove this, it suffices to show that it is nontrivial because, by Proposition 10, there is a unique nontrivial invariant in $\text{Inv}^3(H^1(\text{GSp}(A,\theta)), M)$. Therefore, our goal is to find a field $L \in \text{Fields}_F$ and two symplectic involutions $\sigma_1, \sigma_2$ on $A_L$ such that $e_3(T^+_\sigma_1 - T^+_\sigma_2) \neq 0$. If $\theta$ is any involution on $A$, the equality

$$e_3(T^+_\sigma_1 - T^+_\theta) = e_3(T^+_\sigma_1 - T^+_\sigma_2) + e_3(T^+_\sigma_2 - T^+_\theta)$$

shows that at least one of the terms $e_3(T^+_\sigma_1 - T^+_\theta)$ and $e_3(T^+_\sigma_2 - T^+_\theta)$ is nonzero, hence the invariant $\sigma \mapsto e_3(T^+_\sigma - T^+_\theta)$ is nontrivial.

After scalar extension to the function field of a suitable generalized Severi-Brauer variety (see [2]), we may assume that $\text{ind } A = 2$ i.e., that $A$ is Brauer-equivalent to a quaternion division algebra $Q$ over $F$. Then, denoting by $V$ an $F$-vector space of dimension $m$, we obtain

$$A \simeq Q \otimes_F \text{End}_F V.$$

For the rest of this section, we fix an isomorphism identifying $A$ with $Q \otimes \text{End}_F V$. Let $b$ be a symmetric nondegenerate bilinear form on $V$. The symmetric square $S^2 V$ and the exterior square $\wedge^2 V$ are endowed with symmetric bilinear forms $b^{S^2}$ and $b^{\wedge^2}$ respectively, defined by

$$b^{S^2}(x_1 \cdot x_2, y_1 \cdot y_2) = b(x_1, y_1)b(x_2, y_2) + b(x_1, y_2)b(x_2, y_1)$$

and

$$b^{\wedge^2}(x_1 \wedge x_2, y_1 \wedge y_2) = b(x_1, y_1)b(x_2, y_2) - b(x_1, y_2)b(x_2, y_1).$$

**Lemma 11.** Let $Q = (\alpha, \beta)_F$. On $A = Q \otimes_F \text{End}_F V$, consider the symplectic involution $\sigma = \gamma \otimes \text{ad}_b$, where $\gamma$ is the quaternion conjugation on $Q$ and $\text{ad}_b$ is the (orthogonal) involution adjoint to $b$. Then, the bilinear form $B^+_\sigma(x, y) = \text{Trd}_A(xy)$ on $\text{Sym}(A, \sigma)$ (which is the polar form of the quadratic form $T^+_\sigma$) decomposes as an orthogonal sum

$$B^+_\sigma = b^{S^2} \perp (-\alpha, -\beta, \alpha \beta) \cdot b^{\wedge^2}.$$

**Proof.** Let $\text{Skew}(\text{End}_F V, \text{ad}_b)$ be the $F$-vector space of endomorphisms $f$ of $V$ such that $\text{ad}_b(f) = -f$, and let $Q^0$ be the $F$-vector space of pure quaternions in $Q$. A straightforward calculation shows that the decomposition

$$\text{Sym}(A, \sigma) = (F \otimes \text{Sym}(\text{End}_F V, \text{ad}_b)) \oplus (Q^0 \otimes \text{Skew}(\text{End}_F V, \text{ad}_b))$$
is orthogonal with respect to the form $B_\theta^+$. Let $B_\theta^+$ and $B_\theta^-$ be the restrictions of the bilinear trace form $B(f,g) = \text{tr}(fg)$ to $\text{Sym}(\text{End}_F V, \text{ad}_b)$ and $\text{Skew}(\text{End}_F V, \text{ad}_b)$ respectively. The decomposition above yields

$$B_\theta^+ = \langle 2 \rangle \cdot B_\theta^+ \perp \langle 2\alpha, 2\beta, -2\alpha\beta \rangle \cdot B_\theta^-.$$  

The lemma follows, since, by [6, (11.4)] $B_\theta^+ \simeq \frac{1}{2} b^{S^2}$ and $B_\theta^- \simeq -\frac{1}{2} b^{\wedge 2}$. □

If $b = \langle a_1, \ldots, a_m \rangle$ is a diagonalization of $b$, it is easily verified that

$$b^{S^2} \simeq m(2) \perp \left( \bigwedge_{1<i<j} \langle a_i a_j \rangle \right) \quad \text{and} \quad b^{\wedge 2} \simeq \bigwedge_{1<i<j} \langle a_i a_j \rangle$$

(cf [6, p. 135]). The formula of the preceding lemma can then be written as

$$B_\theta^+ \simeq m(2) \perp \langle 1, -\alpha, -\beta, \alpha\beta \rangle \cdot b^{\wedge 2},$$

hence, in terms of quadratic forms,

$$T_\theta^+ = m(2) + n_Q \cdot q^{\wedge 2}$$

where $n_Q$ denotes the norm form of $Q$ and $q^{\wedge 2}$ is the quadratic form defined by $q^{\wedge 2}(x) := b^{\wedge 2}(x, x)$.

Let $b_1$ and $b_2$ be two nonsingular symmetric bilinear forms on $V$, and let

$$\sigma_1 = \gamma \otimes \text{ad}_{b_1}, \quad \sigma_2 = \gamma \otimes \text{ad}_{b_2}$$

be the symplectic involutions on $A = Q \otimes \text{End}_F V$ constructed as in the preceding lemma. Observe that $T_{\sigma_1}^+ - T_{\sigma_2}^+ = n_Q \cdot (q_1^{\wedge 2} - q_2^{\wedge 2})$, hence

$$e_3(T_{\sigma_1}^+ - T_{\sigma_2}^+) = [Q] \cup \text{disc}(q_1^{\wedge 2} - q_2^{\wedge 2}).$$

Explicit calculation shows that

$$\text{disc}(q_1^{\wedge 2} - q_2^{\wedge 2}) = \det b_1^{\wedge 2} \cdot \det b_2^{\wedge 2} = (\det b_1 \cdot \det b_2)^m - 1.$$  

Adjoining an indeterminate to $F$ if necessary, we may assume that there exists an element $t \in F^\times$ not belonging to $\text{Nrd}(Q)$. By a theorem of Merkurjev, this element satisfies $[Q] \cup \langle t \rangle_2 \neq 0$. It is easy to find two bilinear forms $b_1$ and $b_2$ on $V$ such that $\det b_1 \cdot \det b_2 = t F^\times$. Since $m$ is even, it follows from (4) and (5) that the corresponding involutions $\sigma_1$ and $\sigma_2$ satisfy

$$e_3(T_{\sigma_1}^+ - T_{\sigma_2}^+) \neq 0.$$  

This completes the proof of Theorem 4.

We now turn to Proposition 6 and assume $A = M_2(F) \otimes_F A_0$. Since all hyperbolic involutions are conjugate, we may assume moreover $\theta = \gamma \otimes \theta_0$ for some orthogonal involution $\theta_0$ on $A_0$, where $\gamma$ is the unique symplectic involution on $M_2(F)$ (which is hyperbolic). As in Lemma 11, we have an orthogonal decomposition

$$\text{Sym}(A, \theta) = (F \otimes \text{Sym}(A_0, \theta_0)) \oplus (\text{Skew}(M_2(F), \gamma) \otimes \text{Skew}(A_0, \theta_0))$$

which yields

$$T_\theta^+ = \langle 2 \rangle \cdot T_{\theta_0}^+ \perp \langle 2, -2, -2 \rangle \cdot T_{\theta_0}^-.$$
Therefore, $T^+_\theta$ is Witt-equivalent to $\langle 2 \rangle \cdot (T^+_\theta - T^-_{\theta_0}) = \langle 2 \rangle \cdot T_{A_0}$. Since the adjoint involution to $T_\theta$ is $\theta \otimes \theta$, by [6, (11.1)], it is clear that $T_\theta$ is hyperbolic when $\theta$ is hyperbolic. If $\deg A \equiv 2 \mod 4$, then $\deg A_0$ is odd, hence $A_0$ is split. Therefore, $A$ is also split. The other statements in Proposition 6 follow from Corollary 5.

4. Discriminant and decomposability of involutions.

Our first goal in this section is to give a proof of Theorem 7. As observed in Section 1.3, the theorem is easy if $\ind A = 2$. Therefore, we assume $\ind A = 4$. We may then represent $A$ as

$$A = \text{End}_D V$$

where $D$ is a division algebra of degree 4 and $V$ is a 2-dimensional $D$-vector space. Let $\theta_0$ be an arbitrary symplectic involution on $D$. The involution $\sigma$ is adjoint to a Hermitian form $h$ on $V$ (with respect to $\theta_0$). Using an orthogonal basis of $V$ relative to $h$, we may identify

$$A = M_2(D) \quad \text{and} \quad \sigma = \text{Int diag}(u_1, u_2) \circ \hat{\theta}_0$$

for some $u_1, u_2 \in \text{Sym}(D, \theta_0)^\times$, where

$$\hat{\theta}_0((a_{ij})_{1 \leq i, j \leq 2}) = (\theta_0(a_{ij}))^t_{1 \leq i, j \leq 2},$$

i.e., $\hat{\theta}_0 = t \otimes \theta$ on $A = M_2(F) \otimes_F D$. Substituting $\text{Int}(u_1) \circ \theta_0$ for $\theta_0$, we may assume $u_1 = 1$. By [6, (16.16)], we may find a decomposition of $D$ into a tensor product of quaternion subalgebras stable under $\theta_0$,

$$D = Q_1 \otimes_F Q, \quad \theta_0 = \theta_1 \otimes \gamma$$

where $\theta_1$ is an orthogonal involution on $Q_1$ and $\gamma$ is the canonical involution on $Q$. Moreover, we may assume $u_2 \in Q_1$. Then

$$\sigma = \text{Int diag}(1, u_2) \circ (t \otimes \theta_1 \otimes \gamma) = \sigma_0 \otimes \gamma$$

with $\sigma_0 = \text{Int diag}(1, u_2) \circ t \otimes \theta_1$ on $M_2(F) \otimes Q_1$. Theorem 7 is thus proved. Note that the quaternion algebra $Q$ is not uniquely determined by [6, (16.16)].

Let us now prove Theorem 8, starting with the following general remark:

**Lemma 12.** For $i = 1, 2$, let $A_i$ be a central simple $F$-algebra with involutions $\sigma_i, \theta_i$. Assume:

(a) $\deg A_1 \equiv 2 \mod 4$ and $\sigma_1, \theta_1$ orthogonal, and
(b) $\deg A_2 \equiv 0 \mod 4$ and $\sigma_2, \theta_2$ symplectic.

Then $\Delta_{\theta_1 \otimes \theta_2}(\sigma_1 \otimes \sigma_2) = 0$.

**Proof.** Consider $u_1 \in \text{Sym}(A_1, \theta_1)^\times$ and $u_2 \in \text{Sym}(A_2, \theta_2)^\times$ such that

$$\sigma_1 = \text{Int}(u_1) \circ \theta_1 \quad \text{and} \quad \sigma_2 = \text{Int}(u_2) \circ \theta_2,$$
hence
\[ \sigma_1 \otimes \sigma_2 = \text{Int}(u_1 \otimes u_2) \circ (\sigma_1 \otimes \sigma_2). \]
Then \( \Delta_{\theta_1 \otimes \theta_2}(\sigma_1 \otimes \sigma_2) = (\text{Nrp}_{\theta_1 \otimes \theta_2}(u_1 \otimes u_2))_2 \cup [A_1 \otimes A_2] \), and Lemma 9(d) yields
\[ (6) \quad \text{Nrp}_{\theta_1 \otimes \theta_2}(u_1 \otimes u_2) = \text{Nrd}_{A_1}(u_1)^{\frac{1}{2}\deg A_2} \text{Nrd}_{A_2}(u_2)^{\frac{1}{2}\deg A_1}. \]
Since \( \deg A_2 \equiv 0 \mod 4 \), the first factor is a square. Moreover, since \( \sigma_2 \) and \( \theta_2 \) are of symplectic type,
\[ \text{Nrd}_{A_2}(u_2) = \text{Nrp}_{\theta_2}(u_2)^2. \]
Therefore, Equation (6) shows that \( \text{Nrp}_{\theta_1 \otimes \theta_2}(u_1 \otimes u_2) \in F^\times 2 \), so that
\[ \Delta_{\theta_1 \otimes \theta_2}(\sigma_1 \otimes \sigma_2) = 0. \]

\[ \square \]

Even in the case \( \deg A = 8 \), there may be symplectic involutions \( \theta, \sigma \) on \( A \) which do not decompose as in Lemma 12, even though \( \Delta_\theta(\sigma) = 0 \). Indeed, there are examples of algebras with involution which do not contain any invariant quaternion subalgebra on which the restriction of the involution is of orthogonal type. Suppose \( \theta \) is a symplectic involution on a central simple algebra \( A \) of degree 8, and \( A_1 \subset A \) is a quaternion subalgebra on which \( \theta \) restricts to an orthogonal involution \( \theta_1 \). The restriction of \( \theta \) to the centralizer of \( A_1 \) is then symplectic, hence [6, (16.16)] yields a decomposition
\[ (A, \theta) = (A_1, \theta_1) \otimes (A_2, \theta_2) \otimes (A_3, \gamma_3), \]
where \( A_1, A_2 \) and \( A_3 \) are quaternion algebras, \( \theta_1 \) and \( \theta_2 \) are orthogonal involutions on \( A_1 \) and \( A_2 \) respectively, and \( \gamma_3 \) is the canonical involution on \( A_3 \). This implies, in particular, that the signature of \( \theta \) with respect to every ordering of the field \( F \) is either 0 or 8. For example, if \( F = \mathbb{R} \) is the field of real numbers and \( \theta \) is the involution adjoint to the Hermitian form \( \langle 1, 1, 1, -1 \rangle \) on the usual quaternion algebra \( \mathbb{H} \), then \( \text{sgn} \theta = 4 \), and so \( A = M_4(\mathbb{H}) \) has no quaternion subalgebras on which \( \theta \) restricts to an orthogonal involution. Therefore, even though \( \Delta_\theta(\theta) = 0 \), there is no decomposition as in Lemma 12. (See Example 13 for a subtler example.)

Returning to the proof of Theorem 8, we suppose until the end of this section that \( A \) is a central simple \( F \)-algebra of degree 8, with index dividing 4. Let \( \sigma \) be a symplectic involution on \( A \), and suppose \( A_1 \simeq M_2(F) \) is an invariant subalgebra on which the restriction of \( \sigma \) is an orthogonal involution. In this situation, we have a decomposition \( A = A_1 \otimes A_1' \), where \( A_1' \) denotes the centralizer of \( A_1 \), and \( \sigma = \sigma_1 \otimes \sigma_1' \), where \( \sigma_1 \) and \( \sigma_1' \) are the restrictions of \( \sigma \) to \( A_1 \) and \( A_1' \) respectively. As \( A_1 \simeq M_2(F) \), we can find a hyperbolic orthogonal involution \( \theta_1 \) on \( A_1 \) and set
\[ \theta = \theta_1 \otimes \sigma_1'. \]
The involution $\theta$ is hyperbolic of symplectic type and, by Lemma 12, we have
\[
\Delta(\sigma) = \Delta_\theta(\sigma) = 0.
\]
Conversely, let $\sigma$ be a symplectic involution on $A$ such that $\Delta(\sigma) = 0$. To prove that $\sigma$ leaves invariant a subalgebra of $A$ isomorphic to $M_2(F)$ on which it restricts to an orthogonal involution, we consider separately various cases, depending on the index of $A$. If $A$ is split, every symplectic involution is hyperbolic and the property is a consequence of [1, Theorem 2.2]. If $\text{ind} A = 2$, we can always represent $A$ in the form $A = \text{End}_Q V$ where $Q$ is a quaternion algebra and $V$ is a 4-dimensional vector space over $Q$. The involution $\sigma$ is then adjoint to a Hermitian form $h$ on $V$ (with respect to the canonical involution $\gamma$ on $Q$). Let $e$ be an orthogonal basis for $h$. Since $h$ is determined by $\sigma$ up to a factor in $F^\times$, we may assume that the diagonalization of $h$ with respect to the basis $e$ is $\langle 1, \alpha_1, \alpha_2, \alpha_3 \rangle$ with $\alpha_1, \alpha_2, \alpha_3 \in F^\times$. Let $V_0 \subset V$ be the $F$-subspace with basis $e$. We have $V = V_0 \otimes_F Q$ and $A = (\text{End}_F V_0) \otimes_F Q$, $\sigma = \sigma_0 \otimes \gamma$ where $\sigma_0$ is the orthogonal involution on $\text{End}_F V_0$ adjoint to the bilinear symmetric form $\langle 1, \alpha_1, \alpha_2, \alpha_3 \rangle$. As in Example 2, $\Delta(\sigma) = (\alpha_1 \alpha_2 \alpha_3)^2 \cup [Q]$. Therefore, the condition $\Delta(\sigma) = 0$ implies, by a theorem of Merkurjev, that $\alpha_1 \alpha_2 \alpha_3 \in \text{Nrd}_Q(Q^\times)$. Changing basis if necessary, we may assume that $\alpha_3 = \alpha_1 \alpha_2$. Then
\[
\langle 1, \alpha_1, \alpha_2, \alpha_3 \rangle = \langle 1, \alpha_1 \rangle \otimes \langle 1, \alpha_2 \rangle.
\]
This implies $\sigma_0 = \sigma_1 \otimes \sigma_2$ on $\text{End}_F V_0 \simeq M_2(F) \otimes M_2(F)$, where $\sigma_1$ and $\sigma_2$ are the involutions adjoint to the bilinear forms $\langle 1, \alpha_1 \rangle$ and $\langle 1, \alpha_2 \rangle$, respectively. This proves the theorem in this case.

Finally, suppose $\text{ind} A = 4$. As in the proof of Theorem 7 given at the beginning of this section, we may then represent $A$ as $A = \text{End}_D V$ where $D$ is a division algebra of degree 4 and $V$ is a 2-dimensional $D$-vector space. For the rest of the proof, we use the same notation as in the proof of Theorem 7. We may thus assume $A = M_2(D) = M_2(F) \otimes_F D$ and $\sigma = \text{Int diag}(1, u_2) \circ \theta_0$ for some symplectic involution $\theta_0$ on $D$ and $\hat{\theta}_0 = t \otimes \theta_0$. The involution $\theta = \text{Int diag}(1, -1) \circ \theta_0$ is hyperbolic, and $\sigma = \text{Int diag}(1, -u_2) \circ \theta$. By Lemma 9(e), we have
\[
\text{Nrp}_\theta(\text{diag}(1, -u_2)) = \text{Nrp}_{\theta_0}(-u_2) = \text{Nrp}_{\theta_0}(u_2),
\]

hence
\[
\Delta(\sigma) = \Delta_\theta(\sigma) = (\text{Nrp}_{\theta_0}(u_2))^2 \cup [D].
\]
Therefore, the condition $\Delta(\sigma) = 0$ implies by [6, (16.19)] that the involution $\text{Int}(u_2) \circ \theta_0$ on $D$ is conjugate to $\theta_0$. We may then find $v \in D^\times$ such that
\[
\text{Int}(u_2) \circ \theta_0 = \text{Int}(v) \circ \theta_0 \circ \text{Int}(v)^{-1} = \text{Int}(v \theta_0(v)) \circ \theta_0,
\]
hence $u_2 = v \theta_0(v) \lambda$ for some $\lambda \in F^\times$. The involution $\sigma$ is conjugate to
\[
\text{Int}(1, v)^{-1} \circ \sigma \circ \text{Int}(1, v) = \text{Int}(1, \lambda) \circ \theta_0,
\]
which restricts to $\text{Int}(1, \lambda) \circ t$ on $M_2(F) \subset M_2(D)$. Therefore, $\sigma$ leaves the subalgebra $A_1 = \text{diag}(1, v) M_2(F) \text{diag}(1, v^{-1})$ invariant and restricts to an orthogonal involution $\sigma_1$ on that subalgebra. We thus have a decomposition
\[
(A, \sigma) = (A_1, \sigma_1) \otimes (A_1', \sigma_1')
\]
where $A_1'$ is the centralizer of $A_1$ and $\sigma_1'$ is the restriction of $\sigma$ to $A_1'$. The involution $\sigma_1'$ is symplectic, hence, by [6, (16.16)], there is a decomposition
\[
(A_1', \sigma_1') = (A_2, \sigma_2) \otimes (A_3, \gamma_3).
\]
The proof of Theorem 8 is thus complete.

**Example 13.** The following is another example where the discriminant vanishes even though there is no decomposition as in Lemma 12. Consider three quadratic extensions $K_1$, $K_2$, $K_3$ of a field $k$,
\[
K_i = k(\sqrt{a_i}) \quad \text{for some } a_i \in k,
\]
such that $K_1 \otimes_k K_2 \otimes_k K_3$ is a field, and let $F = k(x_1, x_2, x_3)$ be the field of rational fractions in three indeterminates over $k$. For $i = 1, 2, 3$, consider $K_i$ as a subfield of the quaternion algebra $A_i = (a_i, x_i)_F$. On the tensor product
\[
A = A_1 \otimes_F A_2 \otimes_F A_3,
\]
consider the symplectic involution
\[
\theta = \theta_1 \otimes \theta_2 \otimes \gamma_3,
\]
where $\gamma_3$ is the conjugation involution on $A_3$ and $\theta_1$ (resp. $\theta_2$) is an orthogonal involution on $A_1$ (resp. $A_2$) which is the identity on $K_1$ (resp. $K_2$).

Let $\lambda \in N_{K_1/k}(K_1^\times) \cap N_{K_2/k}(K_2^\times)$. By a well-known property of biquadratic extensions (see for instance [4, 2.13]), we may find $u \in K_1 \otimes_k K_2$ and $v \in k^\times$ such that
\[
\lambda = v^2 N_{K_1 \otimes_k K_2/k}(u).
\]
Viewing $u = u \otimes 1$ as an element of $A$, we let
\[
\sigma = \text{Int}(u) \circ \theta.
\]
By Lemma 9(d), we have $\text{Nr} \theta(u) = \text{Nrd}_{A_1 \otimes_F A_2}(u) = N_{K_1 \otimes_k K_2/k}(u)$, hence
\[
\Delta_{\theta}(\sigma) = (\lambda) \cup [A].
\]
Since $\lambda$ is a norm from $K_1$ and $K_2$, hence a reduced norm from $A_1$ and $A_2$, it follows that
\[ \Delta^\theta(\sigma) = \langle \lambda \rangle \cup [A_3]. \]
Hence, $\Delta^\theta(\sigma) = 0$ if and only if $\lambda \in N_{K_1/k}(K_1^\times) \cap N_{K_2/k}(K_2^\times) \cap N_{K_3/k}(K_3^\times)$.

Now, suppose $\sigma = \text{Int}(w) \circ \sigma' \circ \text{Int}(w)^{-1}$ for some involution $\sigma'$ leaving $A_1$ invariant. Then $\sigma' = \text{Int}(w^{-1}u\theta(w)^{-1}) \circ \theta$, and the proof of Lemma 12 shows that $N_{r^p}(w^{-1}u\theta(w)^{-1}) \in F^\times 2$. Since
\[ N_{r^p}(w^{-1}u\theta(w)^{-1}) = N_{r^p}(w) N_{r^d}(w)^{-1} = \lambda v^{-2} N_{r^d}(w)^{-1}, \]
it follows that $\lambda \in F^\times 2 \cdot N_{r^d}(A^\times)$.

By Proposition 9 of [3], we then have
\[ \lambda \in k^\times 2 \cdot N_{M/k}(M^\times) \quad \text{with} \quad M = K_1 \otimes_k K_2 \otimes_k K_3. \]
Therefore, examples of triquadratic extensions $M = K_1 \otimes_k K_2 \otimes_k K_3/k$ such that
\[ N_{K_1/k}(K_1^\times) \cap N_{K_2/k}(K_2^\times) \cap N_{K_3/k}(K_3^\times) \neq k^\times 2 \cdot N_{M/k}(M^\times) \]
yield examples of involutions $\sigma$ for which $\Delta^\theta(\sigma) = 0$ even though $\sigma$ is not conjugate to an involution leaving $A_1$ invariant. Triquadratic extensions of this type were constructed in [9] (see also [5, Proposition 3]).

References


Received March 15, 2002 and revised May 20, 2002. The authors thank the referee for his/her careful reading and constructive remarks. They gratefully acknowledge support from the TMR network “K-theory and linear algebraic groups” (ERB FMRX CT97-0107). The third author is also grateful to the National Fund for Scientific Research (Belgium) for partial support.

Département de mathématiques
École Polytechnique Fédérale de Lausanne
CH-1015 Lausanne
Switzerland
E-mail address: gregory.berhuy@epfl.ch

Département de mathématiques
École Polytechnique Fédérale de Lausanne
CH-1015 Lausanne
Switzerland
E-mail address: marina.monsurro@epfl.ch

Institut de Mathématique Pure et Appliquée
Université catholique de Louvain
B-1348 Louvain-la-Neuve
Belgium
E-mail address: tignol@math.ucl.ac.be
FUSION AND FISSION IN GRAPH COMPLEXES

JAMES CONANT

We analyze a functor from cyclic operads to chain complexes first considered by Getzler and Kapranov and also by Markl. This functor is a generalization of the graph homology considered by Kontsevich, which was defined for the three operads Comm, Assoc, and Lie. More specifically we show that these chain complexes have a rich algebraic structure in the form of families of operations defined by fusion and fission. These operations fit together to form uncountably many Lie$_\infty$ and co-Lie$_\infty$ structures. In particular, the chain complexes have a bracket and cobracket which are compatible in the Lie bialgebra sense on a certain natural subcomplex.

1. Introduction.

More than a decade ago Maxim Kontsevich [K] considered graph homology as a tool for studying and computing the homology of many seemingly disparate objects. One version of the graph complex computes, via work of R.C. Penner [P], the homology of the moduli space (or equivalently mapping class group) of surfaces. Another version computes, via work of M. Culler and K. Vogtmann [CuV], the homology of the group of outer automorphisms of the free group. There is also a version which is related to finite type invariants of three-manifolds. On the other hand, these three graph complexes compute the homology of three infinite dimensional Lie algebras, leading to quite unexpected isomorphisms. Kontsevich’s graph complexes were generalized to the case of modular operads by Getzler and Kapranov[GK2], and were considered for the special case of cyclic operads by Martin Markl [M].

In [CV] Karen Vogtmann and I showed that the commutative graph complex carries the structure of both a Lie algebra and a Lie coalgebra. These are compatible as a bialgebra on a certain natural subcomplex. In this paper I will generalize these two operations to the case of any cyclic operad, and show that they are each first in a series of higher order operations which fit together nicely and vanish on homology.

Let the graph complex corresponding to a cyclic operad $\mathcal{O}$ be denoted by $G^\mathcal{O}$. I will define a sequence of “higher order brackets”

$$\phi_n: S^nG^\mathcal{O} \to G.$$
The map $\phi_n$ is defined by fusing together $n$ graphs along a $2n$-gon in all possible ways (Figure 4). Extending each $\phi_n$ as a coderivation to $SG^O$, these maps are all compatible with each other in a very strong sense (Theorem 1). For any subset $I \subset \mathbb{N}$, define $\phi_I = \sum_{i \in I} \phi_i$. Theorem 1 implies that $\phi_2^2 = 0$. This is precisely the definition of a Lie$_\infty$ (strong homotopy Lie) structure. In this way we get uncountably many Lie$_\infty$ structures.

Let $PG^O$ denote the subcomplex of the graph complex spanned by connected graphs. I will define a sequence of “higher order cobrackets”

$$\theta_n : PG^O \to S^n PG^O.$$ 

The map $\theta_n$ is defined by fissioning a graph into $n$ graphs along a $2n$-gon (Figure 5). The $\theta_n$ maps, extended to $SG^O$ as derivations, are compatible in a strong sense also (Theorem 2). For any $I \subset \mathbb{N}$, $\theta_I$ is defined as above, and Theorem 2 implies each $\phi_I$ is a co-Lie$_\infty$ structure.

Trouble arises, as was foreshadowed in [CV] in the compatibility between brackets and cobrackets. In [CV] we were able to avoid difficulty by restricting to connected graphs without separating edges, and indeed in this context $\theta_2, \phi_2$ are compatible in a Lie bialgebra sense (Theorem 3). But there appears to be no similar way out for the higher order operations. The higher order brackets and cobrackets simply fit together in a more complicated way than one would guess, even on graphs without separating edges.

All of the operations are highly nontrivial on chains, and are compatible with the boundary operator. Indeed they vanish canonically on the level of homology. Thus these operations can be thought of as “generalized Schouten brackets,” since in the case of Lie algebras, the Schouten bracket is an operation on the Chevalley-Eilenberg complex which vanishes canonically upon application of the homology functor.

Moira Chas and Dennis Sullivan [CS] define similar structures on string homology, the homology of a free loop space. They define an uncountable family of Lie$_\infty$ structures, indexed by sets of positive integers, on string homology which obey the same compatibility relations as the ones found here (Theorem 1). They also find a Lie bialgebra structure [C] and [CS2]. Drawing the analogy further, one is led to speculate that string homology has an uncountable infinity of co-Lie$_\infty$ structures. It would be interesting to know whether such co structures, if they exist, are compatible in a nice way with the Lie structures, or if they mirror more complicated graph interactions.

2. Cyclic operads and graphs.

We begin by briefly reviewing the salient features of a cyclic operad, and proceed to give Markl’s construction of graph complexes. A good introduction to these objects can be found in the recent book by Markl, Shnider and Stasheff [MSS].

Kontsevich’s three graph complexes are associated to the commutative, associative and Lie operads. Each of these operads $O = \oplus O(n)$ has a
description as a vector space spanned by different flavors of rooted trees with labelled leaves.

Figure 1. Elements of Comm[3], Assoc[3], and Lie[3] respectively.

- The $n$th degree part of the commutative operad $\text{Comm}(n)$ has a basis consisting of rooted trees which have one internal vertex and $n$ labelled leaves. Hence $\text{Comm}[n]$ is 1 dimensional! The composition law

$$\text{Comm}[m] \otimes \text{Comm}[n_1] \otimes \cdots \otimes \text{Comm}[n_m] \to \text{Comm}[n_1 + \cdots + n_m]$$

is defined on $c \otimes c_1 \otimes \cdots \otimes c_m$ by grafting the root of each $c_i$ onto the leaf of $c$ labelled by $i$ for each $i$, and suitably relabelling the leaves. The composition is completed by contracting all edges not adjacent to a root or a leaf.

- The $n$th degree part of the associative operad $\text{Assoc}(n)$ has a basis consisting of rooted trees with one internal vertex which have a specified cyclic ordering of the edges incident to the vertex, and which have $n$ labelled leaves. Composition is again by grafting and contracting created edges, with the proviso that the cyclic ordering is respected.

- The $n$th degree part of the Lie operad $\text{Lie}(n)$ is actually easiest to describe as a quotient space $\text{Lie}(n)/\text{AS} + \text{IHX}$. $\text{Lie}(n)$ has a basis given by rooted trivalent trees with $n$ labelled leaves, where each vertex has a specified cyclic order of adjacent edges. The AS subspace is spanned by sums $T_1 + T_2$, where $T_{1,2}$ are identical except for a cyclic ordering on some vertex. Modding out by AS says that Lie algebras are anti-symmetric. The IHX subspace is spanned by sums $T_1 - T_2 + T_3$, where $T_{1,2,3}$ are identical trees except at one spot where they are as in Figure 2. On the level of Lie algebras this is the Jacobi relation. Composition is via grafting, but without the contraction step.

Notice that in each of these cases the action of the symmetric group $S_n$ which permutes the labels of the leaves can be extended to an action of $S_{n+1}$. This is by thinking of the root as another labelled leaf, say labelled by 0. (One must check in the Lie case that the IHX subspace is preserved by this action.) Operads where this extension is possible are called cyclic [GK], provided that the extension satisfies appropriate axioms. Other examples of cyclic operads are the endomorphism operad and the Poisson operad.

As a general philosophy, one can think of cyclic operads as consisting of unrooted trees, with composition given by some version of grafting. The idea is to plug these in to the nodes of a graph to obtain different types of
Figure 2. The IHX relation. Each term represents a piece of a graph which is identical outside of the pictured spot.

decorations on a graph. Plugging in a basis element from Comm($n$) at each vertex of valence $n$, one simply gets an undecorated graph. Plugging in an element of Assoc($n$) one gets a cyclic order at the vertex. This is often called a ribbon graph. Plugging in an element of Lie($n$) gives a relatively strange object. By definition it is obtained from some unrooted ribbon trivalent trees by joining the leaves together with edges. See Figure 3. Thus one may think of it as a trivalent graph with a special distinguished subset where IHX and AS relations may take place. It is reminiscent of the diagram algebras that appear in the theory of Vassiliev invariants of low dimensional objects, which consist of (uni)trivalent graphs, but where the AS and IHX relations are not restricted to a distinguished subset.

In general, let $H(v)$ be the set of half-edges incident to a vertex. Let $L$ be a labelling of the elements of $H(v)$ by $0, \ldots, n$, where $n+1$ is the valence of $v$. Now define

$$O((H(v))) = (\oplus_L O(n))_{S_{n+1}}$$

which is the set of coinvariants under the action of $S_{n+1}$, which acts as follows. If $o \in O(n)$, let $o_L$ denote putting $o$ in the $L$th summand of the direct sum. If $\sigma \in S_{n+1}$ then $\sigma \cdot o_L = (\sigma \cdot o)_{\sigma L}$. When $O$ is an operad of trees, $O((H(v)))$ is isomorphic to the space of identifications of the leaves and root of elements of $O[n]$ with the half-edges incident to $v$.

Now we define an $O$-labelling of a graph to be a choice of element $o_v \in O((H(v)))$ for each vertex, $v$, of the graph. Graphically, we put a circle at each vertex to represent the operad element.

In addition we would like a notion of “orientation” of a graph, which will make it possible to define a boundary operator. This is analogous to the need for an orientation of the simplices of a simplicial complex in order to do the same. There are many equivalent notions, perhaps the most intuitive is the following.

**Definition.** An orientation of a graph is an ordering of the vertices and a choice of direction for each edge, modulo the even action of $S_V \times \mathbb{Z}_2$. Here
$V$ and $E$ are the number or vertices and edges of the graph, respectively. An element of $S_V \times \mathbb{Z}_2^E$ is called even if it is a product of an even number of elements each of which is either a transposition in $S_v$ or an element of the form $(0, \ldots, 1, \ldots, 0) \in \mathbb{Z}_2^E$.

Notice that any graph has exactly two orientations. Let $-$ indicate the map switching orientations.

**Remark.** Lie graphs actually have a much simpler description, because the orientations of the graph and vertices cancel out to a large degree. Namely, one can think of a Lie graph as a trivalent graph with a distinguished subforest, whose edges are ordered modulo even permutations. The IHX relation in the Lie operad becomes the condition that the three terms in an IHX relation of the trivalent graph sum to zero provided the edge involved is in the forest. This will be explained carefully in [CV2].

**2.1. Chain complexes.** Now for any cyclic operad $\mathcal{O}$ we are ready to define $\mathcal{O}$-graph complexes.

Define $G^\mathcal{O}_v$ to be the span of $\mathcal{O}$-labelled oriented graphs with vertices all of valence $\geq 3$, modulo the relation $(G, \text{or}) = -(G, -\text{or})$ and also modulo multilinearity of the $\mathcal{O}$-labels. More precisely, we set

$$G^\mathcal{O}_v = \bigoplus_{(G, \text{or}) \in V(G)} \mathcal{O}(H(v)) \bigg/ \{(G, \text{or}) = -(G, -\text{or})\}$$

where the direct sum is over oriented graphs with vertices of valence $\geq 3$, and where $V(G)$ is the set of vertices of $G$. Define $G^\mathcal{O}_v$ to be the part of $G^\mathcal{O}$ with $v$ vertices.

For each edge, $e$, in a graph $(G, \text{or})$ we define contraction along that edge $(G, \text{or})_e$ to be the graph where the two operad elements at each endpoint of $e$ are composed along $e$. The induced orientation can be fixed by assuming that the endpoints of $e$ are labelled 1 and 2 and the edge direction is from 1 to 2. The new vertex, which results from composing the two operad elements, is labelled 1, and all other indices are reduced by 1. If $e$ is a loop, then define $(G, \text{or})_e = 0$. In the commutative case, $(G, \text{or})_e$ is defined by simply contracting the edge of the (undecorated) graph. In the associative case the cyclic orders at both endpoints of an edge are joined together to give a cyclic order at the vertex resulting from the edge collapse. For an example in the Lie case, see Figure 3.

Define $\partial_G : G^\mathcal{O}_v \to G^\mathcal{O}_{v-1}$ by $\partial_G(G, \text{or}) = \sum_{e \in E(G)} (G, \text{or})_e$, where $E(G)$ is the set of edges of $G$.

**Remark.** In the simpler version of the Lie case, the boundary operator adds an edge to the forest in all possible ways, with the edge’s number coming directly after the edge numbers in the original forest.
$G^O$ is a graded commutative algebra under disjoint union. It is also a graded commutative coalgebra, defining the coproduct such that connected graphs are primitive, and extending multiplicatively. Thus we may write $PG^O$ for the subspace generated by connected graphs. Let $P^{(n)}G^O$ be the subspace generated by connected graphs with $b_1 = n$.

Even though this paper is concerned with chain complexes and not their homology per se, it is still useful to record the following facts.

Let $\text{Out}(F_n)$ denote the group of outer automorphisms of the free group $F_n$, and let $\mathcal{M}'_{g,m}$ denote the moduli space of a surface of genus $g$ with $m$ unlabelled punctures.

Then

$$H_k(P^{(n)}G^{\text{Assoc}}) = \bigoplus_{m \geq 1, 2g + m - 1 = n} H^{4g - 4 + 2m - k}(\mathcal{M}'_{g,m}; \mathbb{Q})$$

$$H_k(P^{(n)}G^{\text{Lie}}) = H^{2n - 2 - k}(\text{Out}(F_n); \mathbb{Q}).$$

In addition, part of commutative graph cohomology plays a role in the theory of finite type invariants of homology 3-spheres. More precisely, we have that

$$\bigoplus_{n \geq 2} H^{2n - 2}(P^{(n)}G^{\text{Comm}}) \cong PA(\emptyset)$$

where $PA(\emptyset)$ is the diagram algebra where the logarithm of the Aarhus version of the LMO invariant takes values [B-NGRT].

The first two statements above are due to Kontsevich, being implicit in his paper [K]. A more detailed explanation of these two facts and their

\begin{figure}
\centering
\includegraphics[width=\textwidth]{lie_graphs.png}
\caption{The Lie graphs $G$ and $G_e$.}
\end{figure}
FUSION AND FISSION IN GRAPH COMPLEXES

proofs will appear in [CV2]. The last statement, the relation to finite type
invariants, is essentially content-free, being a trivial isomorphism, at least
modulo equivalences of various notions of orientation.

2.2. Cohomology. In at least two interesting cases, it is possible to define
graph cohomology. The coboundary operator $\delta$ is the sum of inserting an
edge in all possible ways. In the commutative and associative cases this
makes perfect sense. Unfortunately, in the Lie case an insertion, which is
essentially the deletion of an edge from the forest, does not preserve the
IHX subspace and is not well-defined. In the cases where $\delta$ can be defined
the boundary and coboundary are adjoint with respect to the inner product
$\langle G, H \rangle = |\text{Aut}(G)| \delta_{GH}$. This can be seen by applying the argument of [CV],
Proposition 12 mutatis mutandis.

3. Fusion.

We start with an oriented labelled $2n$-gon. Label every other edge on its
perimeter consecutively by the numbers $1 \ldots n$, consistent with the orienta-
tion. Now fix $n$ directed edges $e_1, \ldots, e_n$ of a graph $G$. Define $G \langle e_1, \ldots, e_n \rangle$
to be the graph formed in the following way. First, for each $i$, glue the edge
marked $i$ of the $2n$-gon to the edge $e_i$ of the graph. Second, delete these
edges along which the $2n$-gon was attached leaving $n$ new edges. This is
illustrated in Figure 4. The graph $G \langle e_1, \ldots, e_n \rangle$ has an induced orientation
which can be easily described. Fix a labelling of the graph such that the
directions of the edges $e_1, \ldots, e_n$ are both consistent with the graph’s orien-
tation and with the directions which correspond to the gluing. The $n$ new
edges have orientations induced by the $n$-gon. Switch all of these, as is usual
with a cobordism.

![Figure 4. One term in $\phi_3(G_1, G_2, G_3)$.](image)

Now, for any $n \in \mathbb{N}$ we define an operation

$$\phi_n : S^n G^\mathcal{O} \to G^\mathcal{O}$$

by $\phi_n(G_1 \odot \cdots \odot G_n) = \sum (G_1 \cdot G_2 \cdots \cdot G_n) \langle e_1, \ldots, e_n \rangle_e$, where:
• The sum is over all \( n \)-tuples of directed edges \((e_1, e_2, \ldots, e_n)\) all of which Lie in separate \( G_i \).
• The notation “\( \odot \)” denotes “graded symmetric tensor product.”
• The edge \( e \) which is contracted is the edge coming from the boundary of the \( 2n \)-gon between “1” and “2.”

Thus \( \phi_n \) is a type of fusion operation which takes \( n \) graphs and fuses them together along a \( 2n \)-gon.

Extend \( \phi_n \) to \( S\mathcal{G}^O \) as a coderivation. That is
\[
\phi_n(G_1 \odot \cdots \odot G_k) = \sum_{I \cup J} \epsilon(I, J) \phi_n(G_I) \odot G_J,
\]
where \( I, J \) is an ordered partition of \( 1, \ldots, k \), with \(|I| = n\), and \( \epsilon(I, J) \) is the sign defined by the equation \( G_1 \odot \cdots \odot G_k = \epsilon(I, J) G_I \odot G_J \). Notice that \( \phi_1 \) by definition glues on a bigon to an edge, which doesn’t change the edge, and then contracts it. That is, \( \phi_1 = \partial G \). Notice that it doesn’t matter whether we extend \( \partial G \) to \( S\mathcal{G}^O \) as a derivation or a coderivation, since they are equivalent in this case!

**Theorem 1.** The following equations hold:

a) \( \forall i \phi_i^2 = 0 \),

b) \( \forall i \neq j \phi_i \phi_j + \phi_j \phi_i = 0 \).

**Corollary 1.** Let \( I \) be a subset of \( \mathbb{N} \), finite or infinite. Let \( \phi_I = \sum_{i \in I} \phi_i \). Then \( \phi_I^2 = 0 \).

**Proof of Theorem 1.** First we show \( \phi_n^2|_{S\mathcal{G}^O} = 0 \). We only need consider the case when \( k = 2n - 1 \), which implies the higher degree cases.
\[
\phi_n^2(G_1 \odot \cdots \odot G_{2n-1}) = \sum_{I \cup J = [2n-1]} \phi_n(\phi_n(G_I) \odot G_J).
\]
Thus we are attaching a disk to \( G_i \) where \( i \in I \) along its \( n \) subarcs. We then attach a disk to the result together with the other \( n - 1 \) graphs. If the second disk attaches to an edge not involved in the first disk, then this gives the same unoriented result as attaching the disks in the other order. Keeping track of the orientation, we see that the two orders of attaching the disk cancel. The other possibility is that the second disk attaches to the first. This can be thought of as attaching a \( 4n - 2 \)-gon to the \( 2n - 1 \) graphs, with a separating arc along the \( 4n - 2 \)-gon, and two ordered edges marked for collapse. We can simplify the combinatorics somewhat by shrinking the complement of the \( 2n - 1 \) attaching regions for the disk, to get a \( 2n - 1 \)-gon with an arc joining two vertices and two vertices marked for collapse. The sorts of configurations that arise are exactly recorded by the concept of *admissible* defined below. The lemma now follows from the following analysis of \( 2n - 1 \)-gons.
Define Conf\((2n-1,n)\) be the set of admissible configurations of a 2\(n-1\)-gon. An admissible configuration consists of an embedded arc on the 2\(n-1\)-gon between two of the vertices, thereby partitioning the 2\(n-1\) vertices into two sets of \(n-1\) and \(n-2\) respectively, on each side of the arc. There are also two vertices labelled by 1 and 2, the 1 must be in the set of \(n-1\) and the 2 must be among the \(n-2\) or it could be one of the endpoints of the arc. We claim that the subset of Conf\((2n-1,n)\) where two specific vertices are marked 1 and 2 is bijective with the subset where these vertices are marked 2 and 1, respectively. This follows from the fact that there is a unique automorphism exchanging any two vertices of a 2\(n-1\)-gon. This induces a bijection between the two types of configurations. Keeping track of orientations, we see that the terms of corresponding to elements of Conf\((2n-1,n)\) cancel in pairs.

The fact that \(\phi_i, \phi_j\) anti-commute follows from the following similar facts about configurations of \(i+j-1\)-gons, Conf\((i+j-1,i,j)\). The arc in this case will partition the vertices into a set of \(i-1\), and a set of \(j-2\), where the 1 vertex must lie in the \(i-1\) and the 2 elsewhere. We claim there is a bijective correspondence between subsets of Conf\((i+j-1,i,j)\) where two fixed vertices are labelled 1 and 2 and the subsets of Conf\((i+j-1,j,i)\) where these vertices are labelled 2 and 1. To see this, fix an automorphism of the \(i+j-1\)-gon, exchanging the two given vertices. This will carry one set of configurations onto the other. \(\square\)

**Proposition 1.** \(\phi_n\) is canonically zero at the level of homology.

**Proof.** The fact that \(\phi_n\) is even compatible with homology is the fact

\[\partial_G \circ \phi_n + \phi_n \circ \partial_G = 0\]

where \(\partial_G\) is extended to \(SG^O\) as a derivation. This follows since \(\partial_G = \phi_1\).

It remains to show that it vanishes canonically. Consider the map

\[\mu_n : S^n G^O \to G^O\]

which is defined by gluing in a 2\(n\)-gon in all possible ways, but without contracting an edge. Then a straightforward argument shows that \(\phi_n = \partial_G \mu_n - \mu_n \partial_G\). Thus if the input to \(\phi_n\) consists of \(n\) cycles, the \(\mu_n \partial_G\) term in this equation vanishes, and what is left expresses \(\phi_n\) as a boundary. \(\square\)

4. Fission.

In this section, for simplicity, we restrict ourselves to connected graphs, although much of it can be generalized to the nonconnected case. In particular, when edge insertions make sense, one can dualize and prove Theorem 2 analogously to Proposition 11 of [CV].

Note that \(G^O \cong S(PG^O)\). Denote this isomorphism by \(S\). Let

\[\pi_i : S(PG^O) \to S^i(PG^O)\]
be the natural projection. Define the map
\[ \partial_i : \mathcal{G}_v^O \to \mathcal{G}_{v-1}^O \]
by summing over all ways of attaching a $2i$-gon to the edges of an $O$-graph, and then contracting the edge between 1 and 2. The behavior of this operator (which does not have square zero) is complicated, but it becomes better behaved if we look at the part which disconnects the graph the most.

**Definition.** The map
\[ \theta_i : P\mathcal{G}^O \to S^i(P\mathcal{G}^O) \]
is defined as the composition $\frac{1}{2} \pi_i \circ S \circ \partial_i$.

The operator $\theta_i$ can be thought of as a type of fission, where a graph splits up into $i$ particles. See Figure 5.

**Figure 5.** A term in $\theta_3(G)$. The middle picture represents a term in $\partial_3(G)$, and the final picture is a result of applying $S$.

Extend $\theta_i$ to
\[ \theta_i : S(P\mathcal{G}^O) \to S(P\mathcal{G}^O) \]
as a derivation. Notice that $\theta_1 = \partial G = \phi_1$.

**Theorem 2.** The following identities hold:

a) \( \forall i \neq j \theta_i \theta_j + \theta_j \theta_i = 0 \),
b) \( \forall i \theta_i^2 = 0 \).

**Proof.** We prove a). Statement b) is similar. We show that
\[ \theta_i \theta_j + \theta_j \theta_i : \mathcal{G}^O \to S^{i+j}(\mathcal{G}^O) \]
is zero, which is enough. If the $i$-gon and $j$-gon attach to two different sets of edges, they can be applied in either order to get the same (unoriented) result. Keeping track of orientation, one sees that they anticommute. Attaching one disk, and then the other to an edge of the original disk is the same as adding a bigger disk with an ordered pair of two sides marked for collapse. We may now apply our analysis from the proof of Theorem 1 to show that the terms cancel in pairs.
Corollary 2. Let $I$ be a subset of $\mathbb{N}$, finite or infinite. Let $\theta_I = \sum_{i \in I} \theta_i$. Then $\theta_I^2 = 0$.

Proposition 2. $\theta_i$ is canonically zero at the level of homology.

Proof. That $\theta_i$ is compatible with homology follows since $\theta_1 = \partial_G$.

A similar argument to Proposition 1 shows that $\theta_i$ vanishes canonically on homology. □

The operator $\theta_i$ can be defined for disconnected graphs as well, as we alluded to earlier. Suppose we start with a graph with $k$ connected components. A $2i$-gon attaches to one of these and it fissions into $i$ components.

In order to get a well-defined map, the remaining $k-1$ components must be distributed with the $i$ fission components in all possible ways, which leads to more complicated formulas.

5. Compatibility.

It is unclear if there is a theory of Lie$_\infty$ bialgebras; a search of MathSciNet yields no hits. Under some obvious generalizations of the definition of Lie bialgebra to the case of higher order operations on the symmetric algebra, the higher degree fusion operations are not compatible with the higher degree fission operations. Interestingly, degree 2 fission is compatible with degree 2 fusion on the subcomplex of connected graphs with no separating edges. As was noted in [CV] this is not the case on the full complex $G^O$.

Definition. Let $P^{\text{irred}}G^O$ be the subcomplex of $G^O$ spanned by connected (primitive) graphs with no separating edges (irreducible).

Theorem 3. On $P^{\text{irred}}G^O$ the following equation holds:

$$
\theta_2 \phi_2(X \circ Y) + \phi_2(\theta_2(X) \circ Y) + (-1)^x \phi_2(X \circ \theta_2(Y)) = 0.
$$

Proof. The bracket $\phi_2$ and cobracket $\theta_2$ coincide with the operations $[\cdot, \cdot]$ and $\theta$ defined in [CV] for the commutative operad. In that paper, we defined everything in terms of contracting pairs of half-edges, but the operations are easily seen to match. (In fact, we mentioned a “dotted line notation” in that paper which is very close to the definition of $\phi_2$ considered here.) Now use the argument from [CV] Theorem 1, which holds even if the vertices are labelled by the operad $O$. □

Acknowledgements. It is a pleasure to thank Karen Vogtmann for many discussions. I also wish to thank Swapneel Mahajan for his perceptive input. Credit also goes to the anonymous referee who noticed an error in the original manuscript and suggested many expositional improvements.
References


Received April 4, 2002 and revised July 25, 2002. This work was partially supported by NSF VIGRE grant DMS-9983660.

Department of Mathematics
Cornell University
Ithaca, NY 14853-4201
E-mail address: jconant@math.cornell.edu
SOME PLANAR ALGEBRAS RELATED TO GRAPHS

Brian Curtin

Let $X$ denote a finite nonempty set, and let $W$ denote a matrix whose rows and columns are indexed by $X$ and whose entries belong to some field $\mathbb{K}$. We study three planar algebras related to $W$. Briefly, a planar algebra is a graded vector space $\mathcal{V} = \bigcup_{n \in \mathbb{Z}^+} \mathcal{V}_n$ which is closed under “planar” operators.

The first planar algebra which we study, $\mathcal{F}_W = \bigcup_{n \in \mathbb{Z}^+} \mathcal{F}_n$, is defined by the group theoretic properties of $W$. For $n \in \mathbb{Z}^+$, $\mathcal{F}_n$ is the vector space of functions from $X^n$ to $\mathbb{K}$ which are constant on the $\text{Aut}(W)$-orbits of $X^n$, and $\mathcal{F}_n^+$ and $\mathcal{F}_n^-$ are identified with $\mathbb{K}$. The second planar algebra, $\mathcal{P}_W = \bigcup_{n \in \mathbb{Z}^+} \mathcal{P}_n$, is the planar algebra generated by $W$. We define it combinatorially: $\mathcal{P}_n$ is spanned by functions from $X^n$ to $\mathbb{K}$ defined via statistical mechanical sums on certain planar open graphs. The third planar algebra, $\mathcal{O}_W = \bigcup_{n \in \mathbb{Z}^+} \mathcal{O}_n$, differs from $\mathcal{P}_W$ only in that the open graphs defining the functions need not be planar.

It turns out that $\mathcal{P}_W \subseteq \mathcal{O}_W \subseteq \mathcal{F}_W$. We show that $\mathcal{P}_W = \mathcal{O}_W$ if and only if $\mathcal{P}_4$ contains a single special function known as the “transposition”. We show that $\mathcal{O}_W = \mathcal{F}_W$ whenever $|X|!$ is not divisible by the characteristic of $\mathbb{K}$.

1. Introduction.

Planar algebras were introduced by V.F.R. Jones [15] to study the structure of subfactors. A planar algebra is a graded vector space $\mathcal{V} = \bigcup_{n \in \mathbb{Z}^+} \mathcal{V}_n$ over some field $\mathbb{K}$ which is closed under certain operators. True to its operator algebra origins, an emphasis is placed upon the interactions of the operators. These operators are defined diagrammatically by objects known as planar tangles. We recall relevant definitions in Section 2. The study of a planar algebra via the dependencies of these operators has a knot theoretic flavor, very much like Conway’s tangles and skein relations [4]. This is no coincidence, as planar algebras were influenced by the deep relationship between subfactors and knots [12] and [14].

When $\dim \mathcal{V}_n$ is finite for all $n$, it is natural to ask for the exact value. We shall consider this problem for some combinatorial planar algebras. In our examples, $\mathcal{V}_n$ is a vector space of functions from $X^n$ to $\mathbb{K}$ for some fixed,
finite, nonempty set $X$. The action of the planar tangles is defined via the statistical mechanical construction known as a partition function.

In Section 3 we introduce three planar algebras related to a matrix $W$ whose rows and columns are indexed by $X$ and entries are in $\mathbb{K}$. In the first planar algebra defined from $W$, $\mathcal{F}^W = \cup \mathcal{F}^W_n$, the vector space $\mathcal{F}^W_n$ consists of those functions from $X^n$ to $\mathbb{K}$ which are constant on the orbits of $X^n$ under the action of $\text{Aut}(W)$. The second is a singly generated planar algebra, $\mathcal{P}^W = \cup \mathcal{P}^W_n$, whose vector space $\mathcal{P}^W_n$ is spanned by functions from $X^n$ to $\mathbb{K}$ defined via statistical mechanical state sums on the planar graphs derived from planar tangles. The third planar algebra, $\mathcal{O}^W = \cup \mathcal{O}^W_n$, differs from $\mathcal{P}^W$ only in that the graphs defining the functions need not be planar.

It turns out that $\mathcal{P}^W_n \subseteq \mathcal{O}^W_n \subseteq \mathcal{F}^W_n$. It is easy to compute $\dim \mathcal{F}^W_n$ using the Cauchy-Frobenius-Burnside formula for group characters. However, we are more interested in $\dim \mathcal{P}^W_n$, and it is generally very difficult to compute. Thus we consider when $\mathcal{P}^W_n = \mathcal{F}^W_n$ for all $n$. The planar algebra $\mathcal{O}^W$ plays an important role in this problem. In Section 4, we show that $\mathcal{P}^W = \mathcal{O}^W$ if and only if $\mathcal{P}^W_4$ contains a particular element called the “transposition”. The proof of this result is essentially skein theoretic in nature. Then in Section 5, show that $\mathcal{O}^W = \mathcal{F}^W$ whenever $|X|$ is not divisible by the characteristic of $\mathbb{K}$. Since this condition holds in characteristic zero, the most important case is thus treated. To prove this result, we encode $\mathcal{O}^W_n$ into polynomials and then appeal to results concerning the polynomial invariants of a finite group.

These results are related to Theorem 4.3 of [16] concerning a certain planar algebra $\mathcal{P}^\sigma$ which contains $\mathcal{P}^W$. This result asserts that any planar subalgebra of $\mathcal{P}^\sigma$ which contains a transposition is the set of elements of $\mathcal{P}^\sigma$ which are fixed under the action of some group $\mathcal{G}$ such that $\mathcal{G}(W) \subseteq \mathcal{G} \subseteq \mathcal{S}_X$. This result relies on the theory of subfactors, and so it is only applicable when the ground field is the real or complex numbers and when the matrix $W$ is symmetric. By introducing the intermediate planar algebra $\mathcal{O}^W$, we have extended this result (as applied to $\mathcal{P}^W$) to almost any field and to any matrix. In this case we also know precisely which group is involved. Moreover, the proof given here is combinatorial in nature, where the original was very non-combinatorial.


Planar algebras were introduced to study the structure of subfactors. True to their operator algebra origins, planar algebras are defined in terms of operators on vector spaces. These operators are defined diagrammatically by objects known as planar tangles. A planar tangle can be presented in several ways. We shall use a slight variation of the operadic definition of [15] (see also [16]). From this point of view, a planar tangle consists of a collection
of disjoint disks which are joined by disjoint smooth curves, together with a coloring of the regions formed by the strings and disks. Various constraints on this collection arise from their subfactor origins; however, no knowledge of subfactors is necessary to proceed.

We begin with a definition of a planar tangle. Let $D_0$ denote the unit disk. Pick disjoint disks $D_1, D_2, \ldots, D_n$ in the interior of $D_0$. Form a finite collection of disjoint “strings” (simple smooth curves) in the interior of $D_0 \setminus \bigcup_{i=1}^{n} D_i$, all of whose endpoints meet the boundary of some disk transversally. There may be some closed loops which touch no disk. Further assume that an even number of strings touch each disk, say $2k_i$ touching $D_i$. Color the regions interior to $D_0 \setminus \bigcup_{i=1}^{n} D_i$ formed by the strings black and white so that regions on either side of a string have opposite colors. Call the points on the boundary of each disk where a string touches “marked”. The marked points divide the boundary of each disk into intervals. On each disk, one of the intervals which touches a white region is chosen to be “privileged”. The entire boundary of a disk with no marked points is either privileged or not according to whether it touches a white region or a black region. Specifying the privileged intervals makes the coloring data redundant. It will sometimes be convenient to number the marked points on each disk consecutively in a clockwise direction where the marked point at the clockwise end of the privileged interval is numbered one.

The smooth isotopy class of this collection of disks, strings, coloring, and privileged intervals is called a planar $k_0$-tangle. There is a natural composition for planar tangles. If $S$ is a planar $k$-tangle with an internal disk $D_i$ with $2k_i$ marked points and $T$ is a planar $k_i$-tangle, then we may replace $D_i$ with a rescaled and isotoped version of $T$ without its unit disk, by matching corresponding marked points (first to first, etc.) and smoothing the connections of the strings. The coloring conventions are preserved by composition. The collection of planar tangles with this composition is called the planar operad.

A general planar algebra is a graded vector space $V_k$ for $k > 0$ and two vector spaces $V_+$ and $V_-$ such that every element $T$ of the planar operad determines a multilinear map from a tensor product of these vector spaces, one for each internal disk of $T$, to the vector space corresponding to the boundary of $T$. We require a natural homomorphism property. Given planar tangles $T_1, T_2, T_3$ which admit compositions of $T_2$ into $T_1$ and $T_3$ into this composition, the net result of these compositions in the planar operad does not depend upon the order in which they are carried out: The same must be true for the corresponding multilinear maps on the planar algebra. We also impose a condition on $V_+$ and $V_-$. We view these two vector spaces as corresponding to the two colorings of any planar 0-tangle–$V_+$ to those colored black next to the unit disk and $V_-$ to those colored white next to the unit disk. Observe that surrounding the interior component with a
closed string reverses its coloring. We require that surrounding the interior components of a planar 0-tangle with two closed strings yields a multilinear map to $V_+$ or $V_-$ which differs from the original multilinear map only by a fixed scalar multiple.

Let $V$ denote a planar algebra. Then $V$ is said to be finite dimensional whenever the vector spaces $V_+, V_-$, and $V_k$ ($k > 0$) are all finite dimensional. All of the examples that we shall consider in this paper are finite dimensional. In fact, $V_+$ and $V_-$ will both be one-dimensional, making the examples planar algebras in the sense of [15]. Given that $V$ is finite dimensional, it is natural to compute $\dim V_k$ for all $k$. This problem motivates the results of this paper. We are interested in a singly generated planar algebra $P^W$ which is contained in a planar algebra $F^W$ whose dimensions we can compute. We shall consider when these two planar algebras are equal. We now describe the planar algebras which we shall study.


3.1. The planar algebra of functions on a finite set. We present a very simple planar algebra. We are more interested in some of its planar subalgebras, but we take this opportunity to describe the multilinear map corresponding to each planar tangle with no other distractions. This correspondence will be the same in all planar algebras which follow.

Let $X$ denote a finite, nonempty set, and let $K$ denote a field. For each positive integer $k$, let $X_k$ be the vector space of all functions from $X$ to $K$ ($k > 0$), with $X_+$, $X_-$ identified with $K$. Then $X = \bigcup X_k$ is a planar algebra. Let $T$ denote a planar $k_0$-tangle with internal disks $D_1, D_2, \ldots, D_n$ with respectively $2k_1, 2k_2, \ldots, 2k_n$ many marked points. Then $T$ defines a multilinear map $\bigotimes_{i=1}^n X_{k_i} \to X_{k_0}$ as follows. Index the black regions of $T$ by $1, 2, \ldots, m$. For all $i$ ($0 \leq i \leq n$) and for all $j$ ($1 \leq j \leq k_i$), let $S_{ij}$ be the index of the $j$th black incident with $D_i$ when traversing the boundary of $D_i$ clockwise so that the privileged interval is traversed last. Given $f_i \in X_{k_i}$ ($1 \leq i \leq n$), define a function $Z_T^{(f_1, f_2, \ldots, f_n)} : X_{k_0} \to K$ which, when evaluated at $(x_1, x_2, \ldots, x_{k_0})$, returns

$$\sum_{\sigma} \prod_{i=1}^n f_i(\sigma(S_{i1}), \sigma(S_{i2}), \ldots, \sigma(S_{ik_i})),$$

where $\sigma$ runs over all maps from $\{1, 2, \ldots, m\}$ to $X$ with $\sigma(S_{0j}) = x_j$.

Extend $Z_T$ multilinearly to a map $\bigotimes_{i=1}^n X_i \to X_{k_0}$. The homomorphism property of planar algebras follows since the function only depends upon the incidences of the black regions and composition merges regions with the same color. Enclosing a planar 0-tangle with two closed strings preserves the color of the interior (reverses it twice) but adds an isolated black band. This modified tangle gives a multilinear map which is $|X|$ times the multilinear map corresponding to the original tangle. Thus $X$ is a planar algebra.
3.2. Planar algebras constructed via finite group action. Let $X$ denote a finite, nonempty set. Let $\mathfrak{S}_X$ denote the symmetric group on $X$. For each positive integer $k$, extend the action of $\mathfrak{S}_X$ to $X^k$ in the natural fashion: For all $g \in \mathfrak{S}_X$ and for all $(x_1, x_2, \ldots, x_k) \in X^k$, let $g(x_1, x_2, \ldots, x_k) = (g(x_1), g(x_2), \ldots, g(x_k))$. Let $\mathfrak{S} \subseteq \mathfrak{S}_X$ denote a subgroup so that $\mathfrak{S}$ is a permutation group on $X$. By a $\mathfrak{S}$-orbit of $X^k$, we mean a nonempty subset $Y \subseteq X^k$ such that $\vec{x}, \vec{y} \in Y$ if and only if there exists $g \in \mathfrak{S}$ such that $\vec{x} = g(\vec{y})$.

Let $K$ denote a field. For each positive integer $k$, let $F_k(\mathfrak{S}, X)$ denote the vector space of functions from $X^k \to K$ which depend only upon the $\mathfrak{S}$-orbit of their inputs. We identify the vector spaces $F_{i}(\mathfrak{S}, X)$ with the field $K$ (constant functions). Together, these vector spaces form a planar algebra $\mathcal{F}(\mathfrak{S}, X)$ with the same planar structure as $X$. That is to say, (1) defines a map $\bigotimes_{i=1}^{n} F_{k_i}(\mathfrak{S}, X) \to F_{k_0}(\mathfrak{S}, X)$. To see that this is so, pick $f_i \in F_{k_i}(\mathfrak{S}, X)$ (1 $\leq i \leq n$), and define $f_0 : X^{k_0} \to K$ by (1). To see that $f_0$ is constant on each $\mathfrak{S}$-orbit of $X^{k_0}$, consider replacing each map $\sigma$ in (1) by $g \circ \sigma$. The effect of this change on the boundary leads (1) to return $f_0(g(x_1, x_2, \ldots, x_{k_0}) = f_0(x_1, x_2, \ldots, x_{k_0})$ since $f_i \in F_{k_i}(\mathfrak{S}, X)$ (1 $\leq i \leq n$). Thus $f_0 \in F_{k_0}(\mathfrak{S}, X)$. Hence $\mathcal{F}(\mathfrak{S}, X) = \cup F_k(\mathfrak{S}, X)$ is a planar algebra. $\mathcal{F}(\mathfrak{S}, X)$ is called the fixed-point planar algebra of $\mathfrak{S}$ acting on $X$. This planar algebra is discussed in [15].

The vector spaces of $\mathcal{F}_k(\mathfrak{S}, X)$ are finite dimensional, and their dimension can be computed via the Cauchy-Frobenius-Burnside formula for the characters of the group, which we briefly recall now. See [11], for example. The permutation representation of $\mathfrak{S}$ acting on $X$ is the map $g \mapsto R(g) \in \mathcal{M}_X(\mathbb{C})$ with $(x, y)$-entry equal to 1 if $y = g \cdot x$ and 0 otherwise $(x, y \in X)$. The permutation character of $\mathfrak{S}$ acting on $X$ is the map $\pi : \mathfrak{S} \to \mathbb{C}$ given by $\pi(g) = \text{Tr} R(g)$ $(g \in \mathfrak{S})$. Fix a positive integer $k$. Then the number of orbits of $X^k$ under the action of $\mathfrak{S}$ is

$$\dim \mathcal{F}_k(\mathfrak{S}, X) = \frac{1}{|\mathfrak{S}|} \sum_{g \in \mathfrak{S}} (\pi(g))^k. \quad (2)$$

3.3. A planar algebra $\mathcal{P}^W$. Fix a field $K$. Let $X$ denote a finite, nonempty set. Let $\mathcal{M}_X(K)$ denote the set of matrices with rows and columns indexed by $X$ and entries in $K$. Pick $W \in \mathcal{M}_X(K)$. Given $k > 0$, we describe a rule using $W$ which maps any planar $k$-tangle all of whose internal disks have exactly 4 marked points to a function $X^k \to \mathbb{C}$. The vector space $\mathcal{P}_k^W$ spanned by these functions will be part of the grading of a planar algebra $\mathcal{P}^W$. A similar rule gives the vector spaces $\mathcal{P}^W_+$ and $\mathcal{P}^W_-$, which turn out to be isomorphic to $K$.

Let $T$ denote planar $k$-tangle in the unit disk $D_0$ with $n$ internal disks $D_1, D_2, \ldots, D_n$ each having exactly 4 marked points. As in Subsection 3.1, label the black regions of $T$ with indices 1, 2, \ldots, $m$ and for all $i$
(0 ≤ i ≤ n) and for j = 1, 2, let $S_{ij}$ denote the index of the $j^{th}$ black region incident with $D_i$ when traversing the boundary of $D_i$ clockwise so that the privileged interval is traversed last. Define a function $Z^W_T : X^k \to \mathbb{K}$ which, when evaluated at $(x_1, x_2, \ldots, x_k)$, returns

$$Z^W_T(x_1, x_2, \ldots, x_k) = \sum_{\sigma} \prod_{i=1}^{n} W(\sigma(S_{i1}), \sigma(S_{i2})),$$

where $\sigma$ runs over all maps from $\{1, 2, \ldots, m\}$ to $X$ with $\sigma(S_{0j}) = x_j$. Let $P_k^W$ denote the $\mathbb{K}$-linear span of all functions which arise in this fashion from a planar $k$-tangle. For $k = 0$, we use the same rule to define functions, but place them in $P^W_+$ or $P^W_-$ when the color of the 0-tangle near the unit circle is black or white, respectively.

Now $P^W = \bigcup P^W_n$ is a planar subalgebra of $X$ with closure under (1) assured since the composition of planar tangles yields a planar tangle. The planar algebra $P^W$ is called the planar algebra generated by $W$. This planar algebra is also discussed in [15]. Singly generated planar algebras are considered in [3] as well. Not all of the planar algebras considered in [3] are generated by a matrix, however.

By an automorphism of $W$, we mean a permutation $s$ of $X$ such that $W(u, v) = W(s(u), s(v))$ for all $u, v \in X$. Let $\mathfrak{A}(W)$ denote the full group of automorphisms of $W$. Observe that $P^W \subseteq \mathcal{F}(\mathfrak{A}(W), X)$ since the definition of the functions in $P^W$ depend only upon the structure of $W$. Thus (2) gives an upper bound for dim $P_k^W$. Our main result concerns the case of equality. The proof compares $P^W$ and $\mathcal{F}(\mathfrak{A}(W), X)$ to an intermediate planar algebra which we now describe.

### 3.4. A planar algebra $O^W$.

We use the language of graph theory to generalize the construction of $P^W$ of the previous subsection. We begin with some graph theoretic terminology.

By a multi-digraph, we mean a pair $\Delta = (V, E)$, where $V$ is a nonempty set and $E$ is a multiset of ordered pairs of (not necessarily distinct) elements of $V$. Let $\Delta = (V, E)$ be a multi-digraph. The elements of $V$ are called the vertices of $\Delta$, and an ordered pair $(u, v) \in E$ is called a (directed) edge from $u$ to $v$. We say that there are multiple edges from $u$ to $v$ whenever the multiset $E$ contains two or more copies of $(u, v)$. Throughout this paper we shall assume that all multi-digraphs have finite vertex and edge sets. Fix a nonnegative integer $n$. By an open graph of boundary size $n$, we mean a triple $\Gamma = (V, E, \vec{b})$, where $(V, E)$ is a multi-digraph and $\vec{b}$ is an $n$-tuple of elements of $V$, called the boundary vector of the open graph. Let $O_n$ denote the set of all open graphs of boundary size $n$.

Let $\Gamma = (V, E, \vec{b})$ denote an open graph of boundary size $n$. $\Gamma$ is said to be planar if the multi-digraph $(V, E)$ has a plane embedding (no crossing edges) into the interior of an $n$-gon with clockwise ordered vertices $b'_1, b'_2,
... such that each $b_i$ can be joined to $b'_i$ in a planar way. A planar open graph may be viewed as a patch which has been cut out of a plane embedded planar graph. The boundary vertices may have neighbors in the larger graph, while all neighbors of non-boundary vertices must appear in the open graph. Let $\mathcal{P}_n$ denote the set of all planar open graphs of boundary size $n$.

As in the previous subsection, we fix a field $\mathbb{K}$, a finite, nonempty set $X$, and a matrix $W \in \mathcal{M}_X(\mathbb{K})$. The planar tangles which define $\mathcal{P}_W$ can be interpreted as open graphs. Let $T$ denote a planar $k$-tangle all of whose internal disks have exactly 4 marked points, and adopt the notation of Subsection 3.1. Define a multi-digraph whose vertex set $V$ consists of the black regions of $T$, whose edge set $E$ consists of the pairs $(S_{i1}, S_{i2})$ as $i$ runs over indices of the internal disks, and whose boundary vector is $\vec{b} = (S_{01}, S_{02}, \ldots, S_{0k})$. It is not difficult to see that $(V, E, \vec{b})$ is a planar open graph and that every planar open graph arises in this way.

The data $\Gamma = (V, E, \vec{b})$ suffices to define the multilinear map $Z^W_\Gamma : X^k \to \mathbb{K}$ corresponding to the planar $k$-tangle $T$ as in (3). The evaluation of $Z^W_\Gamma$ at $(x_1, x_2, \ldots, x_k) \in X^k$ is

$$Z^W_\Gamma(x_1, x_2, \ldots, x_k) = \sum_{\sigma} \prod_{(u,v) \in E} W(\sigma(u), \sigma(v)),$$

where $\sigma$ runs over all maps from $V$ to $X$ with $\sigma(b_i) = x_i$ ($0 \leq i \leq n$).

The construction (4) is well-known in statistical mechanics [1], [2], [23] and [24]. Thus we adopt the following terminology: The elements of $X$ are called spins, and $W$ is called the Boltzman weight matrix. A map $\sigma : V \to X$ with $\sigma(b_i) = x_i$ ($1 \leq i \leq n$) is called a state compatible with the boundary condition $\sigma(b_i) = x_i$. The formula (4) is called the partition function of $\Gamma$ with respect to $W$.

If we restrict $\Gamma$ to planar open graphs, then (4) and (1) agree when $\Gamma$ is produced from $T$ as above. Thus $\mathcal{P}_W$ is the vector space spanned by the functions defined by (4) as $\Gamma$ runs over all planar open graphs of boundary size $k$. However, the planar structure is not necessary in (4). Let $\mathcal{O}^W_k$ denote the vector space of functions $X^k \to \mathbb{K}$ spanned by the functions defined by (4) as $\Gamma$ runs over all open graphs of boundary size $k$. Then $\mathcal{O}^W = \cup \mathcal{O}^W_k$ is a planar subalgebra of $\mathcal{X}$. The closure of $\mathcal{O}^W$ under the multilinear maps defined by planar tangles follows since such operations just combine graphs to form a new graph. We call $\mathcal{O}^W$ the open graph planar algebra of $W$.

By construction $\mathcal{P}_W \subseteq \mathcal{O}^W \subseteq F_k(\text{Aut}(W), X)$. Our main results describe when $\mathcal{P}_W = \mathcal{O}^W$ and when $\mathcal{O}^W = F_k(\text{Aut}(W), X)$. Before proceeding to these results, we present a graph theoretic interpretation of the partition function. Let $\Gamma = (V, E)$ and $\Xi = (X, R)$ denote graphs. By a graph homomorphism from $\Gamma$ into $\Xi$, we mean a map $\sigma : V \to X$ such that if
Lemma 3.1. Suppose \( W \) is the adjacency matrix of a graph \( \Xi = (X, R) \), and let \( \Gamma = (V, E, \vec{b}) \) denote an open graph of boundary size \( n \) for some fixed nonnegative integer \( n \). Then for all \( \vec{p} \in X^n \), \( Z_{\Xi}^\Gamma(\vec{p}) \) equals the number of graph homomorphisms from \( (V, E) \) to \( \Xi \) which map \( \vec{b} \) to \( \vec{p} \) coordinate-wise.

Proof. Each state \( \sigma \) over which the sum in (4) runs maps \( \vec{b} \) to \( \vec{p} \) element-wise. If \( \sigma \) is not a graph homomorphism, then \( (\sigma(u), \sigma(v)) \notin R \) for some \( (u, v) \in E \), so \( W(\sigma(u), \sigma(v)) = 0 \) and the state contributes nothing to the partition function. If \( \sigma \) is a graph homomorphism, then \( W(\sigma(u), \sigma(v)) = 1 \) for all \( (u, v) \in E \), so the state adds one to the partition function. \( \square \)

The problem of determining if there is a graph homomorphism into a fixed graph \( H \) (the so-called \( H \)-coloring problem) is NP-complete in general [9] and [10]. In particular, one cannot expect to find a particularly efficient means of computing the partition functions of open graphs with respect to a fixed matrix \( W \).

4. When \( P^W = O^W \).

We consider when \( P^W = O^W \). Of course this is the case when the partition function with respect to \( W \) of every open graph is a linear combination of the partition functions with respect to \( W \) of some planar open graphs. We give a more practical characterization involving just one special open graph.

Let \( \Phi \) denote the (non-planar) open graph of boundary size 4 consisting two isolated vertices \( v_1 \) and \( v_2 \) and boundary vector \((v_1, v_2, v_1, v_2)\). We call \( \Phi \) the transposition. We picture \( \Phi \) as a 4-tangle in Figure 1(b)—two black “ribbons” which cross, one above the other, without interacting. When drawing our tangles, we shall avail ourselves of the fact that they are determined only up to isotopy and draw the disks as squares. We mark the privileged interval on each with a \( \diamond \), so it is unnecessary to draw the coloring of the regions.

![Image of open graph and tangle](image)

Figure 1. Two views of \( \Phi \).

We now use the transposition \( \Phi \) to build (non-planar) tangles which define operators on open graphs which transpose elements of the boundary vector.
For all $n \geq 4$ and $m$ ($1 \leq m \leq n$), form an $n$-tangle $\Phi^n_m$ with one interior disk $D_1$ with $2n$-marked points by joining the $i$th marked point of $D_1$ to the $i$th marked point of the unit circle of $\Phi^n_m$ for all $i$ except $2m$, $2m+1$, $2m+2$, and $2m+3$ (taken mod $2n$). The $2m$th, $2m+1$st, $2m+2$nd, and $2m+3$rd marked points of $D_1$ are joined to the $2m+2$nd, $2m+3$rd, $2m$th, and $2m+1$st marked points on the unit disk of $\Phi^n_m$, respectively (see Figure 2). Observe that each $\Phi^n_m$ is formed by composing a planar $n$-tangle with $\Phi$—simply cut out a disk around the transposition.

An examination of the tangle presentation of $\Phi^n_m$ reveals that it transposes the order in which the $m$th and $m+1$st black regions are encountered when the unit disk is traversed clockwise versus their order on the interior disk. Taking composition of planar tangles as the product, the $\Phi^n_m$ generate the symmetric group on the $n$ black regions incident with the unit disk. In particular, the various compositions of the $\Phi^n_m$ give rise to all permutations of the black regions when considering the order in which they appear around the unit disk versus the interior disk. Note that the construction (3) can be used to define an operator on open graphs from $\Phi^n_m$. By the above observations, we see that the resulting open graph operator, swaps the $m$th and $m+1$st (mod $n$) boundary vertices of its input.

\[ \Phi^n_m(V, E, (b_1, b_2, \ldots, b_m, b_{m+1}, \ldots, b_n)) = (V, E, (b_1, b_2, \ldots, b_{m+1}, b_m, \ldots, b_n)) \]

**Figure 2.** The transposition operator $\Phi^n_m$.

**Theorem 4.1.** Let $W$ denote a matrix over any field. Then the following are equivalent:

(i) $P^W = O^W$.

(ii) There exists $n \geq 4$ such that $P_n^W = O_n^W$.

(iii) $P_n^W = O_n^W$.

(iv) $Z^W_n \in P_n^W$.

**Proof.** (i) $\Rightarrow$ (ii): Clear.

(ii) $\Rightarrow$ (iii): Let $\Gamma$ denote an open graph of boundary size 4. Form an open graph $\Gamma'$ of boundary size $n$ by extending the boundary vector of $\Gamma$ by repeating the last vertex $n-4$ times. This is a planar operation corresponding to the planar tangle of Figure 3(a) (this is not the preferred inclusion of [15]). Composing Figure 3(a) into Figure 3(b) returns the original planar
tangle along with some closed loops with white interiors, which can be re-
moved with no effect in our planar algebra. By (ii), there exist open graphs
$\Gamma_1, \Gamma_2, \ldots, \Gamma_k \in \mathcal{P}_n$ such that $Z_{\Gamma}^W = \sum_{j=1}^{k} \alpha_j Z_{\Gamma_j}^W$ for some scalars $\alpha_j$. The
same is true of $\Gamma$ since we may apply the restriction to these functions.

\begin{figure}
\centering
\begin{subfigure}{0.5\textwidth}
\includegraphics[width=\linewidth]{inclusion}
\caption{An inclusion}
\end{subfigure} \hspace{0.5cm}
\begin{subfigure}{0.5\textwidth}
\includegraphics[width=\linewidth]{inverse}
\caption{Its inverse restriction}
\end{subfigure}
\caption{Two planar tangles.}
\end{figure}

(iii) $\Rightarrow$ (iv): Clear.

(iv) $\Rightarrow$ (i): Fix a nonnegative integer $n$ and pick $\Gamma \in \mathcal{O}_n$. We shall show
that there are open graphs $\Gamma_1, \Gamma_2, \ldots, \Gamma_k \in \mathcal{P}_n$ such that $Z_{\Gamma}^W = \sum_{j=1}^{k} \alpha_j Z_{\Gamma_j}^W$
for some scalars $\alpha_j$. This will prove that $\mathcal{O}_n^W = \mathcal{P}_n^W$.

In order for an open graph to be planar, it must be possible to embed it in
the plane so that the positions of its boundary vertices are incident with the
exterior and ordered clock-wise as they appear in the boundary vector. This
may not be the case for $\Gamma$. However, by permuting the boundary vector
this can be corrected. By (iv) and the remarks at the beginning of the section,
there exist transposition operators such that

$$\Gamma = \Phi_{m_1}^n(\Phi_{m_2}^n(\ldots \Phi_{m_j}^n(\hat{\Gamma}) \ldots)),$$

where $\hat{\Gamma}$ is an open graph with the same vertex and edge sets as $\Gamma$ and
boundary vector a re-ordering of that of $\Gamma$ so that all repetitions occur in
cyclically successive positions. It is now possible to embed $\hat{\Gamma}$ with the desired
boundary property. There remains the possibility that $\hat{\Gamma}$ has crossing edges
in any plane embedding with the boundary vertices incident with the exterior
face.

In light of (iv), we now only need to prove that $Z_{\Gamma}^W \in \mathcal{P}_n^W$. Indeed,
suppose that this is the case. Then there exists a set of planar open graphs
$\hat{\Gamma}_1, \hat{\Gamma}_2, \ldots, \hat{\Gamma}_j$ such that $Z_{\Gamma}^W = \sum_{i=1}^{j} \beta_j Z_{\Gamma_i}^W$. Now by (iv) there exists a set
of planar open graphs $\tilde{\Gamma}_1, \tilde{\Gamma}_2, \ldots, \tilde{\Gamma}_j$ such that

$$Z_{\Phi_{m_j}^n(\hat{\Gamma})}^W = \sum_{\ell=1}^{\tilde{\gamma}} Z_{\tilde{\Gamma}_\ell}^W \in \mathcal{P}_n^W.$$
successive positions. Embed $\Delta$ in the plane such that all of its vertices lie evenly spaced on a circle and its boundary vertices are ordered clockwise as they appear in the boundary vector. If no edges cross in this embedding, we are done. Suppose that some edges cross. Among all vertices $p$ and $q$ which are incident with crossing edges $(p, p')$ and $(q, q')$ pick those which are cyclically nearest according to their positions on the circle. Observe that $p$ and $q$ partition the remaining vertices into two sets according to which sides of $p$ and $q$ they lie. Moreover, by the choice of $p$ and $q$ nearest, there are no edges between these two sets. By deforming the edges of this embedding, we can make it so that all edges which cross $(p, p')$ and $(q, q')$ do so between $p'$ and $x$ or between $q'$ and $x$ without creating any new crossings, where $x$ is the point in the plane where $(p, p')$ and $(q, q')$ cross. Factor this crossing as two non-crossing edges (through which all edges crossing $(p, p')$ and $(q, q')$ pass as if nothing has changed) and a transposition—see Figure 4. Now by (iv), the transposition belongs to $\mathcal{P}_W^4$. Thus there exist open graphs $\Delta_1, \Delta_2, \ldots, \Delta_h$ such that

$$Z_W^\Delta = \sum_{\ell=1}^h \gamma_\ell Z_W^{\Delta_\ell} \in \mathcal{P}_n^W,$$

and which differ from $\Delta$ only in that under a similar embedding the crossing $(p, p')$ and $(q, q')$ has been replaced by a planar graph. Proceeding by induction (on the number of vertices between the endpoints of crossing edges as one takes the shortest path along the circle), we may remove all crossings in $\Delta$. Thus $Z_W^\Delta \in \mathcal{P}_n^W$, as desired. \hfill $\square$

![Figure 4. Factoring crossing edges.](image)

The arguments of this section suggest the relations of the planar algebras $\mathcal{O}_W$ and $\mathcal{P}_W$ be interpreted as a graph rewriting system. Let $\Delta_1, \Delta_2, \ldots, \Delta_k$ be open graphs with the same boundary size, and say $\sum_{i=1}^k \alpha_i \Delta_i = 0$ (modulo $W$) when $\sum \alpha_i Z_W^\Delta = 0$. The homomorphism property for planar algebras make this relation a “local rewriting rule”. Suppose $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$ are graphs which are identical everywhere except on patch where the subgraph of $\Gamma_i$ is isomorphic to $\Delta_i$. Then $\sum_{i=1}^k \alpha_i \Gamma_i = 0$ (modulo $W$). Moreover, by construction the linear extension of the partition function is an invariant of the associated graph rewriting system. This sort of graph
relation is similar to the formal combinations of diagrams used by knot theorists, such as in Conway’s tangles and skein relations [4] and the invariant is like a spin model [8] and [13]. Thus, planar algebras provide a foundation for a skein theoretic approach to certain graph rewriting problems (this not the standard notion of graph rewriting [20], [5] and [6], although open graphs are used in [17] to study graph rewriting).

5. When $O^W = F^W$.

Let $\mathbb{K}$ denote a field, let $X$ denote a finite, nonempty set, and pick $W \in \mathcal{M}_\mathbb{K}(X)$. Write $F^W$ in place of $\mathcal{F}(\text{Aut}(W), X)$. We show that $O^W = F^W$ whenever the characteristic of $\mathbb{K}$ does not divide $|X|!$. In particular, $O^W = F^W$ whenever $\mathbb{K}$ has characteristic zero.

Lemma 5.1. Pick $W \in \mathcal{M}_\mathbb{K}(X)$, and fix a nonnegative integer $n$. Then the following are equivalent:

(i) $O^W_n = F^W_n$.

(ii) For all $\vec{p}, \vec{q} \in X^n$, $Z^W_\Delta(\vec{p}) = Z^W_\Delta(\vec{q})$ for all $\Delta \in O_n$ implies that $\vec{p}$ and $\vec{q}$ belong to the same $\text{Aut}(W)$-orbit of $X^n$.

Proof. For all $\vec{p}, \vec{q} \in X^n$, $\vec{p}$ and $\vec{q}$ belong to the same $\text{Aut}(W)$-orbit of $X^n$ if and only if $f(\vec{p}) = f(\vec{q})$ for all $f \in F^W_n$ by the definition of $F^W_n$. The equivalence of (i) and (ii) follows since $O^W_n \subseteq F^W_n$ and $O^W_n = \text{span}\{Z^W_\Gamma \mid \Gamma \in O_n\}$.

We shall prove that Condition (ii) of Theorem 5.1 holds whenever the characteristic of $\mathbb{K}$ does not divide $|X|!$. In fact, we only need to consider $\vec{p}, \vec{q} \in X^n$ which differ by a permutation of $X$.

Lemma 5.2. Pick $\vec{p}, \vec{q} \in X^n$. If $Z^W_\Delta(\vec{p}) = Z^W_\Delta(\vec{q})$ for all $\Delta \in O_n$, then there exists $s \in S_X$ such that $\vec{p} = s\vec{q}$.

Proof. Observe that there exists $s \in S_X$ with $\vec{p} = s\vec{q}$ precisely when $p_i = p_j$ if and only if $q_i = q_j$ $(1 \leq i, j \leq n)$. Suppose there exists some $i, j$ $(1 \leq i < j \leq n)$ such that $p_i = p_j$ but $q_i \neq q_j$. Let $\Gamma$ denote the open graph of boundary size $n$ consisting of $n - 1$ isolated vertices, each appearing once on the boundary except one that is both the $i^{\text{th}}$ and $j^{\text{th}}$ boundary vertex. Then $Z^W_\Gamma(\vec{p}) = 1$ and $Z^W_\Gamma(\vec{q}) = 0$.

The idea behind the following argument is to fix some nonnegative integer $n$ and some $\vec{p} \in X^n$ and then reconstruct $W$ from the data $\{\Gamma, Z^W_\Gamma(\vec{p}) \mid \Gamma \in O_n\}$. This means that this information is sufficient to determine the $\text{Aut}(W)$-orbit of $\vec{p}$. We do this reconstruction by encoding this data as a set of polynomials and then showing that $W$ is essentially the only simultaneous zero of these polynomials (at least when the characteristic of $\mathbb{K}$ does not divide $|X|!$).
Let $\mathbb{K}$ denote the algebraic closure of $\mathbb{K}$. Let $\mathcal{L}$ denote the polynomial ring over $\mathbb{K}$ in the variables $\ell_{uv}$ ($u, v \in X$). We evaluate these polynomials over $\mathcal{M}_X(X)$ since the variables are indexed by $X \times X$. Let $L \in \mathcal{M}_L(X)$ denote the matrix whose $(u, v)$-entry is the variable $\ell_{uv}$. Observe that $s \in \mathcal{G}_X$ acts on $\mathcal{L}$ by $s(\ell_{uv}) = \ell_{s(u)s(v)}$ ($u, v \in X$). Similarly, $s$ acts on $\mathcal{M}_X(X)$ by $(sM)_{u,v} = M_{s(u),s(v)}$ ($u, v \in X$) for all $M \in \mathcal{M}_X(X)$.

For any nonnegative integer $n$ and for all $\vec{p} \in X^n$, let $E(\vec{p}) = \{Z^L_X(\vec{p}) - Z^W_X(\vec{p}) | \Delta \in \mathcal{O}_n\}$. Let $Z(\vec{p})$ denote the affine variety over $\mathbb{K}$ defined by $E(\vec{p})$ (the common zeros of all polynomials in $E(\vec{p})$). We view $Z(\vec{p})$ as a subset of $\mathcal{M}_X(X)$. Observe that $W \in Z(\vec{p})$.

There is a trivial symmetry of $E(\vec{p})$ and $Z(\vec{p})$ which arises because the polynomial $Z^L_X(\vec{p})$ will not change if we permute the spins not in $\vec{p}$. Let $\text{stab}_{\mathcal{G}_X}(\vec{p})$ denote the subgroup of $\mathcal{G}_X$ which fixes the spins in $\vec{p}$ pointwise. Note that $\text{stab}_{\mathcal{G}_X}(\vec{p})$ is isomorphic to $\mathcal{G}_{X\backslash\vec{p}}$. When $n = 0$, $\vec{p}$ is the empty vector and $\text{stab}_{\mathcal{G}_X}(\vec{p}) = \mathcal{G}_X$.

The next result shows that Condition (ii) of Theorem 5.1 can be restated in terms of $Z(\vec{p})$ and $Z(\vec{q})$. In light of Lemma 5.2, we need only consider $\vec{q}$ of the form $\vec{s}\vec{p}$ for some $s \in \mathcal{G}_X$.

**Lemma 5.3.** Pick $s \in \mathcal{G}_X$ and $\vec{p} \in X^n$. The following are equivalent:

(i) $Z^W_X(\vec{p}) = Z^W_X(s\vec{p})$ for all $\Delta \in \mathcal{O}_n$.

(ii) $sE(\vec{p}) = E(s\vec{p})$.

(iii) $sZ(\vec{p}) = Z(s\vec{p})$.

Moreover, (i)-(iii) hold when $s \in \text{stab}_{\mathcal{G}_X}(\vec{p})$ and when $s \in \text{Aut}(W)$.

**Proof.** Observe that for all $s \in \mathcal{G}_X$, $sZ^L_X(\vec{p}) = Z^L_X(s\vec{p})$ since the sum defining the partition function runs over all states satisfying the boundary condition. Thus $s(Z^L_X(\vec{p}) - Z^W_X(\vec{p})) = Z^L_X(s\vec{p}) - Z^W_X(s\vec{p}) \in sE(\vec{p})$, and $Z^L_X(s\vec{p}) - Z^W_X(s\vec{p}) = Z^L_X(s\vec{p}) - Z^W_X(s\vec{p}) \in E(s\vec{p})$. The equivalence of (i)-(iii) follows. Clearly (i) holds when $s \in \text{stab}_{\mathcal{G}_X}(\vec{p})$ and when $s \in \text{Aut}(W)$.

Our problem is now reduced to showing that if $sZ(\vec{p}) = Z(s\vec{p})$ for some $s \in \mathcal{G}_X$, then $\vec{p}$ and $s\vec{p}$ belong to the same $\text{Aut}(W)$-orbit of $X^n$. If $s$ is in either of the groups identified in Lemma 5.3, then $\vec{p}$ and $s\vec{p}$ belong to the same $\text{Aut}(W)$-orbit of $X^n$. We shall show that if the characteristic of $\mathbb{K}$ does not divide $|X|$, then $\text{Aut}(W)\text{stab}_{\mathcal{G}_X}(\vec{p}) := \{st | s \in \text{Aut}(W), t \in \text{stab}_{\mathcal{G}_X}(\vec{p})\}$ is the complete set of permutations $s$ such that $sZ(\vec{p}) = Z(s\vec{p})$. We will then use this fact to complete our proof. Our goal now is to describe $Z(\vec{p})$ exactly. To do so, we use some facts about polynomial invariants of finite groups as applied to $\mathcal{L}$.

For all subgroups $\mathcal{G} \subseteq \mathcal{G}_X$, let $\mathcal{L}^\mathcal{G}$ denote the ring of invariants of $\mathcal{L}$ under the action of $\mathcal{G}$:

$$\mathcal{L}^\mathcal{G} = \{ f \in \mathcal{L} | f(M) = (s(f))(M) \text{ for all } s \in \mathcal{G}, M \in \mathcal{M}_X(X) \}.$$
See [21] and [22] for more on polynomial invariants of finite groups. Noether’s original work on the subject can be found in [18] and [19].

We shall show that under suitable conditions, \( E(\vec{p}) \) actually spans the ring of invariants of \( L \) under the action of \( \text{stab}_{\Theta X}(\vec{p}) \). We will then be able to appeal to the following result to describe \( Z(\vec{p}) \) exactly:

**Lemma 5.4.** Pick \( M \in \mathcal{M}_G(X) \). Then the set of common zeros of \( \{ f - f(M) \mid f \in \mathcal{L}^\Theta \} \) is \( \Theta \cdot M := \{ gM \mid g \in \Theta \} \).

**Proof.** Suppose \( M' \notin \Theta \cdot M \). Then \( \Theta \cdot M \) and \( \Theta \cdot M' \) are disjoint finite sets. Thus there exists a polynomial \( h \in \mathcal{L} \) such that \( h(M') = 1 \) and \( h(M) = 0 \) for all \( g \in \Theta \). Now \( f = \prod_{g \in \Theta} gh \in \mathcal{L}^\Theta \) has the property that \( f(M) = 0 \) and \( f(M') = 1 \). Thus every zero of \( \{ f - f(M) \mid f \in \mathcal{L}^\Theta \} \) is in \( \Theta \cdot M \). The reverse containment is clear, so the result follows.

We now describe a simple criterion which ensures that we may apply the previous theorem. We deduced such a condition from Noether’s work. Let \( [\mathcal{L}^\Theta] \subseteq \mathcal{L}^\Theta \) denote the \( \mathbb{K} \)-linear span of the polynomials of the form \( \sum_{g \in \Theta} gm \), where \( m \) runs over all monomials in the variables \( \ell_{uv} \) \((u,v \in X)\). This sum is, up to a normalization constant, the so-called Reynolds operator applied to \( m \). We have the following result of Noether:

**Theorem 5.5** ([19] (Noether)). If \( \text{Char} \ K \nmid |\Theta| \), then \( [\mathcal{L}^\Theta] = \mathcal{L}^\Theta \).

It is this criterion of Noether which leads to our condition that the characteristic of \( K \) does not divide \( |X|! \). We now sandwich \( \text{span}(E(\vec{p})) \) between \([\mathcal{L}^\Theta]\) and \( \mathcal{L}^\Theta \) for \( \Theta = \text{stab}_{\Theta X}(\vec{p}) \). With the previous two results this gives an exact description of \( Z(\vec{p}) \) when the characteristic of \( K \) does not divide \( |X|! \).

**Lemma 5.6.** With the above notation,

\[
[\mathcal{L}^\text{stab}_{\Theta X}(\vec{p})] \subseteq \text{span}(E(\vec{p})) \subseteq \mathcal{L}^\text{stab}_{\Theta X}(\vec{p}).
\]

**Proof.** We first show that \([\mathcal{L}^\text{stab}_{\Theta X}(\vec{p})]\) is contained in the linear span of \( E(\vec{p}) \). Pick \( f \in [\mathcal{L}^\text{stab}_{\Theta X}(\vec{p})] \), and let \( m = \ell_{u_1v_1}^{\alpha_1} \ell_{u_2v_2}^{\alpha_2} \ldots \ell_{u_jv_j}^{\alpha_j} \) denote a monomial appearing in \( f \) (say with coefficient \( \alpha \in \mathbb{K} \)) having the maximal number of distinct indices not in \( \vec{p} \) appearing on the variables. Let \( \Delta = (U, D, \vec{p}) \) denote the open graph with \( U \) the set of spins which appear in \( \vec{p} \) or as a subscript of some variable in \( m \) and \( D \) the multiset which contains \( n_i \) copies of \((u_i, v_i)\) \((1 \leq i \leq j)\). We show that \( f - \alpha(|X| - |U|)!Z^L_\Delta(\vec{p}) \) has fewer monomials with as many distinct indices on the variables as \( m \) does. It will then follow from induction that \( f \in \text{span}(E(\vec{p})) \).

If every element of \( U \) appears in \( \vec{p} \), then \( Z^L_\Delta(\vec{p}) = m \) and \( \sum_{s \in \text{stab}_{\Theta X}(\vec{p})} sm = |\text{stab}_{\Theta X}(\vec{p})|m \) since \( m \) is fixed by \( \text{stab}_{\Theta X}(\vec{p}) \). This is the base case of the induction. Now suppose that not all indices of the variables in \( m \) are in \( \vec{p} \),
and consider the states $\sigma$ over which the sum in (4) runs. Observe that $\sigma$ is simply a map from $U$ to $X$ with the appropriate boundary condition, and it is either an injection or it is not. Suppose $\sigma$ is an injection. Then there are $(|X| - |U|)!$ many ways to extend $\sigma$ to a permutation of $X$. Any such permutation belongs to $\text{stab}_{\Sigma_X}(\vec{p})$ by the boundary condition, and conversely any element of $\text{stab}_{\Sigma_X}(\vec{p})$ restricts to a valid, injective state. In particular, if $(|X| - |U|)! = 0$, then $m$ cannot appear in $f$ with nonzero coefficient because this number is a factor of the number of repetitions of $m$. If $\sigma$ is not an injection, then fewer indices of variables appear in the corresponding summand of $Z^L_\Delta(\vec{p})$ than in $m$ because two or more have been identified by $\sigma$. Thus $\sum_{s \in \text{stab}_{\Sigma_X}(\vec{p})} g_m - \alpha(|X| - |U|)! (Z^L_\Delta(\vec{p}) - Z^W_\Delta(\vec{p}))$ consists only of monomial terms with fewer distinct indices appearing on the variables than in $m$. By the definition of $|L^{\text{stab}_{\Sigma_X}(\vec{p})}|$, every summand of $\sum_{s \in \text{stab}_{\Sigma_X}(\vec{p})} g_m$ appears in $f$. It follows by induction that $f \in E(\vec{p})$, thus proving the containment $[Z^{\text{stab}_{\Sigma_X}(\vec{p})}] \subseteq \text{span}(E(\vec{p}))$.

We now prove the containment $\text{span}(E(\vec{p})) \subseteq L^{\text{stab}_{\Sigma_X}(\vec{p})}$. Pick an open graph $\Gamma = (V, E, \vec{b})$ of boundary size $n$ and a permutation $s \in \text{stab}_{\Sigma_X}(\vec{p})$. Then applying $s$ to $Z^L_\Delta(\vec{p}) - Z^W_\Delta(\vec{p})$ has the same effect as applying $s$ to each state $\sigma$ over which the sum defining $Z^L_\Delta(\vec{p})$ runs. Since $s$ fixes $\vec{p}$ pointwise, the map $s\sigma$ is also another state satisfying the boundary condition. Thus $Z^L_\Gamma(\vec{p}) - Z^W_\Gamma(\vec{p}) \in L^{\text{stab}_{\Sigma_X}(\vec{p})}$.

Suppose the characteristic of $\mathbb{K}$ does not divide $|X|!$. Then Theorem 5.5 and Lemma 5.6 imply that $\text{span}(E(\vec{p})) = L^{\text{stab}_{\Sigma_X}(\vec{p})}$, so $Z(\vec{p}) = \text{stab}_{\Sigma_X}(\vec{p}) \cdot W$ by Lemma 5.4. It is this fact about $Z(\vec{p})$ which we shall use to complete our proof. We note that the condition on the characteristic of the field is sufficient but it is not necessary. However, this condition always holds in characteristic zero, which we consider the most important case. For the moment, we leave the problem of improving this sufficient condition as an open problem, but proceed with this in mind. Let us say that $\vec{p} \in X^n$ is SSS if $Z(\vec{p}) = \text{stab}_{\Sigma_X}(\vec{p}) \cdot W$. The above discussion gives us the following:

**Lemma 5.7.** Pick $\vec{p} \in X^n$. If $\text{Char} \mathbb{K} \nmid |X|!$, then $\vec{p}$ is SSS.

**Lemma 5.8.** Pick $s \in \Sigma_X$ and $\vec{p} \in X^n$. Suppose that $\vec{p}$ is SSS. Then the following are equivalent:

(i) $s \in \text{Aut}(W)\text{stab}_{\Sigma_X}(\vec{p})$.

(ii) $W \in sZ(\vec{p})$.

**Proof.** (i) $\Rightarrow$ (ii): Since $s \in \text{Aut}(W)\text{stab}_{\Sigma_X}(\vec{p})$, the equivalent conditions of Lemma 5.3 hold for $s \in \Sigma_X$ and $\vec{p} \in X^n$. In particular $sZ(\vec{p}) = Z(s\vec{p})$. Since $W \in Z(s\vec{p})$, (ii) follows.
implies that there exists \( t \in \text{stab}_{\mathcal{O}_X}(\vec{p}) \) such that \( s^{-1}W = tW \). Thus, \( stW = W \), so \( st \in \mathcal{A}(W) \) by definition. Now (i) follows.

\[ \square \]

Lemma 5.9. Pick \( \vec{p}, \vec{q} \in X^n \), and suppose that \( \vec{p} \) and \( \vec{q} \) are SSS. If \( Z^W_\Delta(\vec{p}) = Z^W_\Delta(\vec{q}) \) for all \( \Delta \in \mathcal{O}_n \), then \( \vec{p} \) and \( \vec{q} \) belong to the same \( \mathcal{A}(W) \)-orbit of \( X^n \).

Proof. By Lemma 5.2, there exists \( s \in \mathcal{S}_X \) with \( s\vec{p} = \vec{q} \). Now \( Z^W_\Delta(\vec{p}) = Z^W_\Delta(s\vec{p}) \) for all \( \Delta \in \mathcal{O}_n \), so \( sZ(\vec{p}) = Z(s\vec{p}) \) by Lemma 5.3. In particular, \( W \in sZ(\vec{p}) \) since \( W \in Z(\vec{p}) \) by Lemma 5.8. If \( s \in \text{stab}_{\mathcal{O}_X}(\vec{p}) \), then \( \vec{p} = \vec{q} \). Otherwise, \( s \not\in \text{stab}_{\mathcal{O}_X}(\vec{p}) \), so there must be an automorphism of \( W \) which maps \( \vec{p} \) to \( \vec{q} \). In either case, \( \vec{p} \) and \( \vec{q} \) belong to the same \( \mathcal{A}(W) \)-orbit of \( X^n \).

\[ \square \]

Theorem 5.10. Let \( \mathbb{K} \) denote a field, let \( X \) denote a finite, nonempty set, and pick \( W \in \mathcal{M}_\mathbb{K}(X) \). Suppose that \( \text{Char } \mathbb{K} \mid |X| ! \). Then \( \mathcal{O}^W = \mathcal{F}^W \).

Proof. Immediate from Lemmas 5.1, 5.7 and 5.9.

\[ \square \]

This completes our main results. We now give an example which shows that \( \mathcal{O}^W \) need not equal \( \mathcal{F}^W \) if \( |X| ! \) is divisible by the characteristic of \( \mathbb{K} \).

Example 5.11. Take as the ground field \( \mathbb{F}_2 \), the integers modulo 2. Let \( W \) denote the adjacency matrix of the complete bipartite graph \( K_{1,3} \) on vertex set \( X \). Each partite set is an orbit of \( K_{1,3} \) under the action of its automorphism group, so \( \text{dim } \mathcal{F}_1(X, \mathcal{A}(W)) = 2 \). However, \( \text{dim } \mathcal{O}_1^W = 1 \) since the symmetry of \( K_{1,3} \) implies that given an open graph \( \Delta = (V, E, b) \) of boundary size 1, \( Z^W_\Delta(p) \equiv Z^W_\Delta(q) \mod 2 \) for all vertices \( p, q \) of \( K_{1,3} \). In particular, \( \mathcal{O}_1^W \neq \mathcal{F}_1^W \) over \( \mathbb{F}_2 \). Similar arguments show that when \( W \) is the adjacency matrix of a complete multipartite graph \( K_{n_1, n_2, \ldots, n_m} \) over a field \( \mathbb{K} \) of characteristic \( k > 0 \), \( \text{dim } \mathcal{O}_1^W \) is equal to the number of congruence classes modulo \( k \) appearing among \( n_1, n_2, \ldots, n_m \) while \( \text{dim } \mathcal{F}_1^W \) is equal to the number of distinct numbers among \( n_1, n_2, \ldots, n_m \).

The arguments used in this paper can be extended to planar subalgebras of \( \mathcal{X} \) generated by finitely many functions \( \Omega = \{ f_i : X^{k_i} \to \mathbb{K} \} \). Here the elements of the planar algebra \( \mathcal{P}_\Omega \) are the functions defined from the partition function (1) starting from planar tangles all of whose internal disks are labeled with compatible elements of \( \Omega \). (See [15] for more on labeled planar tangles.) The planar algebra \( \mathcal{O}_\Omega \) can be defined using “open hypergraphs” in a fashion similar to the definition of \( \mathcal{O}^W \) above. Then \( \mathcal{P}_\Omega = \mathcal{O}_\Omega \) if and only if \( \Phi^\Omega \in \mathcal{P}_\Omega^\Omega \). Moreover, \( \mathcal{O}_\Omega = \mathcal{F}(\mathcal{A}(\Omega), X) \) as long as the characteristic of the ground field does not divide the order of \( \mathcal{A}(\Omega) \), where \( \mathcal{A}(\Omega) = \{ s \in \mathcal{S}_X | sf_i = f_i \text{ for all } f_i \in \Omega \} \).
References


Received February 5, 2001 and revised June 27, 2002. Research conducted while the author was an NSF mathematical sciences postdoctoral research fellow at UC-Berkeley.

Department of Mathematics
University of South Florida
4202 E. Fowler Avenue, PHY114
Tampa, FL 33620
E-mail address: bcurtin@math.usf.edu
COMPLETE CONTRACTIVITY OF MAPS ASSOCIATED WITH THE ALUTHGE AND DUGGAL TRANSFORMS

Ciprian Foiaş, Il Bong Jung, Eungil Ko, and Carl Pearcy

For an arbitrary operator $T$ on Hilbert space, we study the maps $\tilde{\Phi} : f(T) \rightarrow f(\tilde{T})$ and $\hat{\Phi} : f(T) \rightarrow f(\hat{T})$, where $\tilde{T}$ and $\hat{T}$ are the Aluthge and Duggal transforms of $T$, respectively, and $f$ belongs to the algebra Hol($\sigma(T)$). We show that both maps are (contractive and) completely contractive algebra homomorphisms. As applications we obtain that every spectral set for $T$ is also a spectral set for $\hat{T}$ and $\tilde{T}$, and also the inclusion $W(f(T))^- \cup W(f(T))^- \subset W(f(T))^-$ relating the numerical ranges of $f(T)$, $f(\tilde{T})$, and $f(\hat{T})$.

1. Introduction.

Let $\mathcal{H}$ be an arbitrary separable, complex Hilbert space whose dimension satisfies $2 \leq \dim \mathcal{H} \leq \aleph_0$, and denote by $\mathcal{L}(\mathcal{H})$ the algebra of all bounded linear operators on $\mathcal{H}$. If $T \in \mathcal{L}(\mathcal{H})$ we shall always write, without further mention, $T = UP$ to be the unique polar decomposition of $T$ (so $P = |T| = (T^*T)^{1/2}$ and $U$ is the appropriate partial isometry satisfying $\ker U = \ker T$ and $\ker U^* = \ker T^*$). Also we write, as usual, $\sigma(T)$ for the spectrum of such a $T$.

In this paper we consider the following two transforms of an arbitrary $T = UP$ in $\mathcal{L}(\mathcal{H})$:

(a) the Aluthge transform $\tilde{T} := P^{1/2}UP^{1/2}$, which was first studied in [1] and which has been studied extensively since, mostly in the context of $p$-hyponormal operators. In particular, some of the present authors studied the map $T \rightarrow \tilde{T}$ for an arbitrary $T$ in $\mathcal{L}(\mathcal{H})$ in [4], [5] and [6]. We obtained in [4] various spectral identities and showed that if $T$ is a quasiaffinity, then the invariant subspace lattice $\text{Lat}(T)$ is nontrivial if and only if $\text{Lat}(\tilde{T})$ is nontrivial, and the same is true of the hyperinvariant subspace lattices $\text{HLat}(T)$ and $\text{HLat}(\tilde{T})$. Furthermore, we showed that the map $T \rightarrow \tilde{T}$ is ($\| \cdot \|$, $\| \cdot \|$) continuous at every $T$ in $\mathcal{L}(\mathcal{H})$ with closed range, and we conjectured that for an arbitrary $T$ in $\mathcal{L}(\mathcal{H})$, where $\mathcal{H}$ is finite dimensional, the sequence $\{\tilde{T}^{(n)}\}$ of Aluthge iterates of $T$, defined by $\tilde{T}^{(0)} = T$ and $\tilde{T}^{(n+1)} = (\tilde{T}^{(n)})^-$ for $n \in \mathbb{N}$, converges to a normal operator. Our study was continued in [5], in which we showed that if $T$ is an arbitrary operator...
in \( L(H) \) such that the spectral picture \( SP(T) \) of \( T \) (or that of \( \tilde{T} \); cf. [9]) contains no pseudoholes, then \( SP(T) = SP(\tilde{T}) \), and we derived connections between \( T \) and \( \tilde{T} \) as consequences of this equality (e.g., \( T \) is quasitriangular if and only if \( \tilde{T} \) is quasitriangular).

Moreover, in [6] we pursued the study of the sequence \( \{\tilde{T}^{(n)}\} \) of Aluthge iterates of an arbitrary \( T \) in \( L(H) \), and we established the validity of [4, Conjecture 1.11] in certain special cases. We also initiated a study of the backward Aluthge iterates of an arbitrary \( T \) in \( L(H) \).

(b) The Duggal transform \( \hat{T} := PU \) (named after Professor B. P. Duggal, who suggested its study to us), has been studied very little.

We will explore below various relations between \( T, \hat{T}, \) and \( \tilde{T} \) by studying maps between the Riesz-Dunford algebras associated with these operators.

It is well-known (and not difficult to see, cf. [4]) that
\[
\sigma(T) = \sigma(\hat{T}) = \sigma(\tilde{T}), \quad T \in L(H).
\]

In what follows, when some \( T \) in \( L(H) \) is under consideration, we denote by \( \text{Hol}(\sigma(T)) \) the algebra of all complex-valued functions which are analytic on some neighborhood of \( \sigma(T) \), where linear combinations and products in \( \text{Hol}(\sigma(T)) \) are defined (with varying domains) in the obvious way. Moreover, the (Riesz-Dunford) algebra \( A_T \subset L(H) \) is defined as
\[
A_T = \{ f(T) : f \in \text{Hol}(\sigma(T)) \},
\]
(where \( f(T) \) is defined by the Riesz-Dunford functional calculus). As our main theorem (Th. 1.1) shows, it is possible to obtain useful information about \( \tilde{T} \) and \( \hat{T} \) by studying maps between the algebras \( A_T, A_{\hat{T}}, \) and \( A_{\tilde{T}} \).

**Theorem 1.1.** For every \( T \) in \( L(H) \), with \( \hat{T}, \tilde{T}, \) and \( \text{Hol}(\sigma(T)) \) as defined above:

a) The maps \( \Phi : A_T \rightarrow A_{\hat{T}} \) and \( \Phi : A_T \rightarrow A_{\tilde{T}} \) defined by
\[
\Phi(f(T)) = f(\hat{T}), \quad \Phi(f(T)) = f(\tilde{T}), \quad f \in \text{Hol}(\sigma(T)),
\]
are well-defined contractive algebra homomorphisms; in particular,
\[
(1) \quad \max\{\|f(\hat{T})\|, \|f(\tilde{T})\|\} \leq \|f(T)\|, \quad f \in \text{Hol}(\sigma(T)).
\]

b) More generally, the maps \( \Phi \) and \( \Phi \) in a) are completely contractive, meaning that for every \( n \in \mathbb{N} \) and every \( n \times n \) matrix \( (f_{ij}) \) with entries from \( \text{Hol}(\sigma(T)) \),
\[
\max\{\|(f_{ij}(\hat{T}))\|, \|(f_{ij}(\tilde{T}))\|\} \leq \|(f_{ij}(T))\|.
\]

c) Every spectral set \( [M\text{-spectral set (for fixed } M > 1\text{)}] \) for \( T \) is also a spectral set \( [\text{respectively, } M\text{-spectral set}] \) for both \( \hat{T} \) and \( \tilde{T} \).
d) If $W(S)$ denotes the numerical range of an operator $S$ in $L(H)$, then
$$W(f(\hat{T}))^- \cup W(f(\hat{T}))^- \subset W(f(T))^-,$$  
$f \in \text{Hol}(\sigma(T))$,

and, moreover, if $T$ belongs to some class $C_\rho$, then $\hat{T}$ and $\tilde{T}$ belong to $C_\rho$ also (see [7, p. 45] for the definition of these classes).

The result d) verifies (except for the closure bar) an earlier conjecture of the authors [4, Conjecture 1.9] and extends recent work of T. Yamazaki [10], who showed that $W(\tilde{T}) \subset W(T)$ if $T$ acts on a finite dimensional space and that $w(\tilde{T}) \subset w(T)$ in complete generality, where, of course, $w(T)$ denotes the numerical radius of $T$.

The proof of a) of Theorem 1.1 requires some lemmas and will be given in Section 2. On the other hand, c) follows immediately from a) and the definitions of spectral and $M$-spectral sets, so no proof of c) need be given. The result d) is also an easy consequence of c), but a proof will be given in Section 3. Finally, b) will be established in Sections 4 and 5.

2. Proof of Theorem 1.1 a).

It is obvious that the maps $\hat{\Phi}$ and $\tilde{\Phi}$ are algebra homomorphisms provided that they are well-defined, and this will follow from the inequalities (1). Thus it suffices to establish (1). As noted above, the proof depends upon several lemmas. The first of these summarizes some easy calculations, so no proof need be given.

Lemma 2.1. For every $T = UP$ in $L(H)$, we have

a) $PT = \hat{T}P$,

b) $TU = U\hat{T}$,

c) $P\frac{1}{2}T = \hat{T}P\frac{1}{2}$, and

d) $P\frac{1}{2}\hat{T} = \hat{T}P\frac{1}{2}$.

Lemma 2.2. For every $T = UP$ in $L(H)$ and every $f \in \text{Hol}(\sigma(T))$, we have

a) $Pf(T) = f(\hat{T})P$,

b) $f(T)U = Uf(\hat{T})$,

c) $P\frac{1}{2}f(T) = f(\hat{T})P\frac{1}{2}$, and

d) $P\frac{1}{2}f(\hat{T}) = f(T)P\frac{1}{2}$.

Proof. If $f$ is a polynomial, the desired relations follow from Lemma 2.1 by trivial calculations. Next suppose that $f = p/q$ is a rational function, where $p$ and $q$ are polynomials such that $q$ doesn’t vanish on $\sigma(T)$. Then $q(T)$ and $q(\hat{T})$ are invertible (for example) and the equation $Pq(T) = q(\hat{T})P$ yields immediately $Pq(T)^{-1} = q(\hat{T})^{-1}P$ (for example), so the desired relations are valid for all rational functions in $\text{Hol}(\sigma(T))$. The lemma now results
easily from Runge’s theorem and the well-known continuity properties of the Riesz-Dunford functional calculus (cf., e.g., [2, Prop. 17.26]). □

Lemma 2.3. For every $T = UP$ in $\mathcal{L}(\mathcal{H})$ and every $f \in \text{Hol}(\sigma(T))$, $f(\hat{T})$ is the (orthogonal) direct sum

\begin{equation}
(2) \quad f(\hat{T}) = EU^*f(T)UE|_{\ker(T)}^\perp \oplus f(0)1_{\ker(T)},
\end{equation}

where $E$ is the (orthogonal) projection $U^*U$ on $(\ker(T))^\perp$, and, consequently,

\begin{equation}
(3) \quad \|f(\hat{T})\| \leq \|f(T)\|.
\end{equation}

Proof. If $T$ has trivial kernel, then $U$ is an isometry, and thus $E = 1_\mathcal{H}$ and $f(\hat{T}) = U^*f(T)U$, so (2) and (3) are satisfied. Thus we may suppose that $0 \in \sigma(T)$, and hence $f$ is analytic at $z = 0$ and $f(0) \in \sigma(f(T))$. An easy calculation shows that $E\hat{T} = \hat{T}E = \hat{T}$, and thus (by writing $f(z) = f(0) + zg(z)$, where $g \in \text{Hol}(\sigma(T)))$ that $f(\hat{T})E = Ef(\hat{T})$. Hence

\[
\begin{align*}
    f(\hat{T}) &= EF(\hat{T})E|_{\ker(T)}^\perp \oplus f(0)1_{\ker(T)} \\
    &= EU^*f(T)UE|_{\ker(T)}^\perp \oplus f(0)1_{\ker(T)},
\end{align*}
\]

from b) of Lemma 2.2, and thus

\[
\|f(\hat{T})\| \leq \max\{\|f(T)\|, |f(0)|\} = \|f(T)\|.
\]

Lemma 2.4. For every $T = UP$ in $\mathcal{L}(\mathcal{H})$ such that $P$ has trivial kernel (which implies, of course, that $U$ is an isometry) and for every $f \in \text{Hol}(\sigma(T))$, $\|f(\hat{T})\| \leq \|f(T)\|$. 

Proof. Suppose first that $P$ is invertible. We use the fact from [3] that if $X \in \mathcal{L}(\mathcal{H})$ and $A$ and $B$ are positive semidefinite operators in $\mathcal{L}(\mathcal{H})$, then

\begin{equation}
(4) \quad \|A^{\frac{1}{2}}XB^{\frac{1}{2}}\| \leq \|AXB\|^{\frac{1}{2}}\|X\|^{\frac{1}{2}}.
\end{equation}

We know from c) of Lemma 2.2 that

\[
    f(\hat{T}) = P^{\frac{1}{2}}f(T)P^{-\frac{1}{2}},
\]

so applying (4) with $A = P$ and $B = P^{-1}$ we obtain

\begin{equation}
(5) \quad \|f(\hat{T})\| = \|P^{\frac{1}{2}}f(T)P^{-\frac{1}{2}}\| \leq \|Pf(T)P^{-1}\|^{\frac{1}{2}}\|f(T)\|^{\frac{1}{2}}.
\end{equation}

Moreover, we know from a) of Lemma 2.2 that

\[
    Pf(T)P^{-1} = f(\hat{T}),
\]

and thus (5) becomes

\[
\|f(\hat{T})\| \leq \|f(\hat{T})\|^{\frac{1}{2}}\|f(T)\|^{\frac{1}{2}} \leq \|f(T)\|,
\]

by Lemma 2.3, and the case in which $P$ is invertible is done.
Now let $P$ be an arbitrary quasiaffinity. Define the sequence $\{Q_n\}$ of positive invertible operators by $Q_n = P + (1/n)1_H$, and set $A_n = UQ_n$ (polar decomposition of $A_n$). Then $A_n = (Q_n)^{1/2}U(Q_n)^{1/2}$, and since $\|Q_n - P\| \to 0$ and $\|(Q_n)^{1/2} - P^{1/2}\| \to 0$, we obtain that $\|A_n - T\| \to 0$ and $\|A_n - \tilde{T}\| \to 0$. By what was proved above, we have $\|f(A_n)\| \leq \|f(A_n)\|$ for all $n$ sufficiently large that $f(A_n)$ is defined. The result follows from the facts that $\|f(A_n) - f(T)\| \to 0$ and $\|f(A_n) - f(\tilde{T})\| \to 0$ (cf., for example, [2, Prop. 17.26]).

**Lemma 2.5.** For every $T = UP$ in $\mathcal{L}(H)$ such that

$$\dim(\ker U^+) \geq \dim(\ker U) > 0$$

and for every $f \in \Hol(\sigma(T))$, $\|f(\tilde{T})\| \leq \|f(T)\|$.

**Proof.** Choose a partial isometry $V$ such that the initial space of $V$ is $\ker U$ and the range of $V$ is a subspace of $\ker U^*$. Define $A_n = T + (1/n)V$ for $n \in \mathbb{N}$, and note that each $|A_n|$ is a quasiaffinity and that $\|A_n - T\| \to 0$. Since the polar decomposition of $A_n$ is $(U + V)|A_n|$, where $|A_n|$ is the direct sum $P|\ker T|^\perp \oplus (1/n)1_{\ker T}$, it follows easily that $\|A_n - \tilde{T}\| \to 0$. From Lemma 2.4 we know that $\|f(A_n)\| \leq \|f(A_n)\|$ and the result now follows as before from, e.g., [2, Prop. 17.26].

To complete the proof of a) of Theorem 1.1, it suffices, in view of Lemma 2.5, to deal with the case in which $T = UP$ and $U$ satisfies $\dim(\ker U^*) < \dim(\ker U)$. Moreover, if $\ker U^*$ is nontrivial, by choosing a partial isometry $W$ whose range is $\ker U^*$ and whose initial space is a subspace of $\ker U$, and considering the sequence $\{T + (1/n)W\}$ as in Lemma 2.5, we can reduce what is to be shown to the case in which $U$ is a nonunitary coisometry.

**Lemma 2.6.** For every $T = UP$ in $\mathcal{L}(H)$ such that $U$ is a nonunitary coisometry and for every $f \in \Hol(\sigma(T))$,

$$\|f(\tilde{T})\| \leq \|f(\tilde{T})\| \leq \|f(T)\|.$$  

**Proof.** Let $U^* := \{z : z \in U\}$, and let $\tilde{f}$ be the analytic function on $U^*$ defined, as usual, by $\tilde{f}(z) := \tilde{f}(\bar{z})$, $z \in U^*$. Recall that in this situation, $\sigma(T^*) \subset U^*$ and $f(T^*) = \tilde{f}(T^*)$, so $\|\tilde{f}(T^*)\| = \|f(T)\|$. Note that $\tilde{T} = PU$, and thus that $(\tilde{T})^* = U^*P$ with $U^*$ an isometry. Define, for $n \in \mathbb{N}$,

$$S_n = U^*(P + (1/n)1_H).$$

Since $P + (1/n)1_H$ is invertible, (6) gives the polar decomposition of $S_n$, and hence

$$S_n = (P + (1/n)1_H)^{1/2}U^*(P + (1/n)1_H)^{1/2}, \quad n \in \mathbb{N}.$$
It follows easily that \( \| S_n - (\hat{T})^* \| \to 0 \) and that \( \| \tilde{S}_n - (\hat{T})^* \| \to 0 \). Thus we have
\[
\| f(\tilde{T}) \| = \| \tilde{f}((\tilde{T})^*) \| = \lim_n \| \tilde{f}(\tilde{S}_n) \| ,
\]
and
\[
\| f(\hat{T}) \| = \| \hat{f}((\hat{T})^*) \| = \lim_n \| \hat{f}(S_n) \|
\]
(again by, e.g., [2, Prop. 17.26]). But Lemma 2.4 applies to each \( S_n \), and thus
\[
\| \tilde{f}(\tilde{S}_n) \| \leq \| \tilde{f}(S_n) \| , \quad n \in \mathbb{N}.
\]
Thus \( \| f(\tilde{T}) \| \leq \| f(\hat{T}) \| \) and the other inequality follows from Lemma 2.3. This completes the proof of Theorem 1.1 a). \( \square \)

3. Proof of Theorem 1.1 d).

In view of Theorem 1.1 c), which follows immediately from Theorem 1.1 a) as noted above, the first statement in d) follows trivially from the following known fact, and the other statements are immediate from Remarks 1, 2 and 3 on pp. 48 and 49 of [7]:

**Proposition 3.1.** For every \( T \in L(H) \), \( W(T)^- \) is the intersection of all closed half-planes \( H \) containing \( W(T) \) such that \( H \) is a spectral set for \( T \).

**Proof.** Since \( W(T)^- \) is convex, and is thus the intersection of all closed half-planes containing \( W(T) \), it suffices to show that if \( H \) is any closed halfplane containing \( W(T) \), then \( H \) is a spectral set for \( T \). By a harmless rotation and translation, we may suppose that \( H \) is the closed right-halfplane \( \{ z : \text{Re } z \geq 0 \} \). Thus, writing \( T = K + iL \), with \( K \) and \( L \) Hermitian, we see that \( K \) is positive semidefinite, and therefore that the Cayley transform of \( T \),
\[
c(T) = (T + 1_H)^{-1}(T - 1_H),
\]
is a contraction (cf., e.g., [7, p. 167]). Hence, by von Neuman’s inequality, the closed unit disc \( \mathbb{D} \) in \( \mathbb{C} \) is a spectral set for \( c(T) \), and thus, by taking inverse Cayley transforms, we obtain that \( H \) is a spectral set for \( T \), as desired. \( \square \)

4. Complete contractivity of \( \hat{\Phi} \).

In this section we prove the following theorem, which establishes a part of Theorem 1.1 b):

**Theorem 4.1.** For every \( T \) in \( L(H) \), the map \( \hat{\Phi} : A_T \to A_{\hat{T}} \) defined in Section 1 is completely contractive.
Recall that this means, by definition, that for every \( n \in \mathbb{N} \) and for every \( n \times n \) matrix \( (f_{ij}) \), where each \( f_{ij} \in \text{Hol}(\sigma(T)) \), the inequality
\[
\| (f_{ij}(T)) \| \leq \| (f_{ij}(T)) \|
\]
is satisfied. (Here of course, the \( n \times n \) operator matrices in (8) act on the Hilbert space \( \mathcal{H}^{(n)} \), the direct sum of \( n \) copies of \( \mathcal{H} \), and the norm indicated is the operator norm on \( \mathcal{L}(\mathcal{H}^{(n)}) \).

**Proof of Theorem 4.1.** Let \( T \in \mathcal{L}(\mathcal{H}) \), let \( n \in \mathbb{N} \), and let \( (f_{ij}) \) be an arbitrary \( n \times n \) matrix with entries from \( \text{Hol}(\sigma(T)) \). Then, with the notation as in Lemma 2.3, it is immediate from (2) that we have the matricial identity
\[
(f_{ij}(T)) = (EU^*f_{ij}(T)UE|_{\ker T})^\perp \oplus (f_{ij}(0)1_{\ker T}),
\]
where, of course, the first [second] matrix on the right acts on the space \( \{\ker T\}^{(n)} \) (respectively, \( \{\ker T\}^{(n)} \)). As in the proof of Lemma 2.3, if \( T \) has trivial kernel, then \( E = 1_{\mathcal{H}} \) and \( U \) is an isometry. Since it is obvious that the inequality
\[
\| (U^*f_{ij}(T)U) \| \leq \| (f_{ij}(T)) \|
\]
holds (the matrix on the left is the product of two diagonal matrices of norm at most one and the matrix on the right), it suffices to treat the case in which \( \ker T \neq (0) \). Moreover, from (9), one sees easily that it is enough to show that
\[
\| (f_{ij}(0)1_{\ker T}) \| \leq \| (f_{ij}(T)) \|.
\]
Since \( f_{ij}(0) \in \sigma(f_{ij}(T)) \), \( f_{ij} \) is analytic at \( z = 0 \) for \( i, j = 1, \ldots, n \). Upon writing
\[
f_{ij}(z) - f_{ij}(0) = g_{ij}(z)z, \quad z \in \text{domain } f_{ij},
\]
we see that \( g_{ij}(z) \in \text{Hol}(\sigma(T)) \) for \( i, j = 1, \ldots, n \), and hence we get the matricial identity
\[
(f_{ij}(T) - f_{ij}(0)1_{\mathcal{H}}) = (g_{ij}(T)T) = (g_{ij}(T)) \text{ Diag } (T, \ldots, T).
\]

Observe next that the matrix \( (f_{ij}(0)1_{\ker T}) \) has the same norm as the matrix \( M = (f_{ij}(0)) \) acting on \( \mathbb{C}^n \). Moreover, there exists a unit vector \( w = (\xi_1, \ldots, \xi_n)^t \in \mathbb{C}^n \) such that \( \| Mw \| = \| M \| \). Now let \( x \) be a unit vector in \( \ker(T) \) and note that if \( s \) is the unit vector
\[
s = (\xi_1x, \ldots, \xi_nx)^t \in \mathcal{H}^{(n)},
\]
then, from (12), we have \( (f_{ij}(T))s = (f_{ij}(0)1_{\mathcal{H}})s \). Write \( Mw = (\gamma_1, \ldots, \gamma_n)^t \), and observe that
\[
\| M \| = \| Mw \| = \| (\gamma_1, \ldots, \gamma_n)^t \| = \| (\gamma_1x, \ldots, \gamma_nx)^t \| = \| (f_{ij}(0)1_{\mathcal{H}})s \| = \| (f_{ij}(T))s \| \leq \| (f_{ij}(T)) \|,
\]
which is the desired inequality. 

5. Complete contractivity of $\tilde{\Phi}$.

In this section we prove the following analog of Theorem 4.1 for the mapping $\tilde{\Phi}$, and thus complete the proof of Theorem 1.1 b).

**Theorem 5.1.** For every $T$ in $\mathcal{L}(\mathcal{H})$, the map $\tilde{\Phi} : \mathcal{A}_T \to \mathcal{A}_{\tilde{T}}$ defined in Section 1 is completely contractive.

Let $n \in \mathbb{N}$ and let $(f_{ij})$ be an arbitrary $n \times n$ matrix with entries from Hol($\sigma(T)$). As noted above, we must show that

\[(13) \quad \| (f_{ij}(\tilde{T})) \| \leq \| (f_{ij}(T)) \|.
\]

To establish (13) we need some lemmas. The following lemma will simplify greatly the remainder of the argument:

**Lemma 5.2.** Suppose $n \in \mathbb{N}$ and $(f_{ij})$ is an $n \times n$ matrix with entries from Hol($\sigma(T)$). Let $T \in \mathcal{L}(\mathcal{H})$ and suppose that there exists a sequence $\{A_n\}$ in $\mathcal{L}(\mathcal{H})$ such that:

a) $\| A_n - T \| \to 0$,  

b) $\| \tilde{A}_n - \tilde{T} \| \to 0$, and  

c) $\| (f_{ij}(A_n)) \| \leq \| (f_{ij}(A_n)) \|$ for all $n$ sufficiently large.

Then (13) is satisfied.

**Proof.** By the upper semicontinuity of the spectrum,

$\sigma(A_n) \subset \bigcap_{i,j=1}^{n} (\text{domain } f_{ij})$

for $n$ sufficiently large, so $f_{ij}(A_n)$ and $f_{ij}(\tilde{A}_n)$ are defined for such $n$. Moreover, as noted several times above,

$\| f_{ij}(A_n) - f_{ij}(T) \| \to 0, \quad \| f_{ij}(\tilde{A}_n) - f_{ij}(\tilde{T}) \| \to 0, \quad i, j = 1, \ldots, n.$

Since there are only a finite number of functions $f_{ij}$, it follows easily that

$\| (f_{ij}(A_n)) - (f_{ij}(T)) \| \to 0, \quad \| (f_{ij}(\tilde{A}_n)) - (f_{ij}(\tilde{T})) \| \to 0,$

and these facts, together with c) above, yield the result. 

**Lemma 5.3.** With the notation as above, if $T = UP$ and $P$ has trivial kernel, then (13) holds.

**Proof.** Suppose first that $P$ is invertible. By c) of Lemma 2.2,

$$(f_{ij}(\tilde{T})) = \left(P^{\frac{1}{2}} f_{ij}(T) P^{-\frac{1}{2}} \right)$$

$$= \text{Diag} \left( P^{\frac{1}{2}}, \ldots, P^{\frac{1}{2}} \right)(f_{ij}(T)) \text{Diag} \left( P^{-\frac{1}{2}}, \ldots, P^{-\frac{1}{2}} \right).$$
Thus, utilizing (4), Lemma 2.2 a), and Theorem 4.1, we obtain
\[
\| (f_{ij}(\tilde{T})) \| \leq \| (P f_{ij}(T) P^{-1}) \| \cdot \\| (f_{ij}(T)) \|^{\frac{1}{2}}
\]
\[
= \| (f_{ij}(\tilde{T})) \|^{\frac{1}{2}} \cdot \| (f_{ij}(T)) \|^{\frac{1}{2}}
\]
\[
\leq \| (f_{ij}(T)) \|,
\]
as desired. Now let \( P \) be an arbitrary quasi-afiinity, and let the sequences \( \{Q_n\} \) and \( \{A_n\} \) be as defined in the proof of Lemma 2.4, so we have a) and b) of Lemma 5.2 satisfied. Since each \( |A_n| \) is invertible by construction, by what was just shown,
\[
\| (f_{ij}(\tilde{A}_n)) \| \leq \| (f_{ij}(A_n)) \|,
\]
so c) of Lemma 5.2 is satisfied and the result follows from that lemma. □

**Lemma 5.4.** Let \( n \in \mathbb{N} \), let \( (f_{ij}) \) be any \( n \times n \) matrix with entries from \( \text{Hol}(\sigma(T)) \) and suppose \( T = UP \) is any operator in \( \mathcal{L}(\mathcal{H}) \) such that
\[
\dim(\ker U^*) \geq \dim(\ker U) > 0.
\]
Then (13) is satisfied.

**Proof.** Let the sequence \( \{A_n\}_{n=1}^{\infty} \) be as defined in Lemma 2.5, and observe that from the proof of that lemma, we know that a) and b) of Lemma 5.2 are satisfied. Moreover, since each \( |A_n| \) is a quasi-affinity, Lemma 5.3 yields that c) of Lemma 5.2 is satisfied, and the result follows from Lemma 5.2. □

In view of the discussion preceding Lemma 2.6, the proof of Theorem 5.1 (and thus the proof of Theorem 1.1 b) is completed by the following:

**Lemma 5.5.** For every \( T = UP \) in \( \mathcal{L}(\mathcal{H}) \) such that \( U \) is a nonunitary coisometry, for every \( n \in \mathbb{N} \), and for every \( n \times n \) matrix \( (f_{ij}) \) with entries from \( \text{Hol}(\sigma(T)) \), (13) is satisfied.

**Proof.** Let the sequence \( \{S_n\}_{n=1}^{\infty} \) be as defined in the proof of Lemma 2.6, and observe from that proof that \( \| S_n - (\tilde{T})^* \| \to 0 \) and \( \| \tilde{S}_n - (\tilde{T})^* \| \to 0 \). Moreover, since \( |S_n| \) is an isometry for each \( n \), Lemma 5.4 applies to give that
\[
\| (\tilde{f}_{ij}(\tilde{S}_n)) \| \leq \| (\tilde{f}_{ij}(S_n)) \|
\]
so that a), b) and c) of Lemma 5.2 are satisfied (with \( S_n \to A_n \) (i.e., \( S_n \) replaces \( A_n \)), \( \tilde{T}^* \to T, \tilde{T}^* \to \tilde{T}, \) and \( \tilde{f}_{ij} \to f_{ij} \)), so
\[
\| (\tilde{f}_{ij}(\tilde{T}^*)) \| \leq \| (\tilde{f}_{ij}(\tilde{T}^*)) \|.
\]
Upon taking adjoints in (14) and using Theorem 4.1, the result follows. □
Of course, one reason for establishing that the maps $\widetilde{\Phi}$ and $\hat{\Phi}$ are completely contractive is that the extension theorems of Arveson and Stinespring can be applied to obtain the structure of such maps (cf., e.g., [8]), and thus we get the following:

**Theorem 5.6.** Let $T$ be an arbitrary operator in $L(H)$, and let $\widetilde{\Phi}$ and $\hat{\Phi}$ be the maps defined in Theorem 1.1. Then there exist Hilbert spaces $\widetilde{K} = \widetilde{K}_T$ and $\hat{K} = \hat{K}_T$ containing $H$, and $C^*$-homomorphisms $\widetilde{\Psi} : C^*(T) \to L(\widetilde{K})$ and $\hat{\Psi} : C^*(T) \to L(\hat{K})$ (where $C^*(T)$ is the smallest unital $C^*$-algebra containing $A_T$) such that for every $f$ in $\text{Hol}(\sigma(T))$,

$$\widetilde{\Phi}(f(T)) = P_H \widetilde{\Psi}(f(T))|_H$$

and

$$\hat{\Phi}(f(T)) = P^{(2)}_H \hat{\Psi}(f(T))|_H,$$

where $P^{(1)}_H$ and $P^{(2)}_H$ are the orthogonal projections of $\widetilde{K}$ and $\hat{K}$, respectively, onto $H$.

The implications of Theorem 5.6 for the Aluthge and Duggal transforms will be the subject of a forthcoming paper by the authors.

**Acknowledgement.** The second and third authors were supported by KOSEF Research Project No. R01-2000-00003. The fourth author acknowledges the support of the National Science Foundation.

**References**


Received April 4, 2002.

Department of Mathematics
Texas A&M University
College Station, TX 77843
E-mail address: foias@math.tamu.edu

Department of Mathematics
Kyungpook National University
Taegu 702-701
Korea
E-mail address: ibjung@kyungpook.ac.kr

Department of Mathematics
Ewha Women’s University
Seoul 120-750
Korea
E-mail address: eiko@mm.ewha.ac.kr

Department of Mathematics
Texas A&M University
College Station, TX 77843
E-mail address: pearcy@math.tamu.edu
SOMMES DE MODULES DE SOMMES D’EXPONENTIELLES

ETIENNE FOUVRY AND PHILIPPE MICHEL

Let $Kl(a,b;n)$ be the usual Kloosterman sum modulo $n$, with coefficients $a$ and $b$. We give upper and lower bounds for the sum $\sum_{n \leq x} |Kl(1,1;n)|/\sqrt{n}$, and for related sums, by using large sieve techniques and Deligne-Katz theory of exponential sums. Extensions to more general exponential sums of dimension one are also given.

1. Introduction.

Soient $a$ et $b$ des entiers et $n$ un entier $\geq 1$. On rappelle que la somme de Kloosterman $Kl(a,b;n)$, de dénominateur $n$ et de coefficients $a$ et $b$, est définie par

$$Kl(a,b;n) = \sum_{(m,n)=1} \exp\left(2\pi i \frac{am + bm}{n}\right).$$

Systématiquement, le symbole $\overline{m}$ dans la fraction $\frac{\overline{m}}{n}$ désigne l’inverse multiplicatif de $m$ modulo $n$, $\omega(n)$ est le nombre de facteurs premiers distincts de l’entier $n$, et on réserve la lettre $p$ aux nombres premiers. Les sommes de Kloosterman jouent un rôle crucial dans l’actuelle théorie analytique des nombres, au confluent de la géométrie algébrique et de la théorie des formes modulaires. Rappelons que ce sont des nombres réels non nuls qui vérifient, entre autres, les propriétés suivantes:

- **Multiplicativité croisée**: Pour $(n_1,n_2) = 1$, on a la relation

$$Kl(a,b;n_1n_2) = Kl(a,\overline{m_2};n_1)Kl(a,\overline{m_1};n_2).$$

- **Majoration individuelle**:

$$|Kl(a,b;n)| \leq (a,b,n)^{\frac{1}{2}}2^{\omega(n)}n^{\frac{1}{2}},$$

conséquence de la multiplicativité croisée, de la démonstration par Weil, de l’hypothèse de Riemann pour les courbes sur les corps finis (qui donne (1.1) lorsque $n = p$) et de l’étude, due à divers auteurs, des sommes de Kloosterman de dénominateur $p^k$ ($k \geq 2$).
L’objet de cet article est de s’intéresser, lorsque $a$ et $b$ sont fixés (disons $a = b = 1$, pour fixer les idées), à l’optimalité de l’inégalité (1.1), lorsque $n$ parcourt l’ensemble des entiers. Dans ce but on introduit les quantités

$$A^*(x) = \sum_{n \leq x} \left| \frac{\text{Kl}(1, 1; n)}{2^{\omega(n)\sqrt{n}}} \right|,$$

et

$$\tilde{A}(x) = \sum_{n \leq x} \left| \frac{\text{Kl}(1, 1; n)}{\sqrt{n}} \right|,$$

qui vérifient donc les inégalités triviales

(1.2) \hspace{1cm} A^*(x) \leq x \quad (x \geq 1),

et

(1.3) \hspace{1cm} \tilde{A}(x) \leq \sum_{n \leq x} 2^{\omega(n)} \leq \left( \frac{6}{\pi^2} + o(1) \right) x \log x \quad (x \to \infty).

Les parties droites des inégalités (1.2) et (1.3) peuvent être améliorées d’une constante multiplicative, en injectant des majorations plus précises que (1.1) dans le cas où $n = p^k$ ($k \geq 2$), mais nous sommes concernés par des gains plus substantiels, puisque nous montrerons les encadrements:

**Théorème 1.1.** Il existe une constante absolue $c_1^*$ et, pour tout $k$, une constante $c_0^*(k) > 0$, telles que, pour $x \geq 3$, on ait les inégalités

(1.4) \hspace{1cm} c_0^*(k) \frac{x}{\log x} (\log \log x)^k \leq A^*(x) \leq c_1^* x \left( \frac{\log \log x}{\log x} \right)^{1 - \frac{4}{\pi^2}}.

**Théorème 1.2.** Il existe une constante absolue $\tilde{c}_1$ et, pour tout $k$, une constante $\tilde{c}_0(k) > 0$, telles que, pour $x \geq 3$, on ait les inégalités

(1.5) \hspace{1cm} \tilde{c}_0(k) \frac{x}{\log x} (\log \log x)^k \leq \tilde{A}(x) \leq \tilde{c}_1 x (\log \log x) \left( \frac{\log \log x}{\log x} \right)^{1 - \frac{8}{3\pi^2}}.

Ainsi, par rapport aux majorations triviales (1.2) et (1.3), on gagne respectivement $(\log x)^{-0.575587\ldots}$ et $(\log x)^{-1.151174\ldots}$. À notre connaissance, la première majoration non triviale de $A^*(x)$ ou de $\tilde{A}(x)$ est due à Hooley ([Ho] Theorem 3), où il montre, pour $u$ et $v$ entiers la relation

(1.6) \hspace{1cm} \sum_{n \leq x} \text{Kl}(u, v; n) = O \left( x^3 \left( \sum_{d|v} d^{-\frac{1}{2}} \right) (\log x)^{\sqrt{2}-1}(\log \log x)^c_2 \right),

où $c_2$ est une certaine constante. La preuve par Hooley de (1.6) donne en fait une majoration de $\sum_{n \leq x} |\text{Kl}(u, v; n)|$, qui conduit donc, dans notre cas, à

(1.7) \hspace{1cm} \tilde{A}(x) \leq c_3 x (\log x)^{\sqrt{2}-1}(\log \log x)^{c_2},
pour un certain $c_3 > 0$, c’est-à-dire un gain de $(\log x)^{-0.585786\ldots}$, par rapport à la majoration triviale (1.3).

La minoration (1.4) de $A^*(x)$ améliore notablement celle en

\begin{equation}
\widetilde{A}(x) \geq c_4 x / \log^2 x,
\end{equation}

conséquence directe de ([Mi1] Théorème 1) où il est montré, que pour $x \to \infty$, on a

\[ \# \{(p_1, p_2) : x < p_1, p_2 < 2x, |\text{Kl}(1, 1; p_1 p_2)| \geq 0, 64 \sqrt{p_1 p_2} \} \gg \frac{x^2}{\log^2 x}. \]

Autant on devine une différence de comportement à l’infini des quantités $A^*(x)$ et $\widetilde{A}(x)$, dans les majorations (1.4) et (1.5), autant notre preuve ne permet guère de différencier les minorations de ces fonctions, si ce n’est qu’au §4, nous obtiendrons les constantes $\tilde{c}_0(k) = 2^{k+3} c_0^*(k)$. Rappelons que l’on sait, grâce à la théorie des formes modulaires, qu’il y a d’énormes compensations entre les signes des sommes de Kloosterman, puisqu’on a pour tout $\epsilon > 0$, la majoration ([Ku] et [D-I])

\begin{equation}
\sum_{n \leq x} \frac{\text{Kl}(1, 1; n)}{\sqrt{n}} = O\left(x^{\frac{3}{2} + \epsilon}\right),
\end{equation}

qu’il convient de comparer avec (1.3), alors que la conjecture de Linnik-Selberg prédit même une majoration en $O(x^{\frac{3}{2} + \epsilon})$. Enfin, on ne sait pas si la majoration (1.9) continue d’être vraie si on insère au dénominateur de la partie gauche, le facteur arithmétique $2^{\omega(n)}$.

Ainsi le Théorème 1.2 répond de façon plus précise que (1.7) et (1.8), à la question de l’origine des compensations dans (1.9): de façon succinette, on peut dire que le fait que les modules $|\text{Kl}(1, 1; n)| (n \leq x)$ soient petits en moyenne ne fait gagner, par rapport à la majoration triviale

\[ \left| \sum_{n \leq x} \frac{\text{Kl}(1, 1; n)}{\sqrt{n}} \right| \leq \widetilde{A}(x) = O(x \log x) \]

qu’un certain facteur $X$, vérifiant $\log^{-2} x \ll X \ll (\log x)^{-1,151174}$. En conclusion, on peut affirmer que dans (1.9), la plus grande partie des compensations provient des changements de signe des sommes de Kloosterman.

En fait, comme nous le fit remarquer R. de la Bretèche, on peut, dans les minorations (1.4) et (1.5), donner explicitement les fonctions $c_0^*(k)$ et $\tilde{c}_0(k)$, puis prendre $k$ comme fonction de $x$. Nous donnerons, au paragraphe 5, des indications menant au:
Théorème 1.3. Il existe $\delta$, strictement supérieur à $5/12$, tel que, pour $x \geq 3$, on ait les minorations

$$A^*(x) \gg \frac{x}{\log x} \exp \left( (\log \log x)^\delta \right),$$

et

$$\tilde{A}(x) \gg \frac{x}{\log x} \exp \left( (\log \log x)^\delta \right).$$

À la différence des Théorèmes 1.1 et 1.2, la méthode menant au Théorème 1.3 n’est pas directement transposable au cas des sommes d’exponentielles plus générales (cf. infra Théorème 1.5 et §6), ce qui explique pourquoi nous avons séparé ces divers énoncés.

La minoration de $\tilde{A}(x)$ donnée au Théorème 1.3 répond aussi de façon plus précise que (1.8), à une question de Serre évoquée dans ([Sa] p. 33), sur le comportement asymptotique de $Kl(1,1;n)$:

Corollaire 1.4. Pour $n$ tendant vers l’infini, on a la relation

$$Kl(1,1;n) = \Omega \left( \frac{\exp \left( (\log \log n)^\frac{5}{12} \right)}{\log n} \right).$$

La démonstration des Théorèmes 1.1 et 1.2 repose essentiellement sur les propriétés multiplicatives statistiques des entiers, la multiplicativité croisée des sommes de Kloosterman, des majorations de crible et surtout, sur une troisième propriété de ces sommes, découverte par Katz ([Ka2] Example 13.6):

– loi de Sato-Tate verticale: Soit $\theta_{p,m}$ défini par l’égalité

$$\frac{Kl(1,m;p)}{2\sqrt{p}} = \cos \theta_{p,m} \ (0 \leq \theta_{p,m} \leq \pi),$$

alors, pour $p \to \infty$, l’ensemble d’angles $\{\theta_{p,m}; 1 \leq m \leq p - 1\}$ est équiréparti sur $[0,\pi]$, suivant la mesure de Sato-Tate $\frac{2}{\pi} \sin^2 \theta \, d\theta$, c’est-à-dire que pour tout $0 \leq \alpha < \beta \leq \pi$, on a, pour $p \to \infty$

$$\frac{1}{p - 1} \{1 \leq m \leq p - 1 ; \alpha \leq \theta_{p,m} \leq \beta\} \to \frac{2}{\pi} \int_{\alpha}^{\beta} \sin^2 t \, dt.$$

En fait nous aurons même besoin du même résultat d’équirépartition, mais pour les angles $\theta_{p,m^2}$ (voir Lemme 2.1).

Rappelons que la loi de Sato-Tate horizontale, dans la stricte direction de laquelle, il n’y a pour l’instant aucun résultat non trivial, prédit que, pour $x \to \infty$, l’ensemble d’angles $\{\theta_{1,p}; p \leq x\}$, est équiréparti sur $[0,\pi]$,
suivant la mesure de Sato-Tate ([Ka1] Conj. 1.2.5). L’exactitude de cette conjecture entraînerait, pour \( x \to \infty \), la relation
\[
\sum_{p \leq x} \left| \frac{\text{Kl}(1,1;p)}{2\sqrt{p}} \right| \sim \frac{4}{3\pi} \cdot \frac{x}{\log x}.
\]
La recherche d’un équivalent asymptotique des sommes \( A^*(x) \) et \( \tilde{A}(x) \) paraît ainsi comme un problème très ardu, malgré l’encadrement étroit que fournisent, pour chacune de ces sommes, les Théorèmes 1.1, 1.2 et 1.3.

On peut étendre les résultats précédents à des sommes de la forme
\[
\sum_{n \leq x} \left| \frac{\text{Kl}(1,1;n)}{2\omega(n)\sqrt{n}} \right|^\alpha,
\]
ou
\[
\sum_{n \leq x} \left| \frac{\text{Kl}(1,1;n)}{\sqrt{n}} \right|^\alpha,
\]
avec \( \alpha \) réel positif fixé, mais il est beaucoup plus intéressant d’étudier les sommes trigonométriques plus générales
\[
S_f(m;n) = \sum_{x \pmod n \atop f(x) \neq \infty} \exp \left( \frac{2\pi i \cdot m f(x)}{n} \right) = \sum_{x \pmod n \atop (Q(x),n)=1} \exp \left( \frac{2\pi i \cdot m P(x) Q(x)}{n} \right),
\]
pour:
- \( n \geq 1, m \in \mathbb{Z} \),
- \( f = \frac{P}{Q} \) fraction rationnelle, quotient de deux polynômes \( P \) et \( Q \), de \( \mathbb{Z}[X] \), premiers entre eux, à coefficients premiers entre eux.

Les sommes \( S_f(m;n) \) vérifient elles-aussi la multiplicativité croisée
\[
S_f(m; n_1 n_2) = S_f(m \overline{m_1}; n_2) S_f(m \overline{m_2}; n_1), \quad \text{pour} \ (n_1,n_2)=1,
\]
et une majoration due à Weil (cf. [D] Formule (3.5.2) p. 191):
\[
|S_f(m;p)| \leq k_f \sqrt{p} \quad (p \nmid m),
\]
où \( k_f \) étant un entier parfaitement défini en termes de la géométrie de la fraction \( f \), c’est-à-dire
\[
k_f = \max \{ \deg P, \deg Q \} + \sharp \{ \text{racines distinctes de } Q \} - 1.
\]
Le cas des \( S_f(m; p^a) \), \( a \geq 2 \) mène à des situations délicates à traiter en toute généralité, nous préférons les éviter en ne considérant que des \( n \) sans facteur carré. Après ces diverses considérations, nous étudions, pour \( x \to \infty \), les sommes
\[
A_f^*(x) := \sum_{n \leq x} \mu^2(n) \left| \frac{S_f(1;n)}{k_f^{\omega(n)} \sqrt{n}} \right|,
\]
et
\[ \tilde{A}_f(x) := \sum_{n \leq x} \mu^2(n) \left| \frac{S_f(1; n)}{\sqrt{n}} \right|, \]
dont des majorations triviales sont respectivement \( O(x) \) et \( O(x \log^{-1} k_f x) \).
Katz ([Ka3] 7.9, 7.10, 7.11; voir aussi le début du §5) a prouvé aussi une loi de Sato-Tate verticale pour la plupart des sommes \( S_f \). Ainsi, si on pose
\[ \left| \frac{S_f(m; p)}{k_f \sqrt{p}} \right| = \cos \theta_{f,p,m} \left( 0 \leq \theta_{f,p,m} \leq \frac{\pi}{2} \right), \]
Katz a montré que, pourvu que \( k_f \geq 2 \) et pourvu que \( f \) vérifie des hypothèses très générales concernant essentiellement la nature et la disposition des zéros de \( f' \), l'ensemble des angles \( \{ \theta_{f,p,m} ; 1 \leq m \leq p-1 \} \) est équiréparti sur \([0, \frac{\pi}{2}]\), lorsque \( p \rightarrow \infty \), suivant une certaine mesure que nous décrivons ci-dessous et que, par un certain abus de langage, nous appellerons aussi mesure de Sato-Tate. De façon plus précise, si \( f \) est une fraction rationnelle comme ci-dessus, on note \( Z_f' \) l'ensemble des zéros de \( f' \) dans \( \mathbb{P}^1(\mathbb{C}) \) et on pose \( C_f = f(Z_f') \). On désigne par \( H \) l'ensemble des hypothèses suivantes:

\begin{itemize}
  \item[H.1:] Les zéros de \( f' \) sont simples, (autrement dit \( \sharp Z_f' = k_f \)).
  \item[H.2:] \( f \) sépare les zéros de \( f' \) (autrement dit, pour \( z \) et \( z' \in Z_f' \), on a l'implication \( f(z) = f(z') \Rightarrow z = z' \)).
  \item[H.3:] On a l'implication
    \[ \left\{ \begin{array}{l}
      s_1, s_2, s_3, s_4 \in C_f \\
      s_1 - s_2 = s_3 - s_4
    \end{array} \right\} \Rightarrow \left\{ \begin{array}{l}
      s_1 = s_2 \text{ et } s_3 = s_4 \\
      s_1 = s_3 \text{ et } s_2 = s_4.
    \end{array} \right. \]
  \item[H.3'] \( f \) est impaire et on a l'implication
    \[ \left\{ \begin{array}{l}
      s_1, s_2, s_3, s_4 \in C_f \\
      s_1 - s_2 = s_3 - s_4
    \end{array} \right\} \Rightarrow \left\{ \begin{array}{l}
      s_1 = s_2 \text{ et } s_3 = s_4 \\
      s_1 = s_3 \text{ et } s_2 = s_4 \\
      s_1 = -s_4 \text{ et } s_2 = -s_3.
    \end{array} \right. \]
\end{itemize}
On désigne par \( \mathcal{H} \) l'ensemble des conditions H.1, H.2 et H.3 et par \( \mathcal{H}' \) l'ensemble des conditions H.1, H.2 et H.3'.
Si la fraction rationnelle \( f \) vérifie \( \mathcal{H} \), on pose
\[ G_f = SU_{k_f}(\mathbb{C}), \]
et si $f$ vérifie $\mathcal{H}'$, on pose

$$G_f = \text{USp}_{k_f}(\mathbb{C}).$$

Si $G$ désigne l’un des groupes compacts $\text{SU}_k(\mathbb{C})$ ou $\text{USp}_k(\mathbb{C})$, on note $\mu^\text{Haar}_G$, la mesure de Haar sur le groupe $G$ et $\mu_G$ l’image directe de $\mu^\text{Haar}_G$ dans $[0, \frac{\pi}{2}]$, par l’application

$$G \rightarrow [0, \frac{\pi}{2}],$$
$$A \rightarrow \text{Arc cos} \left(\frac{|\text{tr} A|}{k}\right).$$

Avec ces conventions, une des conséquences des travaux de Katz est que, si la fraction rationnelle $f$ avec $k_f \geq 2$, vérifie $\mathcal{H}$ ou $\mathcal{H}'$, l’ensemble des angles $\{\theta_{f,p,m} ; 1 \leq m \leq p - 1\}$ est équiréparti sur $[0, \frac{\pi}{2}]$, lorsque $p \rightarrow \infty$, suivant la mesure $\mu_G$.

Au paragraphe §6, nous démontrerons:

**Théorème 1.5.** Soit $f$ une fraction rationnelle comme auparavant, avec $k_f \geq 2$, vérifiant les conditions $\mathcal{H}$ ou $\mathcal{H}'$. Il existe des constantes, $c^*_6$ et $\tilde{c}_6$ et, pour tout $k$, des constantes $c^*_5(k)$ et $\tilde{c}_5(k)$ telles qu’on ait les inégalités

$$c^*_5(k) \frac{x}{\log x} (\log \log x)^k \leq A^*_f(x) \leq c^*_6 x \left(\frac{\log \log x}{\log x}\right)^{1 - \frac{1}{k_f}}$$

et

$$\tilde{c}_5(k) \frac{x}{\log x} (\log \log x)^k \leq \tilde{A}_f(x) \leq \tilde{c}_6 x (\log \log x)^{k_f - 1}.$$

On sait que, génériquement, l’hypothèse $\mathcal{H}$ est vérifiée et qu’il en est de même pour $\mathcal{H}'$ dans l’ensemble des fractions rationnelles impaires dès lors que deg $P >$ deg $Q$. Des familles explicites, vérifiant ces conditions ont été exhibées (voir [Ka3] Theorem 7.10.5 et 7.10.6, [Mi2] p. 229):

- La famille

$$f(X) = aX^{\ell+1} + bX,$$

avec $ab \neq 0$, avec $\ell$ entier impair, vérifiant $|\ell| \geq 3$ vérifie $\mathcal{H}$.

- La famille $f(X)$ avec $f$ polynôme de degré $\ell + 1 \geq 6$, tel que $f'$ soit proportionnel à un polynôme unitaire, irréductible, ayant pour groupe de Galois, le groupe de permutations $S_\ell$, vérifie $\mathcal{H}$.

- La famille

$$f(X) = aX^{\ell+1} + bX,$$

avec $ab \neq 0$, avec $\ell$ entier pair non nul, vérifie les conditions $\mathcal{H}'$.

(Signalons que pour tout $f$ appartenant à l’une des trois familles évoquées précédemment, on a $k_f = |\ell|$.) Ainsi le Théorème 1.5 montre que, à mesure que $k_f$ croît, on a un encadrement de plus en plus précis de $A^*_f(x)$. Enfin, on constate que sous les mêmes conditions, le gain par rapport à la majoration
triviale de $\tilde{A}_f(x)$ est d’autant plus important. Pour illustrer ce qui précède, nous énonçons:

**Corollaire 1.6.** Pour tout entier $\ell \geq 2$, on a la majoration

$$\sum_{n \leq x} \mu^2(n) \left| \sum_{1 \leq t \leq n} \exp \left( 2\pi i \frac{\ell t + t}{n} \right) \right| \ll_{\ell} x^{3/2}(\log \log x)^{-2/3}.$$

2. **Lemmes préparatoires.**

Dans cette partie, nous indiquons les résultats nécessaires à la preuve des Théorèmes 1.1 et 1.2, relatifs aux sommes de Kloosterman. Les généralisations, requises pour la preuve du Théorème 1.4, seront présentées au §5.

Le premier outil est issu de la géométrie algébrique, plus précisément de la théorie des sommes d’exponentielles comme l’ont développée Deligne et Katz. On a:

**Lemme 2.1.** Soit $\text{sym}_k \theta = \frac{\sin(k+1)\theta}{\sin \theta}$ la $k$-ième fonction symétrique correspondant à la mesure de Sato-Tate $\frac{2}{\pi} \sin^2 \theta \, d\theta$, associée au groupe $\text{SU}_2(\mathbb{C})$.

Il existe une constante absolue $c_7$, telle que, pour tout $k \geq 1$, tout $p$, on a l’inégalité

$$\left| \sum_{1 \leq m \leq p-1} \text{sym}_k(\theta_{p,m^2}) \right| \leq c_7 kp^{1/2}.$$ 

**Preuve.** D’après, par exemple, [Mi2] (Corollaire 2.4, avec $\psi$ le caractère trivial) on a la relation

$$\left| \sum_{1 \leq m \leq p-1} \text{sym}_k(\theta_{p,m^2}) \right| \leq 3(k+1)p^{1/2},$$

pour tout $k$ non nul. Cet énoncé reste évidemment vrai en remplaçant $\theta_{p,m^2}$ par $\theta_{p,m^2}$.

De ce lemme nous déduisons un calcul de discrépance qui évitera un facteur parasite de la forme $\log x$ à droite de (1.4) et (1.5).

**Lemme 2.2.** Soit $\phi$ une fonction paire, de période $2\pi$, de classe $C^3$, telle que, pour tout $t$ réel, on ait

$$|\phi^{(3)}(t)| \leq \lambda_3.$$ 

On a alors l’égalité

$$\frac{1}{p-1} \sum_{1 \leq m \leq p-1} \phi(\theta_{p,m^2}) = \frac{2}{\pi} \int_{0}^{\pi} \phi(\theta) \sin^2 \theta \, d\theta + O \left( \lambda_3 p^{-1/2} \right),$$

où la constante implicite dans le $O$ peut être prise absolue.
Preuve. On développe la fonction $\phi$ dans la base orthonormée $\{\text{sym}_k, k \geq 0\}$ de l'espace $L^2([0, \pi])$ muni de la mesure de Sato-Tate $\frac{2}{\pi} \sin^2 \theta \, d\theta$:

$$\phi(t) = \sum_{k=0}^{\infty} C_k \, \text{sym}_k(t),$$

avec

$$C_k = \frac{2}{\pi} \int_0^\pi \phi(t) \, \text{sym}_k(t) \, \sin^2 t \, dt = \frac{1}{\pi} \int_0^\pi \phi(t) \, k t \, dt - \frac{1}{\pi} \int_0^\pi \phi(t) \, \cos(k + 2) t \, dt.$$ Intégrant trois fois par parties, on a la relation

$$C_k = O \left( \frac{\lambda_3}{k^3} \right) \quad (k \geq 1).$$

(2.2)

Par (2.1), on a

$$\frac{1}{p-1} \sum_{1 \leq m \leq p-1} \phi(\theta_{p,m^2})$$

$$= \frac{2}{\pi} \int_0^\pi \phi(t) \, \sin^2 t \, dt + \frac{1}{p-1} \sum_{k \geq 1} C_k \sum_{1 \leq m \leq p-1} \text{sym}_k(\theta_{p,m^2})$$

$$= \frac{2}{\pi} \int_0^\pi \phi(t) \, \sin^2 t \, dt + O \left( \frac{\lambda_3 p^{-\frac{3}{2}}}{} \right),$$

d’après le Lemme 2.1 et la relation (2.2).

Une conséquence directe du Lemme 2.2 est obtenue en prenant des fonctions $\phi$ qui encadrent de mieux en mieux la fonction caractéristique d’un intervalle $[\alpha, \beta]$. On a:

Lemme 2.3. Il existe une constante absolue $c_8$, telle que pour tout $0 \leq \alpha \leq \beta \leq \pi$, tout nombre premier $p$ on ait l’inégalité

$$\left| \frac{1}{p-1} \sum_{1 \leq m \leq p-1} \phi(\theta_{p,m^2}) \right| \leq c_8 p^{-\frac{3}{8}}.$$

Preuve. Pour $I \subset \mathbb{R}$, on désigne par $1_I$ sa fonction caractéristique. Soit $\Delta$ un paramètre qui sera fixé par la suite. On suppose qu’on a les inégalités

$$0 \leq \alpha - \Delta < \alpha + \Delta < \beta - \Delta < \beta + \Delta < \pi.$$

(2.3)

On construit deux fonctions $\phi^+$ et $\phi^-$, paires, de période $2\pi$, de classe $C^3$, à supports compacts respectivement égaux (dans $[0, \pi]$) à $[\alpha - \Delta, \beta + \Delta]$ et $[\alpha, \beta]$, vérifiant les inégalités

$$1_{[\alpha + \Delta, \beta - \Delta]} \leq \phi^- \leq 1_{[\alpha, \beta]} \leq \phi^+ \leq 1_{[\alpha - \Delta, \beta + \Delta]}.$$
et vérifiant les hypothèses du Lemme 2.2 avec $\lambda_3 \ll \Delta^{-3}$. Le Lemme 2.2 entraîne l’encadrement
\[
\frac{2}{\pi} \int_0^\pi \phi^-(t) \, dt - O\left(\Delta^{-3}p^{-\frac{1}{2}}\right)
\leq \frac{1}{p-1} \sharp\{m; 1 \leq m \leq p-1, \, \alpha \leq \theta_{p,m^2} \leq \beta\}
\leq \frac{2}{\pi} \int_0^\pi \phi^+(t) \, dt + O\left(\Delta^{-3}p^{-\frac{1}{2}}\right).
\]
Puisqu’on a $\int_0^\pi (\phi^+(t) - \phi^-(t)) \, dt = O(\Delta)$, on déduit l’égalité
\[
\left|\frac{1}{p-1} \sharp\{m; 1 \leq m \leq p-1, \, \alpha \leq \theta_{p,m^2} \leq \beta\} - \frac{2}{\pi} \int_0^\beta \sin^2 t \, dt\right| = O\left(\Delta + \Delta^{-3}p^{-\frac{1}{2}}\right)
\]
d’où le lemme en posant $\Delta = p^{-\frac{1}{8}}$.

On traite de même le cas où (2.3) n’est pas vérifié. □

De la même façon, on étend le Lemme 2.2 à d’autres fonctions $\phi$, moins régulières. Nous nous contenterons de l’extension de ce lemme au cas de la fonction $\phi(t) = |\cos t|$, ce qui nous sera utile par la suite.

**Lemme 2.4.** Il existe une constante $c_9$, telle qu’on ait l’inégalité
\[
\left|\frac{1}{p-1} \sum_{1 \leq m \leq p-1} |\cos \theta_{p,m^2}| - \frac{2}{\pi} \int_0^\pi |\cos t| \sin^2 t \, dt\right| \leq c_9 p^{-\frac{1}{2}}
\]
pour tout $p$.

**Preuve.** La fonction $|\cos t|$ n’est pas dérivable au point $\frac{\pi}{2}$. On encadre cette fonction par deux fonctions plus régulières. Soit $\Delta$ un paramètre dont on fixera la valeur ultérieurement. Il existe deux fonctions $\phi^+$ et $\phi^-$, paires, de classe $C^3$, de période $2\pi$ vérifiant les propriétés suivantes
\[
\phi^-(t) \leq |\cos t| \leq \phi^+(t), \; (t \in \mathbb{R})
\]
et
\[
\int_0^\pi (\phi^+(t) - \phi^-(t)) \, dt \leq 2\Delta^2.
\]
Pour construire ces deux fonctions il suffit d’imposer que $\phi^+(t) = \phi^-(t) = |\cos t|$ si $0 \leq t \leq \frac{\pi}{2} - \Delta$ ou si $\frac{\pi}{2} - \Delta \leq t \leq \pi$ et de compléter la définition de ces fonctions en lissant la fonction $|\cos t|$ sur l’intervalle restant $[\frac{\pi}{2} - \Delta, \frac{\pi}{2} + \Delta]$. Ces fonctions $\phi^+$ et $\phi^-$ vérifient les conditions du Lemme 2.2 avec $\lambda_3 \leq$
10\Delta^{-2}, d'où l'encadrement
\[
\frac{2}{\pi} \int_0^{\pi} \phi^-(t) \sin^2 t \, dt - O \left( \Delta^{-2} p^{-\frac{1}{2}} \right)
\leq \frac{1}{p-1} \sum_{1 \leq m \leq p-1} | \cos \theta_{p,m^2} |
\leq \frac{2}{\pi} \int_0^{\pi} \phi^+(t) \sin^2 t \, dt + O \left( \Delta^{-2} p^{-\frac{1}{2}} \right),
\]
qui mène à l'égalité
\[
\left| \frac{1}{p-1} \sum_{1 \leq m \leq p-1} | \cos \theta_{p,m^2} | - \frac{2}{\pi} \int_0^{\pi} | \cos t | \sin^2 t \, dt \right| = O \left( \Delta^2 + \Delta^{-2} p^{-\frac{1}{2}} \right).
\]
Pour compléter la preuve, il suffit de prendre \( \Delta = p^{-\frac{1}{8}}. \)

Un autre ingrédient important de la preuve est l'inégalité de grand crible sous la forme suivante du théorème de Barban-Davenport-Halberstam:

Lemme 2.5. Il existe deux constantes absolues \( c_{10} \) et \( c'_{10} \), telles que, pour toute fonction arithmétique \( f \), pour tout \( P > 1 \), et tout \( Y \geq 2 \) on ait les inégalités

\[
(2.4) \quad \sum_{P \leq p < 2P} \sum_{0 < a < p} \left| \sum_{n \leq x, \chi(n) = 1} f(n) - \frac{1}{p-1} \sum_{n \leq x, (n,p) = 1} f(n) \right|^2 \leq c_{10} (xP^{-1} + P) \left( \sum_{n \leq x} |f(n)|^2 \right)
\]

et

\[
(2.4') \quad \sum_{P \leq p < 2P} \sum_{0 < a < p} \left| \sum_{n \leq x, n \leq Y/p, \chi(n) = 1} f(n) - \frac{1}{p-1} \sum_{n \leq x, (n,p) = 1} f(n) \right|^2 \leq c'_{10} \log^2 Y (xP^{-1} + P) \left( \sum_{n \leq x} |f(n)|^2 \right).
\]

Preuve. Grâce à l'orthogonalité des caractères, ce qui est à l'intérieur de \(|\ldots|^2\) dans la partie gauche de (2.4), s'écrit sous la forme
\[
\frac{1}{p-1} \sum_{\chi \neq \chi_0} \chi(a) \sum_{n \leq x} f(n) \chi(n) := \frac{1}{p-1} \sum_{\chi \neq \chi_0} \chi(a) c_{\chi},
\]
par définition. Développant le carré et utilisant de nouveau l’orthogonalité des caractères, on a l’égalité

$$\sum_{0 < a < p} | \ldots |^2 = \frac{1}{(p-1)^2} \sum_{\chi, \chi' \neq \chi_0} c_{\chi} \overline{c_{\chi'}} \sum_{1 \leq a \leq p-1} \bar{\chi}(a) \chi'(a)$$

$$= \frac{1}{p-1} \sum_{\chi \neq \chi_0} |c_{\chi}|^2.$$ 

Ainsi, la quantité à gauche de (2.4) est égale à

$$\sum_{P \leq p < 2P} \frac{1}{p-1} \sum_{\chi \neq \chi_0} \left| \sum_{n \leq x} f(n) \chi(n) \right|^2$$

$$\leq \frac{1}{P-1} \sum_{P \leq p < 2P} \sum_{\chi \neq \chi_0} \left| \sum_{n \leq x} f(n) \chi(n) \right|^2$$

$$\leq \frac{1}{P-1} (x + 4P^2) \left( \sum_{n \leq x} |f(n)|^2 \right),$$

par l’inégalité de grand crible multiplicatif. Notre démonstration ne nécessite aucune connaissance de la répartition des valeurs de la fonction dans les progressions arithmétiques de petits modules (énoncés de type Siegel-Walfisz), puisque modulo $p$, tout caractère non principal est primitif. La démonstration en est d’autant simplifiée.

Pour passer de (2.4) à (2.4’), il faut rendre indépendantes les variables $p$ et $n$ liées par la contrainte multiplicative $pn \leq Y$. Parmi les multiples manières de le faire, nous avons choisi la transformée de Mellin, dans la forme que l’on trouve par exemple dans ([D-F-I] Lemma 9), via l’existence d’une fonction $h_Y$ telle que

$$\int_{-\infty}^{\infty} |h_Y(t)| \, dt < \log 6Y,$$

et telle que pour tout $k$ entier $\geq 1$, on ait

$$\int_{-\infty}^{\infty} h_Y(t) t^k \, dt = \begin{cases} 1 & \text{si } k \leq Y \\ 0 & \text{dans le cas contraire.} \end{cases}$$

En posant que $g_p(n, a)$ vaut $0$, $1 - \frac{1}{p-1}$, $-\frac{1}{p-1}$, suivant que $(p, n) > 1$, $n \equiv a \pmod{p}$, ou $n \not\equiv a \pmod{p}$ et $p \nmid n$, on écrit la partie gauche de (2.4’) sous
SOMMES DE MODULES DE SOMMES D’EXPONENTIELLES

la forme

\[ \sum_{p} \sum_{a} \left| \sum_{n} \int_{-\infty}^{\infty} h_Y(t)(pn)^it f(n)g_p(n,a)dt \right|^2 \]

\[ \leq \sum_{p} \sum_{a} \left( \int_{-\infty}^{\infty} |h_Y(t)| \left| \sum_{n} n^it f(n)g_p(n,a) \right| dt \right)^2 \]

\[ \leq \sum_{p} \sum_{a} \left( \int_{-\infty}^{\infty} |h_Y(t)| dt \right) \left( \int_{-\infty}^{\infty} |h_Y(t)| \left| \sum_{n} n^it f(n)g_p(n,a) \right|^2 dt \right) \]

\[ \leq \log 6Y \left( c_{10}(\log 6Y)(xP^{-1} + P) \left( \sum_{n} |f(n)|^2 \right) \right), \]

par l’inégalité de Cauchy-Schwarz, la propriété de la fonction \( h_Y \) et l’inégalité (2.4) appliquée à la fonction \( f(n)n^it \).

Le lemme suivant montre que pour presque tout entier \( n \), le produit des petits facteurs premiers de \( n \) est petit. On a ([Te] Lemme 3, [H-T] Theorem 07, p. 4):

**Lemme 2.6.** Il existe deux constantes absolues \( c_{11} \) et \( c_{12} > 0 \), telles, que pour tout \( 2 \leq u \leq v \leq x \), on a l’inégalité

\[ \sharp \left\{ n \leq x; \prod_{\substack{p \nmid n \\text{ et } p^\nu \geq v \\text{ pour tout } p \leq u}} \right\} \leq c_{11}x \exp \left( -c_{12} \frac{\log v}{\log u} \right). \]

Le lemme suivant est de nature combinatoire, il est obtenu par itération de la formule

\[ \sharp(A \cap B) \geq \sharp A + \sharp B - \sharp \mathcal{E}, \]

valable pour tous sous-ensembles \( A \) et \( B \) d’un ensemble fini \( \mathcal{E} \). On trouve déjà l’utilisation d’une telle inégalité dans ([Mi1] p. 77) pour rechercher des petites sommes d’exponentielles. On a:

**Lemme 2.7.** Soient \( \mathcal{E}_i \) (1 \( \leq i \leq k \)), k sous-ensembles d’un ensemble fini \( \mathcal{E} \)

On a alors l’inégalité

\[ \sharp(\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_k) \geq \sum_{i=1}^{k} \sharp \mathcal{E}_i - (k-1)\sharp \mathcal{E}. \]

Le dernier lemme est un cas particulier d’un résultat de Shiu ([Sh] Theorem 1). Il permet de majorer une fonction multiplicative à comportement raisonnable dans une progression arithmétique et contient, en prenant pour \( f \) la fonction caractéristique des entiers dont les facteurs premiers sont supérieurs à un certain \( z \), les habituelles majorations du crible.
Lemme 2.8. Soit $f$ une fonction multiplicative positive telle:

- Il existe une constante positive $A_1$ telle que, pour tout $p$ et tout $\ell \geq 1$, on ait
  \[ f(p^\ell) \leq A_1^\ell. \]
- Il existe une fonction $A_2 : \mathbb{R}^+ \rightarrow \mathbb{R}$ telle que, pour tout $n \geq 1$, on ait
  \[ f(n) \leq A_2(\varepsilon)n^\varepsilon. \]

Alors pour tous les entiers $a$ et $k$, vérifiant $(a,k) = 1$, tout réel $x$ tel que $x \geq k^{\frac{10}{11}}$ on a la relation

\[ \sum_{n \equiv a \pmod{k}} f(n) \ll \frac{x}{\varphi(k)} \frac{1}{\log x} \exp \left( \sum_{p \leq x, p \nmid k} \frac{f(p)}{p} \right), \]

où la constante implicite du symbole $\ll$, ne dépend que de $A_1$ et de la fonction $A_2$.

3. Preuve de la majoration de $A^*(x)$ et de $\tilde{A}(x)$.

Cette démonstration débute comme [Ho]. Pour alléger les notations, on pose

\[ K^{\ast}(a;n) = \frac{K(1,a;n)}{2^{\omega(n)}\sqrt{n}}, \]

ce qui conduit à l’égalité

\[ A^*(x) = \sum_{n \leq x} |K^*(1;n)|. \]

On pose

\[ Y = \exp \left( \frac{\log x}{c_{13}\log \log x} \right), \quad Z = x^{\frac{1}{2}}, \]

où $c_{13}$ est une constante choisie assez grande. Chaque entier $n$ se factorise de façon unique en

\[ n = n^b n^\flat, \]

avec

\[ n^b = \prod_{p \leq Y, p^\nu \| n} p^\nu. \]

La somme $A^*(x)$ se décompose en

\[ A^*(x) = \sum_{n \leq x, n^b \leq Z} |K^*(1;n)| + O \left( \sum_{n \leq x, n^b > Z} 1 \right), \]
soit encore
\[
A^*(x) = \sum_{n \leq x \atop n^2 \leq Z} |Kl^*(1; n)| + O\left(\frac{x}{\log x}\right),
\]
par le Lemme 2.6, pourvu qu’on ait \(c_{12}c_{13} \geq 4\), ce que nous supposerons par la suite.

On utilise la multiplicativit\'e cro\'ee, sous la forme
\[
|Kl^*(1; n)| = |Kl^*(\bar{n}^2; n^2)|, |Kl^*(\bar{n}^2; n^3)| \leq |Kl^*(n^2; n^3)|.
\]
On incorpore cette in\'egalit\'e dans (3.1) et on regroupe suivant les classes \(a\) modulo \(n^2\) (noter la relation \((n^2, n^3) = 1\)), d’où l’in\'egalit\'e
\[
A^*(x) \leq \sum_{m \leq Z, \atop p|m \Rightarrow p \leq Y} \sum_{a \mod{m} \atop (a, m) = 1} |Kl^*(a^2; m)| \left\{ \frac{x}{m} : r \leq \frac{x}{m}, r \equiv a \mod{m}, \right. \\
\left. p|r \Rightarrow p > Y \right\} + O\left(\frac{x}{\log x}\right).
\]
Puisqu’on a \(m \leq Z < \frac{x}{Z} \leq \left(\frac{x}{m}\right)^{\frac{1}{3}}\), une application classique du crible ([H-R] Théorème 3.6, ou Lemme 2.8, par exemple) donne
\[
A^*(x) \ll \sum_{m \leq Z, \atop p|m \Rightarrow p \leq Y} \sum_{a \mod{m} \atop (a, m) = 1} |Kl^*(a^2; m)| \left(\frac{x}{m\varphi(m)} \cdot \frac{1}{\log Y}\right) + \frac{x}{\log x}
\]
\[
\ll \frac{x \log \log x}{\log x} \sum_{m \leq Z, \atop p|m \Rightarrow p \leq Y} \frac{1}{m} \left(\frac{1}{\varphi(m)} \sum_{a \mod{m} \atop (a, m) = 1} |Kl^*(a^2; m)|\right) + \frac{x}{\log x}.
\]
D’apr\'es le Lemme 2.4, on a l’\'egalit\'e
\[
\frac{1}{\varphi(p)} \sum_{a \mod{p} \atop (a, p) = 1} |Kl^*(a^2; p)| = \frac{1}{\varphi(p)} \sum_{a \mod{p} \atop (a, p) = 1} |\cos \theta_{p, a^2}|
\]
\[
= \frac{2}{\pi} \int_0^\pi |\cos t| \sin^2 t \, dt + O(p^{-\frac{3}{4}})
\]
\[
= \frac{4}{3\pi} + O(p^{-\frac{1}{4}}).
\]
On rappelle aussi la relation triviale
\[ \frac{1}{\varphi(p^k)} \sum_{a \pmod{p^k}} \left| K^*(a^2; p^k) \right| \leq 1. \]

Ainsi, par la multiplicativité croisée, on a, pour tout \( m \geq 1 \), l’inégalité
\[ \frac{1}{\varphi(m)} \sum_{a \pmod{m}} \left| K^*(a^2; m) \right| \leq \kappa(m), \]

où \( \kappa(m) \) est la fonction multiplicative définie par
\[
\begin{cases}
\kappa(p) = \frac{1}{\varphi(p)} \sum_{a \pmod{p}} \left| K^*(a^2; p) \right| \\
\kappa(p^k) = 1 \quad (k \geq 2).
\end{cases}
\] (3.5)

Grâce à (3.4), l’inégalité (3.3) devient alors
\[
A^*(x) \ll \frac{x \log \log x}{\log x} \sum_{p \leq Y} \frac{\kappa(m)}{m} + \frac{x}{\log x} \prod_{p < Y} \left( 1 + \frac{4}{3\pi} + O\left( \frac{1}{p^{1/2}} \right) \right) + \frac{x}{\log x} \ll x, \quad \left( \frac{\log \log x}{\log x} \right)^{1 - \frac{4}{3\pi}},
\]

par la formule de Mertens. Ceci termine la preuve de majoration de \( A^*(x) \), Formule (1.4).

La démonstration de la majoration de \( \widetilde{A}(x) \) est assez proche de la précédente. On pose
\[ \widetilde{Kl}(a; n) = \frac{Kl(1, a; n)}{\sqrt{n}}. \]

Ainsi \( \left| \widetilde{Kl}(a; n) \right| \leq 2^{\omega(n)} \). On a la suite d’égalités
\[
\widetilde{A}(x) = \sum_{n \leq x} \left| \widetilde{Kl}(1; n) \right| = \sum_{n \leq x} \left| \widetilde{Kl}(1, n) \right| + O \left( \sum_{n \leq x} 2^{\omega(n)} \right).
\]
Par l’inégalité de Cauchy-Schwarz et le Lemme 2.6, on a

\[
\sum_{n \leq x, n^\beta > Z} 2^{\omega(n)} \leq \left( \sum_{n \leq x} 4^{\omega(n)} \right)^{\frac{1}{2}} \left( \sum_{n \leq x} 1 \right)^{\frac{1}{2}} = O \left( \frac{(x \log^3 x)^{\frac{1}{2}} }{ \left( x \exp \left( -c_{12} \frac{\log Z}{\log Y} \right) \right)^{\frac{1}{2}}} \right) = O \left( \frac{x}{\log x} \right),
\]

pourvu que \( c_{13} \) soit suffisamment grand (\( c_{12} c_{13} \geq 20 \)). Par la multiplicativité croisée écrite sous la forme

\[
|\tilde{K}(1; n)| \leq 2^{\omega(n)} |\tilde{K}(n^2; n^\beta)|,
\]
on a, de façon similaire à (3.2), l’inégalité

\[
\tilde{A}(x) \leq \sum_{m \leq Z} \sum_{a \equiv 1 (a, m)} |\tilde{K}(a^2; m)| \sum_{r \equiv a (m)} 2^{\omega(r)} + O \left( \frac{x}{\log x} \right)
\]

\[
\ll \sum_{m \leq Z} \sum_{a \equiv 1 (a, m)} |\tilde{K}(a^2; m)| \left( \frac{x}{m \phi(m)} \right)^{\frac{1}{2}} \log x \exp \left( \sum_{Y \leq p \leq x} \frac{2}{p} \right)
\]

\[
+ \frac{x}{\log x}
\]

par le Lemme 2.8. En utilisant la formule

\[
\sum_{p \leq y} \frac{1}{p} = \log \log y + c_{14} + o(1) \quad (y \to \infty),
\]
et la fonction \( \kappa \) introduite en (3.5), on parvient à

\[
\tilde{A}(x) \ll x \left( \log \log x \right)^2 \log x \sum_{m \leq Z} \frac{2^{\omega(m)} \kappa(m)}{m} + \frac{x}{\log x}
\]

\[
\ll x \left( \log \log x \right)^2 \prod_{p \leq Y} \left( 1 + \frac{8}{3\pi} + O(p^{-\frac{1}{4}}) \right) + \frac{x}{\log x}
\]

\[
\ll x \left( \log \log x \right) \left( \frac{\log x}{\log x} \right)^{1 - \frac{8}{3\pi}},
\]
ce qui termine la preuve de la majoration (1.5) de \( \tilde{A}(x) \). \( \square \)
4. Preuve des minorations de $A^*(x)$ et de $\tilde{A}(x)$.

L’outil principal en sera:

**Proposition 4.1.** Pour $P \to \infty$, on a l’égalité

$$
\sum_{p \in \mathcal{P}, \ \theta_p, n^2 \leq Y} f(n) = \left( \sum_{p \in \mathcal{P}, \ \theta_p, n^2 \leq Y} f(n) \right) \left( \frac{2}{\pi} \int_\alpha^\beta \sin^2 \theta \ d\theta + O(P^{-\frac{1}{2}}) \right)
$$

$$+ O \left( \left( \log Y \right) \left( P \# \mathcal{P} \right)^{\frac{1}{2}} \left( \frac{N}{P} + P \right)^{\frac{1}{2}} \left( \sum_n |f(n)|^2 \right)^{\frac{1}{2}} \right),
$$

uniformément sur $0 \leq \alpha < \beta \leq \pi$, sur tout ensemble de nombres premiers $\mathcal{P} \subset [P, 2P]$, toute fonction arithmétique $f$ telle que $f(n) = 0$ si $n > N$ ou $(n, \prod_{p \in \mathcal{P}} p) > 1$ et tout réel $Y \geq 2$.

**Preuve.** C’est une application du Lemme 2.3 (loi de Sato-Tate verticale pour les angles $\theta_p, a$) et du Lemme 2.5 (grand crible). On écrit

(4.1)

$$
\sum_{p \in \mathcal{P}, \ \theta_p, \pi^2 \in [\alpha, \beta]} f(n) = \sum_p \sum_{1 \leq a \leq p-1} \sum_{n \equiv a \ (\text{mod} \ p)} f(n) \sum_{n \leq Y} \left\{ \sum_{n \equiv a \ (\text{mod} \ p)} f(n) - \frac{1}{p-1} \sum_{n \leq Y} f(n) \right\}.
$$

Le premier terme à droite de (4.1) vaut, d’après le Lemme 2.3

$$
\sum_n f(n) \sum_{p \in \mathcal{P}, \ \theta_p, \pi^2 \in [\alpha, \beta]} \frac{1}{p-1} \sum_{1 \leq a \leq p-1} 1
$$

$$= \sum_n f(n) \sum_{p \in \mathcal{P}, \ \theta_p, \pi^2 \in [\alpha, \beta]} \left( \frac{2}{\pi} \int_\alpha^\beta \sin^2 \theta \ d\theta + O(P^{-\frac{1}{2}}) \right),
$$

pour $P \to \infty$, uniformément sur $Y$, $\alpha$ et $\beta$ comme dans l’énoncé. Pour le deuxième terme à droite de (4.1), on utilise l’inégalité de Cauchy-Schwarz.
et le Lemme 2.5, pour écrire que ce terme est

\[ \leq (P_{1}^{2}P)^{\frac{1}{2}} (c'_{10} \log^{2} Y)^{\frac{1}{2}} \left( \frac{N}{P} + P \right)^{\frac{1}{2}} \left( \sum_{n} |f(n)|^{2} \right)^{\frac{1}{2}}, \]

ce qui termine la preuve de la Proposition 4.1.

Passons à la preuve du Théorème 1.1. On fixe l’entier \( k \geq 4 \) puis on définit le réel \( \gamma \left( < \frac{\pi}{2} \right) \) tel que

\[ \frac{2}{\pi} \int_{0}^{\gamma} \sin^{2} \theta \, d\theta = \frac{1}{2} - \frac{1}{4k}. \]

On considère les \( k \)-uplets \((P_{1}, \ldots, P_{k})\) de réels définis comme suit

\[
\begin{align*}
\begin{cases}
4 \leq j \leq k, \quad P_{j} = 2^{\lambda_{j}} \exp \left( \log^{\frac{1}{x+y}} x \right), \quad \lambda_{j} \in \mathbb{N}, \quad P_{j} \leq \frac{1}{4} \exp \left( \log^{\frac{1}{2}} x \right) \\
\quad j = 3, \quad P_{3} = 2^{\lambda_{3} x^{\frac{1}{7}}}, \quad \lambda_{3} \in \mathbb{N}, \quad P_{3} \leq \frac{1}{4} x^{\frac{3}{10}} \\
\quad j = 2, \quad P_{2} = 2^{\lambda_{2} x^{\frac{1}{6}}}, \quad \lambda_{2} \in \mathbb{N}, \quad P_{2} \leq \frac{1}{4} x^{rac{3}{5}} \\
\quad j = 1, \quad P_{1} = 2^{\lambda_{1} x^{\frac{1}{2}}}, \quad \lambda_{1} \in \mathbb{N}, \quad P_{1} \ldots P_{k} \leq x.
\end{cases}
\end{align*}
\]

Ceci étant fixé, on note

\[
E(P_{1}, \ldots, P_{k}) = \{(p_{1}, \ldots, p_{k}) : P_{j} \leq p_{j} < 2P_{j} (1 \leq j \leq k), \ p_{1} \ldots p_{k} \leq x\},
\]

et

\[
E_{j}(P_{1}, \ldots, P_{k}) = \left\{ (p_{1}, \ldots, p_{k}) \in E(P_{1}, \ldots, P_{k}) ; \right. \\
\left. \theta_{p_{j}, p_{1} \ldots p_{j-1} p_{j+1} \ldots p_{k}} \in [0, \gamma] \cup [\pi - \gamma, \pi] \right\}.
\]

Remarquons que les inégalités contenues dans (4.2) entraînent que pour \((p_{1}, \ldots, p_{k})\) et \((p'_{1}, \ldots, p'_{k})\) éléments de \(E(P_{1}, \ldots, P_{k})\), on a \((p_{i}, p'_{j}) = 1\) pour \(i \neq j\), et l’encadrement

\[
(4.3) \quad \exp \left( \log^{\frac{1}{x+y}} x \right) \leq P_{j} \leq (P_{1} \ldots P_{k})^{\frac{1}{2}} x^{-\frac{1}{20}} \quad (1 \leq j \leq k).
\]

La Proposition 4.1 appliquée avec

\[ f(n) = \sharp \left\{ (p_{1}, \ldots, p_{j-1}, P_{j+1}, \ldots, p_{k}) ; \right. \\
\left. n = p_{1} \ldots p_{j-1} P_{j+1} \ldots p_{k}, \ P_{i} \leq p_{i} < 2P_{i} (i \neq j) \right\}, \]
\( Y = x, \mathcal{P} = \{p_j, P_j \leq p_j < 2P_j\} \), la définition de \( \gamma \) et les inégalités (4.3) impliquent la relation

\[
\sharp \mathcal{E}_j(P_1, \ldots, P_k)
= \left(1 - \frac{1}{2k}\right) + o(1) \sharp \mathcal{E}(P_1, \ldots, P_k) \\
+ O \left((\log x)P_j \left(\frac{P_1 \cdots P_{j-1}P_{j+1} \cdots P_k}{P_j} + P_j\right)^{\frac{1}{2}}(P_1 \cdots P_{j-1}P_{j+1} \cdots P_k)^{\frac{1}{2}}\right)
= \left(1 - \frac{1}{2k}\right) + o(1) \sharp \mathcal{E}(P_1, \ldots, P_k) + O \left(P_1 \cdots P_k \exp\left(-\frac{1}{3} \log^{\frac{1}{4}} x\right)\right).
\]

Ainsi, pour \( x \) assez grand, on a, pour tout \( 1 \leq j \leq k \)

\[
\sharp \mathcal{E}_j(P_1, \ldots, P_k) \geq \left(1 - \frac{2}{3k}\right) \sharp \mathcal{E}(P_1, \ldots, P_k) - O \left(x \exp\left(-\frac{1}{3} \log^{\frac{1}{4}} x\right)\right).
\]

Le Lemme 2.7 implique que l’intersection de sous-ensembles de \( \mathcal{E}(P_1, \ldots, P_k) \), notée

\[ \mathcal{F}(P_1, \ldots, P_k) := \mathcal{E}_1(P_1, \ldots, P_k) \cap \cdots \cap \mathcal{E}_k(P_1, \ldots, P_k) \]

est assez grande, puisqu’elle vérifie

\[
\sharp \mathcal{F}(P_1, \ldots, P_k) \geq \frac{1}{3} \sharp \mathcal{E}(P_1, \ldots, P_k) - O \left(x \exp\left(-\frac{1}{3} \log^{\frac{1}{4}} x\right)\right).
\]

Pour \( (p_1, \ldots, p_k) \in \mathcal{F}(P_1, \ldots, P_k) \), on a, par la multiplicativité croisée la minoration

\[
|\text{Kl}^* (1; p_1 \ldots p_k)| = \left| \cos \theta_{p_1, p_2 \ldots p_k} \right| \cdots \left| \cos \theta_{p_k, p_1 \ldots p_{k-1}} \right| \geq \cos^k \gamma.
\]

En sommant sur les \( (P_1, \ldots, P_k) \) vérifiant (4.2), on a la minoration

\[
\sum_{(P_1, \ldots, P_k) \in \mathcal{F}(P_1, \ldots, P_k)} \sum_{(p_1, \ldots, p_k) \in \mathcal{E}(P_1, \ldots, P_k)} |\text{Kl}^* (1; p_1 \ldots p_k)|
\geq \sum_{(P_1, \ldots, P_k) \in \mathcal{F}(P_1, \ldots, P_k)} \sum_{(p_1, \ldots, p_k) \in \mathcal{E}(P_1, \ldots, P_k)} |\text{Kl}^* (1; p_1 \ldots p_k)|
\geq \frac{\cos^k \gamma}{3} \sum_{(P_1, \ldots, P_k)} \sharp \mathcal{E}(P_1, \ldots, P_k) - O_k \left(x \exp\left(-\frac{1}{4} \log^{\frac{1}{4}} x\right)\right).
\]
SOMMES DE MODULES DE SOMMES D’EXPONENTIELLES 281

Ceci conduit donc à la minoration
\[ \sum_{n \leq x} |K^*(1;n)| \]
\[ \geq \frac{\cos^k \gamma}{3} \sum_{(P_1, \ldots, P_k)} \#\mathcal{E}(P_1, \ldots, P_k) - O_k \left( x \exp \left( -\frac{1}{4} \log \frac{1}{11} x \right) \right) \]
\[ \geq \frac{\cos^k \gamma}{3} \sum_{p_k \cdots p_2} \left( \pi \left( \frac{x}{2p_2 \cdots p_k} \right) - \pi(x^\frac{1}{3}) \right) \]
\[ - O_k \left( x \exp \left( -\frac{1}{4} \log \frac{1}{11} x \right) \right), \]

où les variables \( p_i \) vérifient \( x^{3/10} \leq p_2 < x^{1/3}/4, x^{1/4} \leq p_3 < x^{3/10}/4 \) et \( \exp(\log^{1/(j+1)} x) \leq p_j < \frac{1}{2} \exp(\log^{1/j} x) \), pour \( 4 \leq j \leq k \). Par application itérée du théorème des nombres premiers, on obtient la minoration
\[ \sum_{n \leq x} |K^*(1;n)| \gg_k \frac{x}{\log x} \left( \log \log x \right)^{k-3}, \]

ce qui termine la preuve de la minoration de \( A^*(x) \). Enfin on constate que les \( n = p_1 \ldots p_k \) comptés précédemment, sont tels que \( 2^{\omega(n)} = 2^k \), ce qui explique la minoration de \( A(x) \) (Théorème 1.2) avec la valeur annoncée \( \tilde{c}_0(k) = 2^{k+3}c^*_0(k) \).

\[ \square \]

5. Preuve du Théorème 1.3.

Donnons maintenant quelques indications sur la preuve du Théorème 1.3. Elles consistent essentiellement à rendre effective la constante \( c^*_0(k) \) du Théorème 1.1, et à prendre \( k \) comme fonction de \( x \).

Soit \( \nu \) un réel tel que
\[ \frac{\pi}{4} \nu e^{1+\nu} > 1 \text{ et } \nu < \frac{2}{5}. \]

Posons alors
\[ k = \left( \log \log x \right)^{\frac{1}{2+\nu}}. \]

La définition (4.2) des \( k \)-uplets \( (P_1, \ldots, P_k) \) est inchangée pour \( P_1, P_2 \) et \( P_3 \), par contre pour \( 4 \leq j \leq k \), on pose
\[ P_j = 2^{\lambda_j} \exp \left( \log^{\frac{1}{j+1}} x \right), \quad \lambda_j \in \mathbb{N}, \quad P_j \leq \frac{1}{4} \exp \left( \log^{\frac{1}{j}} x \right). \]

Enfin, soit \( \xi = \xi_\nu \), un réel légèrement inférieur à 1/2, dont la valeur sera précisée ultérieurement, et soit \( \gamma < \frac{1}{2} \) tel que
\[ \frac{2}{\pi} \int_0^\gamma \sin^2 \theta \, d\theta = \frac{1}{2} - \frac{\xi}{k}. \]

Notons dès à présent, qu’on a
\[(5.1) \quad \cos \gamma \sim \frac{\pi \xi}{2k}. \]
Définissant de la même manière qu’au §4, les quantités $\mathcal{E}(P_1,\ldots,P_k)$, $\mathcal{E}_j(P_1,\ldots,P_k)$ et $\mathcal{F}(P_1,\ldots,P_k)$, et appliquant de nouveau la Proposition 4.1, on a, pour tout $1 \leq j \leq k$, la minoration

$$\sharp \mathcal{E}_j(P_1,\ldots,P_k) \geq \left(1 - \frac{2\xi}{k} \cdot \frac{1}{\frac{1}{2} + \xi}\right) \sharp \mathcal{E}(P_1,\ldots,P_k) - O\left(x \exp\left(-\frac{1}{2} \exp(\log x)^{\frac{1}{1+\nu}}\right)\right),$$

dont on déduit, grâce au Lemme 2.7, la minoration

$$\sharp \mathcal{F}(P_1,\ldots,P_k) \geq \left(1 - \frac{2\xi}{1+2\xi}\right) \sharp \mathcal{E}(P_1,\ldots,P_k) - O\left(x \exp\left(-\frac{1}{3} \exp(\log x)^{\frac{1}{1+\nu}}\right)\right).$$

Poursuivant la même démarche qu’au §4, on obtient la minoration

$$\sum_{n \leq x} \left| \mathbf{K}_1^*(1;n) \right| \geq (\cos^k \gamma) \left(\frac{1-2\xi}{1+2\xi}\right) \sum_{(P_1,\ldots,P_k)} \sharp \mathcal{E}(P_1,\ldots,P_k) - O\left(x \exp\left(-\frac{1}{3} \exp(\log x)^{\frac{1}{1+\nu}}\right)\right) \geq \left(\cos^k \gamma\right) \left(\frac{1-2\xi}{1+2\xi}\right) \sum_{p_k} \cdots \sum_{p_2} \left(\pi\left(\frac{x}{2p_2\ldots p_k}\right) - \pi(x^{\frac{1}{3}})\right),$$

où les variables $p_i$ vérifient $x^{3/10} \leq p_2 < x^{1/3}/4$, $x^{1/4} \leq p_3 < x^{3/10}/4$ et $\exp(\log x^{\frac{1}{1+\nu'}} x) \leq p_j < \frac{1}{4} \exp(\log x^{\nu'} x)$, pour $4 \leq j \leq k$. On utilise la formule

$$\sum_{\exp(\log x^{\frac{1}{1+\nu'}} x) \leq p_j \leq \frac{1}{4} \exp(\log x^{\frac{1}{j+1}} x)} \frac{1}{p_j} \geq \frac{\nu'}{(j+1)^{\nu'+1}} \log x,$$

valable pour $4 \leq j \leq k$, tout $\nu' < \nu$ et tout $x$ suffisamment grand. Re-groupant (5.1), (5.2) et (5.3), on a, pour une certaine constante $A$ et pour tout $\xi' < \xi$, la minoration

$$\sum_{n \leq x} \left| \mathbf{K}_1^*(1;n) \right| \gg \nu'^k \left(\frac{\pi \xi'}{2k}\right)^k (\log log x)^{k-A} \cdot \frac{x}{\log x} \cdot \prod_{j \leq k} \frac{1}{(j+1)^{\nu'+1}}.$$
La formule de Stirling et la définition de $k$ donnent, pour certaines constantes $A'$ et $A''$, la minoration
\[
\sum_{n \leq x} |Kl^*(1;n)| \gg \left( \frac{\pi \xi' \nu' e^{1+\nu} (\log \log x)}{2k^{2+\nu}} \right)^k \cdot \frac{x}{\log x} \cdot (\log \log x)^{-A'}
\]
\[
\gg \left( \frac{\pi \xi' \nu' e^{1+\nu}}{2} \right)^k \cdot \frac{x}{\log x} \cdot (\log \log x)^{-A''}.
\]
Pour terminer, choisissons $\nu'$ suffisamment proche de $\frac{1}{2}$, pour que l'inégalité
\[
\frac{\pi \xi' \nu' e^{1+\nu}}{2} > 1,
\]
soit vérifiée, et remarquons que, pour $\nu = \frac{2}{5}$, on a $k = \left[ \frac{5}{12} \right]$. □

6. Preuve du Théorème 1.5.

La preuve du Théorème 1.5 n'est pas structurellement différente de celle des Théorèmes 1.1 et 1.2, mais nécessite de placer l'étude des sommes $S_f$ dans le cadre suffisamment général construit par Katz ([K2] et [K3]). Ce cadre apparaît de nouveau dans [Mi2]. Ici, dans un premier temps, nous résumons [F-M] §2.

Pour chaque fraction rationnelle vérifiant $\mathcal{H}$ ou $\mathcal{H}'$, pour chaque nombre premier $p$ assez grand, chaque premier $\ell \neq p$, on construit un $\mathbb{Q}_\ell$-faisceau de rang $k_f$ sur $\mathbb{P}^1_{\mathbb{F}_p}$, noté $\mathcal{S}_f$, qui vérifie entre autres les propriétés:

- Pour tout $a \in \mathbb{F}^*_p$, on a
  \[
  \text{tr} (\text{Frob}_a, \mathcal{S}_f) = \alpha_{f,p,a} \frac{S_f(\pi;p)}{\sqrt{p}}
  \]
  avec $\alpha_{f,p,a}$ nombre complexe de module 1.

- Le groupe de monodromie géométrique $G_{\text{géom}}$ coïncide avec le groupe de monodromie arithmétique $\text{SL}_{k_f}$, si $f$ vérifie $\mathcal{H}$ ou vaut $\text{Sp}_{k_f}$, si $f$ vérifie $\mathcal{H}'$.

Soit $K$ un compact maximal de $G_{\text{géom}}$. On rappelle (voir §1 ci-dessus) qu’on choisit

\[
K = G_f,
\]
avec $G_f = \text{SU}_{k_f}(\mathbb{C})$, si $f$ vérifie $\mathcal{H}$ et $G_f = \text{USp}_{k_f}(\mathbb{C})$, si $f$ vérifie $\mathcal{H}'$. Soit $K^\circ$ l’ensemble des classes de conjugaison de $K$ et soit $\mu_{\text{ST}}$ la mesure image sur $K^\circ$ de la mesure $\mu_{\text{Haar}}^K$ par la projection canonique. Pour tout $a \in \mathbb{F}^*_p$, la classe de Frobenius $\text{Frob}_a$ définit une classe de conjugaison $\theta^a_{p,a} \in K^\circ$, pour
laquelle on a l’égalité
\[ \left| \text{tr} \left( \theta_{p,a}^{\natural} \right) \right| = \frac{|S_f(\pi;p)|}{\sqrt{p}} \left( = k_f \cos \theta_{f,p,\pi} \right). \]

(Voir §1, pour la définition de \( \theta_{f,p,a} \).) Le résultat fondamental de Katz est:

**Proposition 6.1** ([Ka3], 7.9, 7.10). *Sous les hypothèses précédentes, quand \( p \to \infty \), les classes de conjugaison \( \{ \theta_{p,a}^{\natural} \}_{a \in F_p^\times} \subset K^2 \) deviennent équiréparties pour la mesure \( \mu_{ST} \), i.e., pour toute fonction \( g \), continue sur \( K^2 \), on a
\[
\lim_{q \to \infty} \frac{1}{p-1} \sum_{a \in F_p^\times} g(\theta_{p,a}^{\natural}) = \int_{K^2} g(\theta^\natural) \, d\mu_{ST}.
\]

L’emploi de cette proposition conduirait aux encadrements de \( A_f^\ast(x) \) et de \( \tilde{A}_f(x) \) énoncés dans le Théorème 1.5, mais avec un facteur supplémentaire de la forme \( \log^\varepsilon x \) pour chacune des majorations. Pour éviter l’apparition de ce facteur, il est nécessaire de faire un calcul de discrépance dans le même esprit que lors de la preuve des Lemmes 2.2, 2.3 et 2.4, mais dans un cadre plus général. Un tel calcul a été fait dans [F-M] §2, ce qui nous permet d’en esquisser les principales étapes.

Soit \( h \) une fonction radiale sur \( \mathbb{C} \), à support compact, à valeurs dans \( \mathbb{R}^+ \), de classe \( C^\infty \). On considère la fonction \( H \) définie par
\[
K \quad \longrightarrow \quad \mathbb{R}^+
\theta \quad \longmapsto \quad H(\theta) = h \left( \frac{\text{tr} \, \theta}{k_f} \right).
\]

En tout point \( \theta \in K \), on a le développement en série
\[
H(\theta) = \int_K H(\theta') \, d\mu_{K}^{\text{Haar}}(\theta') + \sum_{\rho} \hat{H}(\rho) \text{tr}(\rho(\theta)), \tag{6.1}
\]

où \( \rho \) parcourt l’ensemble des représentations irréductibles non triviales de \( K \) et
\[
\hat{H}(\rho) = \int_K H(\theta') \text{tr}(\rho(\theta')) \, d\mu_{K}^{\text{Haar}}(\theta').
\]

Par sommation de (6.1) sur les \( \theta_{p,a}^{\natural} \), on obtient
\[
\frac{1}{p-1} \sum_{a \in F_p^\times} H(\theta_{p,a}^{\natural}) = \int_K H(\theta') \, d\mu_{K}^{\text{Haar}}(\theta')
\]
\[
\quad + O \left( \frac{1}{p-1} \sum_{\rho} |\hat{H}(\rho)| \left| \sum_{a \in F_p^\times} \text{tr} \left( \rho(\theta_{p,a}^{\natural}) \right) \right| \right).
\]
La majoration ([F-M] Lemme 2.3)

\[ \left| \sum_{a \in \mathbb{F}_p^\times} \text{tr} \left( \rho(\theta^a_{p,a}) \right) \right| \leq k_f \dim \rho \sqrt{p} \]

transforme (6.2) en

\[ \frac{1}{p-1} \sum_{a \in \mathbb{F}_p^\times} H(\theta^a_{p,a}) = \int_K H(\theta') \, d\mu^\text{Haar}_K(\theta') + O \left( \| H \|^2 p^{-\frac{1}{2}} \right), \]

avec

\[ \| H \|^2 = \sum_{\rho} \dim \rho \, |\hat{\rho}|. \]

Pour parfaire l'étude du terme d'erreur de (6.3), nous étudions \( \| H \|^2 \). Dans ce but, nous introduisons certaines hypothèses décrivant la régularité de \( h \):

(6.4) Il existe \( \Delta > 0 \) et des constantes \( c_j \) tels qu'on ait

\[ |h^{(j)}(t)| \leq c_j \Delta^{-j} (\forall t \in \mathbb{R}^+ \quad \forall j \geq 0) \]

et

(6.5) Le support de \( h \) est inclus dans un intervalle de longueur \( L \).

(Notons qu’en raison de l’application à la fonction \( H \), on peut supposer l’inégalité \( L \leq 2k_f \).) La majoration de \( \| H \|^2 \) a été traitée dans le cas particulier où \( L = \Delta \) ([F-M], Formule (2.21) car dans ce travail, on cherchait de petites sommes d'exponentielles). Par une légère généralisation, nous avons:

**Lemme 6.2.** Pour toute fonction \( h \) comme ci-dessus, vérifiant en outre les conditions (6.4) et (6.5), on a l’inégalité

\[ \| H \|^2 \ll \Lambda(L, K) \Delta^{-\frac{\dim K}{2}}, \]

avec

\[ \Lambda(L, K) = \begin{cases} 
L^{\frac{1}{2}} & \text{si } K = \text{USp}_k \text{ et } k \geq 2 \\
L & \text{si } K = \text{SU}_k \text{ et } k \geq 3, \ k \neq 4 \\
L (\log(1/L))^\frac{1}{2} & \text{si } K = \text{SU}_k \text{ et } k = 4.
\end{cases} \]

La constante dans le symbole \( \ll \) ne dépend que de \( k \) et de la suite des \( c_j \) de l’hypothèse (6.4).

Dans l’énoncé du Lemme 6.2, on rappelle les valeurs respectives des dimensions

\[ \dim \text{SU}_k = k^2 - 1, \quad \dim \text{USp}_k = \frac{k}{2}(k + 1), \]
et on remarque l’inégalité \( \Lambda(L,K) = O_{k_f}(1) \). Par la définition de la mesure \( \mu_{G_f} \), donnée au §1, on a l’égalité

\[
\int_K H(\theta') \, d\mu_{K}^{\text{Haar}}(\theta') = \int_0^\pi h(|\cos \theta|) \, d\mu_{G_f}(\theta).
\]

Les remarques précédentes, la Formule (6.2) et le Lemme 6.2 entraînent le résultat suivant, que nous pourrions rendre plus précis mais qui sera satisfaisant pour les applications:

**Lemme 6.3.** Soit \( f \) une fraction rationnelle comme ci-dessus, vérifiant \( k_f \geq 2 \) et vérifiant \( H \) ou \( H' \). Soit \( h : \mathbb{R} \to \mathbb{R}^+ \) une fonction à support compact, de classe \( C^\infty \), vérifiant (6.4) et (6.5), pour un certain \( \Delta > 0 \). Il existe alors une constante \( \delta_f > 0 \), telle qu’on ait l’égalité

\[
\frac{1}{p-1} \sum_{a \in \mathbb{F}_p} h(|\cos \theta_{f,p,a}|) = \int_0^\pi h(|\cos \theta|) \, d\mu_{G_f}(\theta) + O_f \left( p^{-\frac{1}{2}} \Delta^{-\delta_f} \right).
\]

Le Lemme 6.3 joue ainsi le rôle du Lemme 2.2. En recopiant la preuve des Théorèmes 1.1 et 1.2 dans le cadre plus général qui nous intéresse, nous parvenons aux encadrements

\[
(6.6) \quad c_5^*(k) \frac{x}{\log x} (\log \log x)^k \leq A_f^*(x) \leq c_6^* x \left( \frac{\log \log x}{\log x} \right)^{1-I_f}
\]

et

\[
(6.7) \quad \widetilde{c}_5(k) \frac{x}{\log x} (\log \log x)^k \leq \widetilde{A}_f(x) \leq \widetilde{c}_6 x (\log \log x)^{k_f-1} \left( \frac{\log \log x}{\log x} \right)^{1-k_f I_f},
\]

où \( I_f \) désigne l’intégrale

\[
I_f = \int_0^\pi \cos t \, d\mu_{G_f}(t).
\]

Il reste à majorer cette intégrale. Par définition, on a l’égalité

\[
I_f = \frac{1}{k_f} \int_{G_f} |\text{trace } (A)| \, d\mu_{G_f}^{\text{Haar}}(A),
\]

où, suivant les cas \( G_f = \mathrm{SU}_{k_f} \) ou \( G_f = \mathrm{USp}_{k_f} \) et \( d\mu_{G_f}^{\text{Haar}} \) est la mesure de Haar correspondante. Par l’inégalité de Cauchy-Schwarz, on a

\[
(6.8) \quad I_f \leq \frac{1}{k_f} \left( \int_{G_f} d\mu_{G_f}^{\text{Haar}}(A) \right)^{\frac{1}{2}} \left( \int_{G_f} |\text{trace } (A)|^2 \, d\mu_{G_f}^{\text{Haar}}(A) \right)^{\frac{1}{2}}.
\]

Par définition de la mesure de Haar, la première intégrale vaut 1. Pour interpréter la seconde intégrale à droite de (6.8), on écrit que \( A \mapsto |\text{trace } (A)|^2 \) est le caractère de la représentation \( \mathrm{St} \otimes \overline{\mathrm{St}} \) de \( G_f \), où \( \mathrm{St} \) désigne la
représentation standard de $G_f$, (qui agit sur l’espace vectoriel naturel sous-jacent $\mathbb{C}^{k_f}$) et
\[
\int_{G_f} |\text{trace } (A)|^2 \, d\mu_{G_f}^{\text{Haar}}(A)
\]
vaut précisément la dimension des $G_f$-invariants de la représentation $\text{St} \otimes \overline{\text{St}}$. Il est facile de voir que cet espace est de dimension 1. Par (6.8), on a donc la majoration
\[
I_{k_f} \leq \frac{1}{k_f}.
\]
Reportant cette majoration dans (6.6) et (6.7), on termine ainsi la preuve du Théorème 1.5. □

**Remerciements.** Le premier auteur tient à remercier R. de la Bretèche, H. Iwaniec et P. Sarnak pour des remarques concernant une première version de cet article.

**Références**


Received February 4, 2002.

Mathématique
Campus d’Orsay
F-91405 Orsay Cedex
France
E-mail address: Etienne.Fouvry@math.u-psud.fr

Mathématique
Université Montpellier II, CC 051
34095 Montpellier Cedex
France
E-mail address: michel@math.univ-montp2.fr
A ZETA FUNCTION FOR FLIP SYSTEMS

Young-One Kim, Jungseob Lee, and Kyewon K. Park

In this paper, we investigate dynamical systems with flip maps, which can be regarded as infinite dihedral group actions. We introduce a zeta function for flip systems, and find its basic properties including a product formula. When the underlying $\mathbb{Z}$-action is conjugate to a topological Markov shift, the flip system is represented by a pair of matrices, and its zeta function is expressed explicitly in terms of the representation matrices.

1. Introduction.

Let $(X, T)$ be a topological dynamical system, where $X$ is a topological space and $T : X \to X$ a homeomorphism. A homeomorphism $F : X \to X$ is called a flip map or simply a flip for $(X, T)$ if

$$TF = FT^{-1} \quad \text{and} \quad F^2 = \text{id}.$$  

We call the triplet $(X, T, F)$ a flip system. It is easy to see that if $(X, T, F)$ is a flip system, then $(X, T^m, T^n F)$ is also a flip system for any $m, n \in \mathbb{Z}$. Since the infinite dihedral group $D_\infty$ is generated by two elements $a$ and $b$ such that

\begin{equation}
ab = ba^{-1} \quad \text{and} \quad b^2 = 1,
\end{equation}

a flip system can be regarded as a $D_\infty$-action of homeomorphisms.

Two flip systems $(X, T, F)$ and $(X', T', F')$ are said to be conjugate if there is a homeomorphism $\Phi : X \to X'$ such that

$$\Phi T = T' \Phi \quad \text{and} \quad \Phi F = F' \Phi.$$  

In this case, we write $(X, T, F) \cong (X', T', F')$, and $\Phi$ is called a conjugacy from $(X, T, F)$ to $(X', T', F')$. For an arbitrary flip system $(X, T, F)$, $T$ is a conjugacy from $(X, T, F)$ to $(X, T, T^2 F)$ and $F$ is a conjugacy from $(X, T, F)$ to $(X, T^{-1}, F)$.

Since there is a dynamical system $(X, T)$ which is not conjugate to its time reversal $(X, T^{-1})$, not every dynamical system has a flip. See [3, p. 104] and also Example 4.1. On the other hand, any topological Markov shift whose transition matrix is symmetric has a natural flip.

It is well-known that measurable $D_\infty$-actions are isomorphic if the underlying $\mathbb{Z}$-actions are Bernoulli of the same entropy. In [7] it is shown that...
if the underlying \( \mathbb{Z} \)-actions are Kolmogorov and isomorphic, there are examples of non-isomorphic \( D_\infty \)-actions. Unlike the measurable case, we can construct infinitely many non-conjugate flips for a full shift in the topological setting. See Example 4.2.

We establish a zeta function for flip systems which is a conjugacy invariant, and give a finite description of the function when the underlying \( \mathbb{Z} \)-action is conjugate to a topological Markov shift.

The Artin-Mazur zeta function \( \zeta_T \) for a dynamical system \( (X,T) \), found in [1], is defined by

\[
\zeta_T(t) = \exp \left( \sum_{n=1}^{\infty} \frac{p_n t^n}{n} \right),
\]

where

\[
p_n = |\{ x \in X : T^n x = x \}| \quad (n = 1, 2, \ldots).
\]

(We assume that the sequence \( \{p_n\} \) is bounded.) The Artin-Mazur zeta function has the product formula

\[
\zeta_T(t) = \prod_\gamma \frac{1}{1 - t |\gamma|},
\]

where the product is taken over all finite orbits \( \gamma \) of \( T \).

In [5], D. Lind introduced a zeta function for \( \mathbb{Z}^d \)-actions that generalizes the Artin-Mazur zeta function. It is straightforward to extend the notion to the case of general group actions. Let \( G \) be a group, \( X \) a set and \( \alpha : G \times X \to X \) a \( G \)-action on \( X \). Then the zeta function \( \zeta_\alpha \) of the action \( \alpha \) is defined formally by

\[
\zeta_\alpha(t) = \exp \left( \sum_H \frac{p_H |G/H| t^{|G/H|}}{|G/H|} \right),
\]

where

\[
p_H = |\{ x \in X : \forall h \in H \quad \alpha(h,x) = x \}|.
\]

Here, the sum is taken over all finite-index subgroups \( H \) of \( G \), that is, subgroups \( H \) such that \( |G/H| < \infty \), and \( p_H \) is defined by

\[
\zeta_\alpha = \zeta_{\tilde{\alpha}}.
\]

It is easy to see that this zeta function is automorphism-invariant in the following sense: If \( \Psi : G \to G \) is an automorphism and two \( G \)-actions \( \alpha : G \times X \to X \) and \( \tilde{\alpha} : G \times X \to X \) satisfy \( \tilde{\alpha}(g,x) = \alpha(\Psi(g),x) \) for all \( (g,x) \in G \times X \), then \( \zeta_\alpha = \zeta_{\tilde{\alpha}} \).

We define the zeta function \( \zeta_{T,F} \) of a flip system \( (X,T,F) \) to be the zeta function \( \zeta_\alpha \) of the \( D_\infty \)-action \( \alpha : D_\infty \times X \to X \) that is given by

\[
\alpha(a,x) = Tx \quad \text{and} \quad \alpha(b,x) = Fx \quad (x \in X),
\]

where \( a \) and \( b \) are generators of \( D_\infty \) which satisfy (1.1). Since the zeta function is automorphism-invariant, our definition does not depend on the choice of the generators \( a \) and \( b \). Moreover, it is clear that this zeta function
is a conjugacy invariant. There are, however, non-conjugate flip systems with the same zeta function. See Examples 4.3 and 4.4.

In Section 2, we express the zeta function of flip systems in a more tractable form, and establish some of its basic properties including the product formula. In Section 3, we consider the flip systems \((X, T, F)\) such that \((X, T)\) is conjugate to a topological Markov shift. We prove that such a system can be represented by a pair of matrices (Representation Theorem), and express its zeta function in terms of those matrices. Finally, in Section 4, we conclude this paper with some examples.

2. The zeta function of a flip system.

Let \((X, T, F)\) be a flip system, and suppose that \(D_\infty\) is generated by \(a\) and \(b\) satisfying (1.1). Let \(\alpha : D_\infty \times X \to X\) denote the \(D_\infty\)-action defined by (1.5). For a finite-index subgroup \(H\) of \(D_\infty\) set \(p_H = |\{x \in X : \forall h \in H, \alpha(h, x) = x\}|\) and suppose that \(p_H < \infty\) for all finite-index subgroups \(H\) of \(D_\infty\).

In order to express \(\zeta_\alpha\) explicitly, we need to identify all the finite-index subgroups of \(D_\infty\). Suppose that \(H\) is a finite-index subgroup of \(D_\infty\). Then there is an integer \(k \neq 0\) such that \(a^k \in H\), since otherwise we must have \(|H| \leq 2\). Hence, either \(H\) is generated by \(a^i\) for some integer \(i \neq 0\) or by \(a^i\) and \(a^j b\) for some integers \(i\) and \(j\) with \(i \neq 0\).

Let \(H(i)\) denote the subgroup generated by \(a^i\), and \(H(i, j)\) the one generated by \(a^i\) and \(a^j b\). Then it is clear that \(H(i) = H(k)\) if and only if \(|i| = |k|\), and that \(H(i, j) = H(k, l)\) if and only if \(|i| = |k|\) and \(j - l\) is a multiple of \(i\). Moreover, \(|D_\infty/H(i)| = 2|i|\) and \(|D_\infty/H(i, j)| = |i|\) for \(i \neq 0\). Therefore we obtain the following:

**Lemma 2.1.** Let \(n\) be a positive integer. If \(n\) is odd, then

\[
H(n, 0), H(n, 1), \ldots, H(n, n - 1)
\]

are all the subgroups of \(D_\infty\) with index \(n\). In addition to these, there is one more such subgroup \(H(n/2)\) if \(n\) is even.

For convenience, we set \(p_i = p_{H(i)}\) and \(p_{i, j} = p_{H(i, j)}\). Then we have

\[
\begin{align*}
q_i &= |\{x \in X : T^i x = x\}| \quad \text{and} \\
q_{i, j} &= |\{x \in X : T^i x = T^j F x = x\}|.
\end{align*}
\]

Hence (1.4) and Lemma 2.1 imply that

\[
\zeta_{T, F}(t) = \exp\left(\sum_{n=1}^{\infty} \frac{p_n}{2n} t^{2n} + \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} \frac{p_{n, k}}{n} t^n\right).
\]

Now, observe that \(aH(i, j)a^{-1} = H(i, j + 2)\). From this, we see that \(p_{i, j} = p_{i, j+2}\). Moreover, it is clear that \(p_{i, j} = p_{i, i+j}\). Hence we obtain the
following:

\[ \sum_{k=0}^{n-1} \frac{p_{n,k}}{n} = \begin{cases} p_{n,0} & \text{if } n \text{ is odd,} \\ \frac{(p_{n,0} + p_{n,1})}{2} & \text{if } n \text{ is even.} \end{cases} \]  

By (1.2) we have

\[ \exp \left( \sum_{n=1}^{\infty} \frac{p_n}{2n} t^{2n} \right) = \sqrt{\zeta_T(t^2)}. \]  

**Theorem 2.2.** The zeta function \( \zeta_{T,F} \) of the flip system \((X, T, F)\) is given by

\[ \zeta_{T,F}(t) = \sqrt{\zeta_T(t^2)} \exp (G_{T,F}(t)), \]

where \( \zeta_T \) is the Artin-Mazur zeta function of \((X, T)\), and

\[ G_{T,F}(t) = \sum_{m=1}^{\infty} \left( p_{2m-1,0} t^{2m-1} + \frac{p_{2m,0} + p_{2m,1}}{2} t^{2m} \right). \]

**Proof.** The theorem is an immediate consequence of (2.2), (2.3) and (2.4). \( \square \)

**Corollary 2.3.** Let \( R_T \) and \( R_{T,F} \) denote the radii of convergence of the Maclaurin series of \( \zeta_T(t) \) and \( \zeta_{T,F}(t) \), respectively. If \( p_n > 0 \) for some \( n \), then we have

\[ 0 \leq R_T \leq R_{T,F} \leq \sqrt{R_T} \leq 1. \]

**Remark 2.4.** If \((X, T)\) is conjugate to a subshift, then it is easy to see that the radius of convergence of \( G_{T,F} \) is at least \( \exp(h_T/2) \), where \( h_T \) is the topological entropy of \((X, T)\). Moreover, if \((X, T)\) is conjugate to a sofic shift, then \( h_T = \log R_T \) (see [6, Chapter 4]), and hence \( R_{T,F} = \sqrt{R_T} \).

In the remainder of this section, we establish the product formula of the zeta function. Suppose that \( \gamma \) is a finite orbit of \((X, T, F)\). Then there is a point \( x \) such that \( \gamma = \{x, Tx, \ldots, T^{\vert \gamma \vert - 1} x\} \), or there is a point \( x \) such that \( \gamma = \{x, Tx, \ldots, T^{k-1} x\} \cup \{Fx, TFx, \ldots, T^{k-1}Fx\} \) with \( \vert \gamma \vert = 2k \). In the first case, we write \( \gamma \in O_1 \), and in the second case, \( \gamma \in O_2 \). It is obvious that \( O_1 \cap O_2 = \emptyset \). We denote by \( \zeta_{\gamma} \) the zeta function of the flip system \((\gamma, T|_{\gamma}, F|_{\gamma})\).

**Lemma 2.5.** If \( \gamma \in O_1 \),

\[ \zeta_{\gamma}(t) = \sqrt{\frac{1}{1 - t^{2 \vert \gamma \vert}}} \exp \left( \frac{t^{\vert \gamma \vert}}{1 - t^{\vert \gamma \vert}} \right), \]

and if \( \gamma \in O_2 \),

\[ \zeta_{\gamma}(t) = \frac{1}{1 - t^{\vert \gamma \vert}}. \]
Proof. Let
\[ \tilde{p}_i = |\{x \in \gamma : T^ix = x\}| \quad \text{and} \quad \tilde{p}_{i,j} = |\{x \in \gamma : T^ix = T^jFx = x\}|. \]

Assume that \( \gamma \in O_1 \) and \( n \) is a positive integer. If \( n \) is not a multiple of \( |\gamma| \), then no elements of \( \gamma \) are fixed by \( T^n \), and hence \( \tilde{p}_n = 0 \) and \( \tilde{p}_{n,k} = 0 \) for all \( k \). Now suppose \( n \) is a multiple of \( |\gamma| \). Then every element of \( \gamma \) is fixed by \( T^n \), so that \( \tilde{p}_n = |\gamma| \). We can see that if \( |\gamma| \) is odd, \( \tilde{p}_n,0 = 1 \); if \( |\gamma| \) even, either \( \tilde{p}_{n,0} = 0 \) and \( \tilde{p}_{n,1} = 2 \), or \( \tilde{p}_{n,0} = 2 \) and \( \tilde{p}_{n,1} = 0 \). Using (2.2) and (2.3) with \( \tilde{p}_n \) and \( \tilde{p}_{n,k} \) in place of \( p_n \) and \( p_{n,k} \) respectively, we have
\[
\zeta(\gamma)(t) = \exp \left( \sum_{m=1}^{\infty} \frac{1}{2m^2} t^{2m|\gamma|} + \sum_{m=1}^{\infty} t^{m|\gamma|} \right),
\]
from which the first assertion follows.

Next, assume that \( \gamma \in O_2 \). Then for each integer \( j \) no elements of \( \gamma \) are fixed by \( T^jF \). Hence \( \tilde{p}_{n,0} = \tilde{p}_{n,1} = 0 \) for all \( n \). Moreover, it is easy to see that \( \tilde{p}_n = |\gamma| \) if \( n \) is a multiple of \( |\gamma|/2 \), and \( \tilde{p}_n = 0 \) otherwise. Again using (2.2) and (2.3) we have
\[
\zeta(\gamma)(t) = \exp \left( \sum_{m=1}^{\infty} \frac{1}{m^m} t^{m|\gamma|} \right),
\]
from which the second assertion follows. \( \square \)

**Theorem 2.6.** Let \( R_{T,F} \) be the radius of convergence of the Maclaurin series of \( \zeta_{T,F} \), and suppose that \( R_{T,F} > 0 \). Then we have
\[
\zeta_{T,F}(t) = \prod_{\gamma \in O_1} \sqrt{1 - t^{2|\gamma|}} \prod_{\gamma \in O_2} \frac{1}{1 - t^{2|\gamma|}} \quad (|t| < R_{T,F}).
\]

**Proof.** It is clear from the definition that
\[
\zeta_{T,F}(t) = \prod_{\gamma} \zeta(\gamma)(t) \quad (|t| < R_{T,F}),
\]
where the product is taken over all finite orbits \( \gamma \). Now, the result follows from Lemma 2.5. \( \square \)

Let \( O_T \) denote the set of all periodic \( T \)-orbits. It is clear that \( O_1 \subset O_T \), but a periodic \( T \)-orbit may not be an orbit of the flip system \( (X,T,F) \). We restate Theorem 2.6 as follows:

**Theorem 2.7.** Let \( R_{T,F} \) be the radius of convergence of the Maclaurin series of \( \zeta_{T,F} \), and suppose that \( R_{T,F} > 0 \). Then we have
\[
(2.5) \quad \zeta_{T,F}(t) = \prod_{\beta \in O_T} \sqrt{1 - t^{2|\beta|}} \prod_{\gamma \in O_1} \exp \left( \frac{t^{|\gamma|}}{1 - t^{|\gamma|}} \right) \quad (|t| < R_{T,F}).
\]
Proof. Since $O_1 \subset O_T$, we have
\[
\prod_{\beta \in O_T} \sqrt{\frac{1}{1 - t^{2|\beta|}}} = \prod_{\beta \in O_1} \sqrt{\frac{1}{1 - t^{2|\beta|}}} \prod_{\beta \in O_T \setminus O_1} \sqrt{\frac{1}{1 - t^{2|\beta|}}}.
\]
Then the right-hand side of (2.5) is equal to
\[
\prod_{\gamma \in O_1} \sqrt{\frac{1}{1 - t^{2|\gamma|}}} \exp \left( \frac{t|\gamma|}{1 - t|\gamma|} \right) \prod_{\beta \in O_T \setminus O_1} \sqrt{\frac{1}{1 - t^{2|\beta|}}}.
\]
In view of Theorem 2.6, we need only to prove the following:
\[
(2.6) \quad \prod_{\beta \in O_T \setminus O_1} \sqrt{\frac{1}{1 - t^{2|\beta|}}} = \prod_{\gamma \in O_2} \frac{1}{1 - t|\gamma|}.
\]
We note that if $\beta \in O_T \setminus O_1$, then $F\beta \in O_T \setminus O_1$, $\beta \cap F\beta = \emptyset$ and $\beta \cup F\beta \in O_2$. Conversely, if $\gamma \in O_2$, then there is an element $\beta_\gamma \in O_T \setminus O_1$ such that $\gamma = \beta_\gamma \cup F\beta_\gamma$. In this case, we have $|\gamma| = 2|\beta_\gamma| = 2|F\beta_\gamma|$. Thus
\[
\prod_{\beta \in O_T \setminus O_1} \sqrt{\frac{1}{1 - t^{2|\gamma|}}} = \prod_{\gamma \in O_2} \sqrt{\frac{1}{1 - t|\gamma|}} \sqrt{\frac{1}{1 - t^{2|F\beta_\gamma|}}} = \prod_{\gamma \in O_2} \frac{1}{1 - t|\gamma|}.
\]
This proves (2.6). \qed

Corollary 2.8. Let $G_{T,F}$ be as in Theorem 2.2. Then
\[
G_{T,F}(t) = \sum_{\gamma \in O_1} \frac{t|\gamma|}{1 - t|\gamma|}.
\]
Proof. The result is an immediate consequence of (1.3), Theorem 2.2 and the above theorem. \qed

3. Flips for topological Markov shifts.

Let $A$ be a finite discrete topological space. For $x \in A^\Z$ and $i \in \Z$ the $i$-th coordinate of $x$ is denoted by $x_i$, and if $i,j \in \Z$ with $i < j$, the block $x_i x_{i+1} \ldots x_j$ is denoted by $x_{[i,j]}$. For $x \in A^\Z$, we define $\sigma x$ and $\rho x$ by
\[
(\sigma x)_i = x_{i+1} \quad \text{and} \quad (\rho x)_i = x_{-i} \quad (i \in \Z).
\]
Then $\sigma$ and $\rho$ are homeomorphisms of $A^\Z$ onto itself, and satisfy
\[
\sigma \rho = \rho \sigma^{-1} \quad \text{and} \quad \rho^2 = \text{id},
\]
that is, \( \rho \) is a flip for the dynamical system \((A^Z, \sigma)\). This dynamical system is called the full \( A\)-shift. The map \( \sigma \) is called the shift map, and \( \rho \) the reverse map. When we express a point as a bi-infinite sequence, we will underline the 0-th coordinate. For instance, if \( x = \ldots x_{-1}x_0x_1x_2 \ldots \), then \( \sigma x = \ldots x_{-2}x_{-1}x_0x_1x_2 \ldots \) and \( \rho x = \ldots x_2x_1x_0x_{-1}x_{-2} \ldots \).

Let \( A \) be a 0-1, \( A \times A \) matrix, and \((X_A, \sigma_A)\) denote the topological Markov shift whose transition matrix is \( A \). If \( A = A^T \), then \( X_A \) is \( \rho \)-invariant, and hence \( \rho |_{X_A} \) is a flip for \((X_A, \sigma_A)\). More generally, if there is a 0-1, \( A \times A \) matrix \( P \) such that

\[
AP = PA^T \quad \text{and} \quad P^2 = I,
\]

then there is a flip, denoted by \( \phi_{A,P} \), for \((X_A, \sigma_A)\) that is defined as follows: Since \( P \) is 0-1 and \( P^2 = I \), it is a symmetric permutation matrix, that is, \( P = P^T \) and for each \( a \in A \) there is a unique \( a^* \in A \) such that \( P(a, a^*) = 1 \). Then it is easy to see that

\[
(a^*)^* = a \quad (a \in A)
\]

and

\[
A(a,b) = 1 \iff A(b^*, a^*) = 1 \quad (a, b \in A).
\]

For \( x \in X_A \) we define \( \phi_{A,P} x \) by

\[
(\phi_{A,P} x)_i = (x_{-i})^* \quad (i \in \mathbb{Z}).
\]

Then from (3.2) and (3.3) it follows that \( \phi_{A,P} \) is a flip for \((X_A, \sigma_A)\).

The following theorem states that every flip for a topological Markov shift can be represented in this way:

**Theorem 3.1 (Representation Theorem).** Let \((X, T, F)\) be a flip system, and suppose that \((X, T)\) is conjugate to a topological Markov shift. Then there are 0-1 square matrices \( A \) and \( P \) satisfying (3.1) such that \((X, T, F)\) is conjugate to \((X_A, \sigma_A, \phi_{A,P})\).

**Proof.** We suppose that \((X, T)\) is conjugate to a topological Markov shift \((X_M, \sigma_M)\) through a conjugacy \( \Psi \). Set \( \phi = \Psi F \Psi^{-1} \). Then this is a flip for \((X_M, \sigma_M)\), and \((X_M, \sigma_M, \phi)\) is conjugate to \((X, T, F)\). We will construct a finite set \( A \) and two 0-1, \( A \times A \) matrices \( A \) and \( P \) satisfying (3.1) such that \((X_M, \sigma_M, \phi) \cong (X_A, \sigma_A, \phi_{A,P})\).

Since \( \phi \) is continuous, there is a positive integer \( N \) such that

\[
x_{[-N,N]} = y_{[-N,N]} \Rightarrow (\phi x)_0 = (\phi y)_0 \quad (x, y \in X_M).
\]

For \( x \in X_M \) let \( \bar{x} \) denote the bi-infinite sequence defined by

\[
\bar{x} = \ldots (\phi x)_2(\phi x)_1(\phi x)_0(\phi x)_{-1}(\phi x)_{-2} \ldots,
\]

that is, \( \bar{x} = \rho \phi x \). It should be noted that if \( M \) is symmetric, then \( \bar{x} \in X_M \) for all \( x \in X_M \), but in general, this is not the case. For \( x \in X_M \) let \([x]\)
denote the ordered pair of the \((2N + 1)\)-blocks \(x_{[-N,N]}\) and \(\tilde{x}_{[-N,N]}\), and we express \([x]\) as

\[
[x] = \left[ \begin{array}{c} x_{-N} \ldots x_0 \ldots x_N \\ \tilde{x}_{-N} \ldots \tilde{x}_0 \ldots \tilde{x}_N \end{array} \right].
\]

Note that if \(x, y \in X_M\) and \(x_{[-2N,2N]} = y_{[-2N,2N]}\), then \([x] = [y]\).

Now we define \(A\). An ordered pair \(a = \left[ \begin{array}{c} a_{-N} \ldots a_0 \ldots a_N \\ \tilde{a}_{-N} \ldots \tilde{a}_0 \ldots \tilde{a}_N \end{array} \right]\) of \((2N + 1)\)-blocks is an element of \(A\) if and only if \(a = [x]\) for some \(x \in X_M\). It is clear that \(A\) is a finite set. For \(a = \left[ \begin{array}{c} a_{-N} \ldots a_0 \ldots a_N \\ \tilde{a}_{-N} \ldots \tilde{a}_0 \ldots \tilde{a}_N \end{array} \right] \in A\) we define

\[
a^* = \left[ \begin{array}{c} \tilde{a}_N \ldots \tilde{a}_0 \ldots \tilde{a}_{-N} \\ a_{-N} \ldots a_0 \ldots a_N \end{array} \right],
\]

\[
l(a) = \left[ \begin{array}{c} a_{-N} \ldots a_0 \ldots a_{N-1} \\ \tilde{a}_{-N} \ldots \tilde{a}_0 \ldots \tilde{a}_{N-1} \end{array} \right],
\]

\[
r(a) = \left[ \begin{array}{c} a_{-N+1} \ldots a_0 \ldots a_N \\ \tilde{a}_{-N+1} \ldots \tilde{a}_0 \ldots \tilde{a}_N \end{array} \right],
\]

\[
c(a) = a_{-N} \ldots a_0 \ldots a_N \quad \text{and}
\]

\[
b_0(a) = \tilde{a}_0.
\]

Obviously \([x]^* = [\phi x]\) for all \(x \in X_M\). Hence \(a^* \in A\) and \((a^*)^* = a\) for all \(a \in A\). Moreover, (3.4) implies that

\[
(3.5) \quad c(a) = c(b) \Rightarrow b_0(a) = b_0(b) \quad (a, b \in A).
\]

Next, define the matrices \(A\) and \(P\) by

\[
A(a, b) = \delta(r(a), l(b)) \quad (a, b \in A)
\]

and

\[
P(a, b) = \delta(a^*, b) \quad (a, b \in A),
\]

where \(\delta\) denotes the Kronecker delta. Then it is straightforward to check that \(A\) and \(P\) satisfy (3.1).

Finally, define \(\Phi : X_M \rightarrow X_A\) by

\[
(\Phi x)_i = [(\sigma_M)^i x] \quad (x \in X_M, \ i \in \mathbb{Z}).
\]

Then \(\Phi\) is an injective sliding block code of memory and anticipation \(2N\). Moreover a direct calculation shows that \(\Phi \phi = \phi_A \Phi\). It remains only to show that \(\Phi\) is surjective. Let \(y = \ldots a_{-2}a_{-1}a_0a_1a_2 \ldots \) be any point in \(X_A\). Then there is a point \(x \in X_M\) such that

\[
(3.6) \quad x_{[-N+i, N+i]} = c(a_i) \quad (i \in \mathbb{Z}).
\]
Let \( z = \Phi x \), and write \( z = \ldots \) \( b_{-2} b_{-1} b_0 b_1 b_2 \ldots \). Then from the definition of \( \Phi \), we have
\[
x_2^{[-N+i,N+i]} = c(b_i) \quad (i \in \mathbb{Z}).
\]
Hence \( b_0(a_i) = b_0(b_i) \) for all \( i \in \mathbb{Z} \) by (3.5), (3.6) and (3.7). This implies \( y = z \). □

Let \( \zeta_{A,P} \) be the zeta function of the flip system \( (X_A, \sigma_A, \phi_{A,P}) \). In Theorem 3.2 below, we express \( \zeta_{A,P} \) in terms of the matrices \( A \) and \( P \). It is well-known that the Artin-Mazur zeta function \( \zeta_A \) of the topological Markov shift \( (X_A, \sigma_A) \) satisfies
\[
\zeta_A(t) = \frac{1}{\det(I - tA)}.
\]
See Theorem 6.4.6 in [6].

We need some notations. For an \( A \times A \) matrix \( B \), the adjugate of \( B \) is denoted by \( B^\star \), so that \( BB^\star = (\det B)I \), the entry sum \( S[B] \) of \( B \) is defined by
\[
S[B] = \sum_{(a,b) \in A \times A} B(a,b),
\]
and the diagonal projection \( B^\Delta \) of \( B \) is defined by
\[
B^\Delta(a,b) = B(a,b)\delta(a,b) \quad (a,b \in A).
\]

**Theorem 3.2.** If \( A \) and \( P \) are 0-1, square matrices which satisfy (3.1), then
\[
\zeta_{A,P}(t) = \sqrt{\zeta_A(t^2)} \exp \left( \zeta_A(t^2)H_{A,P}(t) \right),
\]
where \( H_{A,P} \) is the polynomial defined by
\[
H_{A,P}(t) = S \left[ tP^\Delta(I - t^2A)^*(AP)^\Delta \right.
\]
\[
+ \frac{t^2}{2} \left\{ P^\Delta A(I - t^2A)^*P^\Delta + (PA)^\Delta(I - t^2A)^*(AP)^\Delta \right\} \right].
\]

**Proof.** For \( i, j \in \mathbb{Z} \) let \( p_{i,j} \) denote the number of points in \( X_A \) that are fixed by \( (\sigma_A)^i \) and \( (\sigma_A)^j \phi_{A,P} \). Set
\[
G_{A,P}(t) = \sum_{m=1}^{\infty} \left( p_{2m-1,0}t^{2m-1} + \frac{p_{2m,0} + p_{2m+1,0}}{2}t^{2m} \right).
\]
Then, in view of Theorem 2.2 and (3.8), we need only to prove the following:
\[
G_{A,P}(t) = \frac{H_{A,P}(t)}{\det(I - t^2A)}.
\]
Let $B_n$ denote the set of all $n$-blocks that occur in points in $X_A$. Then it is easy to see that
\[ \begin{align*}
p_{2m+1,0} &= \left| \{ x_0 \ldots x_m \in B_{m+1} : x_0^* = x_0, A(x_m, x_m^*) = 1 \} \right|, \\
p_{2m,0} &= \left| \{ x_0 \ldots x_m \in B_{m+1} : x_0^* = x_0, x_m^* = x_m \} \right|, \quad \text{and} \\
p_{2m,1} &= \left| \{ x_1 \ldots x_m \in B_m : A(x_1^*, x_1) = A(x_m, x_m^*) = 1 \} \right|.
\end{align*} \]
Recall that for $a \in B_1$, $a^*$ is the unique element of $B_1$ such that $P(a, a^*) = 1$. Moreover, for $a, b \in B_1$ the following are obvious:
\[ a^* = b \iff P(a, b) = 1, \]
\[ A(a, b^*) = 1 \iff AP(a, b) = 1, \quad \text{and} \]
\[ A(a^*, b) = 1 \iff PA(a, b) = 1. \]
Therefore we obtain
\[ (3.11) \quad p_{2m+1,0} = S \left[ P^\Delta A^m (AP)^\Delta \right], \]
\[ p_{2m,0} = S \left[ P^\Delta A^m P^\Delta \right], \quad \text{and} \]
\[ p_{2m,1} = S \left[ (PA)^\Delta A^{m-1} (AP)^\Delta \right]. \]
On the other hand, we have
\[ (3.12) \quad \sum_{m=0}^{\infty} s^m A^m = (I - sA)^{-1} = \frac{1}{\det(I - sA)} (I - sA)^+ \]
\[ (s \in \mathbb{C}, \Lambda |s| < 1), \]
where $\Lambda$ denotes the spectral radius of $A$. Finally, put (3.11) into (3.9), and use (3.12) to obtain (3.10).

4. Examples.

In order for a dynamical system $(X, T)$ to have a flip, it is necessary that $(X, T)$ is conjugate to its time reversal $(X, T^{-1})$. However, it is not known whether the condition is sufficient. The first example shows that there is a dynamical system with no flips.

Example 4.1. Let
\[ A = \begin{bmatrix} 19 & 5 \\ 4 & 1 \end{bmatrix}, \]
and $(X_A, \sigma_A)$ denote the edge shift of $A$. It is known that $A$ is not shift equivalent to its transpose $A^T$ [3, p. 104]. Hence $(X_A, \sigma_A)$ is not conjugate to its time reversal $(X_A, \sigma_A^{-1}) \cong (X_{A^T}, \sigma_{A^T})$. Consequently, $(X_A, \sigma_A)$ does not admit a flip.
In the remainder of this section, we consider various flips on full shifts. We show that some of them are not conjugate by calculating their zeta functions or counting the number of fixed points.

**Example 4.2.** Let \((X, \sigma)\) be the full 2-shift. We will show that there are infinitely many non-conjugate flips for \((X, \sigma)\). For each positive integer \(n\) we define the \((2n + 5)\)-block map \(K_n\) by
\[
K_n(1102n+11) = 1, \quad K_n(110n10n+11) = 0, \quad \text{and} \quad K_n(x_{n-2} \ldots x_0 \ldots x_{n+2}) = x_0 \text{ when the block is not equal to any of the above two.}
\]
Let \(\kappa_n\) denote the sliding block code on \(X\) induced by the block map \(K_n\). Then clearly \(\kappa_n\) is an automorphism of order 2. Let \(\omega_n = \rho \kappa_n\), where \(\rho\) is the reverse map. It is easy to see that \(\omega_n\) is a flip map for \((X, \sigma)\). The flip systems \((X, \sigma, \omega_n), n \geq 1\), are not conjugate to each other. In fact, for \(1 \leq n < m\),
\[
|\{x \in X : \sigma^{2m+5}x = x, \ \omega_n x = x\}| = 2^{m+3} - 2^{m-n+1},
\]
and
\[
|\{x \in X : \sigma^{2m+5}x = x, \ \omega_n x = x\}| = 2^{m+3} - 2.
\]
From this and Theorem 2.2, it also follows that \(\zeta_{\sigma, \omega_n}, n \geq 1\), are all distinct. A long but straightforward calculation using Theorem 3.2 yields that the zeta function for \((X, \sigma, \omega_1)\) is equal to
\[
\sqrt{\frac{1}{1 - 2t^2}} \exp \left( \frac{2t + 3t^2 - 2t^5 - 2t^6 + 2t^7 + 2t^{10} - 2t^{12} - 2t^{14}}{1 - 2t^2} \right).
\]

**Example 4.3.** Let \(n \geq 2\) be an integer, \((X, \sigma)\) the full \(n\)-shift, and \(\rho : X \to X\) the reverse map. As the zeta function is automorphism-invariant, the flip systems \((X, \sigma, \rho)\) and \((X, \sigma, \sigma \rho)\) have the same zeta function, which is
\[
\zeta_{\sigma, \rho}(t) = \sqrt{\frac{1}{1 - nt^2}} \exp \left( \frac{nt + (n + n^2)t^2/2}{1 - nt^2} \right).
\]
They are, however, not conjugate. In fact, we have
\[
|\{x \in X : \sigma^2x = x, \ \rho x = x\}| = n^2,
\]
whereas
\[
|\{x \in X : \sigma^2x = x, \ \sigma \rho x = x\}| = n.
\]

As we have seen in the above examples, a dynamical system may have many non-conjugate flip maps. However the following question still remains to be answered: Let \(A\) and \(B\) be symmetric 0-1 matrices such that \((X_A, \sigma_A) \cong (X_B, \sigma_B)\). Does it follow that \((X_A, \sigma_A, \rho_A) \cong (X_B, \sigma_B, \rho_B)\)?

**Example 4.4.** Let \((X, \sigma)\) be the full 2-shift, and \(\psi : X \to X\) defined by
\[
\psi(x) = \ldots x^5_2 x^4_1 x^3_0 x^2_{-1} x^1_{-2} \ldots,
\]
where $0^* = 1$ and $1^* = 0$. Then $\psi$ is a flip for $(X, \sigma)$. The flips $\psi$ and $\sigma\psi$ are not conjugate since $\psi$ has no fixed points but $\sigma\psi$ has fixed points. But they have the same zeta function

$$\zeta_{\sigma,\psi}(t) = \sqrt{\frac{1}{1 - 2t^2}} \exp\left(\frac{t^2}{1 - 2t^2}\right).$$

On taking $n = 2$ in Example 4.3, we know that $\rho$ and $\sigma\rho$ are not conjugate, and have the same zeta function

$$\zeta_{\sigma,\rho}(t) = \sqrt{\frac{1}{1 - 2t^2}} \exp\left(\frac{2t + 3t^2}{1 - 2t^2}\right).$$

Therefore the four flips $\rho, \sigma\rho, \psi$ and $\sigma\psi$ for $(X, \sigma)$ are not conjugate to each other.

**Example 4.5.** Let $A = \{0, 1, 2, 3\}$. Let $A$ and $P$ be 0-1, $A \times A$ matrices defined by $A(i,j) = 1$ for all $(i,j)$, and $P(i,j) = 1$ if and only if $(i,j) \in \{(0,0), (1,1), (2,3), (3,2)\}$. Then we find that the flip system $(X_A, \sigma, \phi_{A,P})$ has the zeta function

$$\zeta_{A,P}(t) = \sqrt{\frac{1}{1 - 4t^2}} \exp\left(\frac{2t + 4t^2}{1 - 4t^2}\right).$$

Now, we will show that the flips $\phi_{A,P}$ and $\sigma_A\phi_{A,P}$ for the full 4-shift $(X_A, \sigma_A)$ are conjugate. Let $X = \{0, 1\}^\mathbb{Z}$, $\sigma : X \to X$ the shift map, and $\rho : X \to X$ the reverse map. Let $\pi_1 : \{0,1\}^2 \ni ab \mapsto a \in \{0,1\}$, and $\pi_2 : \{0,1\}^2 \ni ab \mapsto b \in \{0,1\}$. Define $f : A \to \{0,1\}^2$ by $f(0) = 00$, $f(1) = 11$, $f(2) = 01$ and $f(3) = 10$, and $\Phi : X_A \to X$ by

$$\Phi(x) = \pi_1 f(x_0) \pi_2 f(x_1) \pi_1 f(x_1) \pi_2 f(x_1) \pi_1 f(x_1) \pi_2 f(x_1) \ldots.$$

We can easily check that $\Phi$ is a conjugacy from $(X_A, \sigma_A, \phi_{A,P})$ to $(X, \sigma^2, \sigma\rho)$, and so one from $(X_A, \sigma_A, \sigma_A\phi_{A,P})$ to $(X, \sigma^3, \sigma^3\rho)$. Trivially $\sigma$ is a conjugacy from $(X, \sigma^2, \sigma\rho)$ to $(X, \sigma^2, \sigma^3\rho)$. Therefore $\Phi^{-1}\sigma\Phi$ is a conjugacy from $(X_A, \sigma_A, \phi_{A,P})$ to $(X_A, \sigma_A, \sigma_A\phi_{A,P})$. This proves the assertion.

**Acknowledgment.** The authors would like to thank Professors R. Burton, K. H. Kim and F. Roush for their helpful comments in the preparation of this paper.

**References**


Received December 26, 2001. This research was supported by the Brain Korea 21 Project.

**School of Mathematical Sciences**
**Seoul National University**
**Seoul 151-747**
**Korea**
*E-mail address: kimyo@math.snu.ac.kr*

**Department of Mathematics**
**Ajou University**
**Suwon 442-749**
**Korea**
*E-mail address: jslee@madang.ajou.ac.kr*

**Department of Mathematics**
**Ajou University**
**Suwon 442-749**
**Korea**
*E-mail address: kkpark@madang.ajou.ac.kr*
ON SOME POINTWISE INEQUALITIES CONCERNING TENT SPACES AND SHARP MAXIMAL FUNCTIONS

Andrei K. Lerner

We consider an abstract analogue of $S^\#_\lambda f$, the truncated square function introduced by J.-O. Strömberg, and show that it is closely related to operators appearing in the theory of tent spaces. We suggest an approach to basic results for these spaces which differs from that due to R.R. Coifman, Y. Meyer and E.M. Stein. Also we discuss pointwise estimates involving $S^\#_\lambda f$ as well as different variants of sharp maximal functions.

1. Introduction.

Let $\mathbb{R}^{n+1}_+ = \{(y, t) : y \in \mathbb{R}^n, t > 0\}$ be the upper half space. For $\alpha, h > 0$ define the truncated cone $\Gamma^h_\alpha(x) = \{(y, t) \in \mathbb{R}^{n+1}_+ : |y - x| < \alpha t, 0 < t < h\}$. Let $\Gamma_\alpha(x) = \Gamma^\infty_\alpha(x)$. When $\alpha = 1$ we simply write $\Gamma^h(x), \Gamma(x)$. Given a ball $B = B(x, r)$ in $\mathbb{R}^n$, centered at $x$ of radius $r$, denote by $T(B)$ the tent over $B$, that is, $T(B) = \{(y, t) : |y - x| + t \leq r\}$. For two quantities $a, b$, we write $a \asymp b$ if there exist absolute constants $c_1, c_2$ such that $c_1 a \leq b \leq c_2 a$.

For any measurable function $F$ defined on $\mathbb{R}^{n+1}_+$, set

$$A(F|h)(x) = \int_{\Gamma^h(x)} |F(y, t)| \frac{dydt}{t^n+1},$$

and

$$C_F(x) = \sup_{B \ni x} \frac{1}{|B|} \int_{T(B)} |F(y, t)| \frac{dydt}{t},$$

where the sup is taken over all balls $B$ containing $x$. Let $AF = A(F|\infty)$; let for $q > 0$, $A_q F(x) = \left(A |F|^q(x)\right)^{1/q}$ and $C_q F(x) = \left(C |F|^q(x)\right)^{1/q}$. Set

$$A_\infty F(x) = \sup_{(y, t) \in \Gamma(x)} |F(y, t)|.$$

In [4], R.R. Coifman, Y. Meyer, and E.M. Stein introduced the tent spaces $T^p_q = T^p_q(\mathbb{R}^{n+1}_+)$. These spaces provide a very useful tool for solving problems in harmonic analysis. When $0 < p, q < \infty$ the space $T^p_q$ consists of all $F$ such that

$$\|F\|_{T^p_q} = \|A_q F\|_{L^p(\mathbb{R}^n)} < \infty.$$
The space $T^p_\infty (p < \infty)$ is defined as the space of continuous functions $F$ which have non-tangential boundary limits a.e., and such that
\[
\|F\|_{T^p_\infty} = \|A\infty F\|_{L^p(\mathbb{R}^n)} < \infty.
\]
The space $T^\infty_q (q < \infty)$ is defined by requiring that
\[
\|F\|_{T^\infty_q} = \|C_q F\|_{L^\infty(\mathbb{R}^n)} < \infty.
\]
Under the latter definition, the pairing
\[
\langle F, G \rangle = \int_{\mathbb{R}^{n+1}_+} F(y,t)G(y,t)\,dydt/t
\]
realizes $T^\infty_q$ as equivalent to the dual of $T^1_q$. Also the above pairing gives that the dual of $T^p_q$ is $T^p_{q'}$, where $1 \leq p < \infty$, $1 < q < \infty$, and, as usual, $1/p + 1/p' = 1$ and $1/q + 1/q' = 1$.

In this paper we show that, besides $AF$ and $CF$, in the study of the tent spaces the operator $A^\#_{q,\lambda} F$, defined by
\[
A^\#_{q,\lambda} F(x) = \sup_{B \ni x} \left( (A(F|_B))\chi_B \right)^*(\lambda|B) \quad (0 < \lambda < 1),
\]
is also of considerable interest. Here the sup is taken over all balls $B$ containing $x$, $r_B$ and $\chi_B$ denote the radius and the indicator function of $B$, respectively, and $f^*(t)$ denotes the standard non-increasing rearrangement of $f$. For $q > 0$ set $A^\#_{q,\lambda} F(x) = (A^\#_{\lambda} |F|^q(x))^{1/q}$. Note that the function $A^\#_{q,\lambda} F$ is particularly interesting if $q = 2$ and $F$ is the Poisson integral of $f$, $F = f \ast P_t(y)$. We denote such a function by $S^\#_{\lambda} f$ and consider it below in connection with some pointwise estimates for $A^\#_{q,\lambda} F$. (Note that the function $S^\#_{\lambda} f$ was introduced and considered by J.-O. Strömberg in [14].)

The paper is organized as follows: In Section 2 we state and discuss our main results which we prove in Section 4. Section 3 contains some additional notation and auxiliary results.

2. Main results: Formulations and discussion.

Let us discuss our main results and their applications to the theory of the tent spaces.

2.1. Relations between the operators $A_q F, C_q F$ and $A^\#_{q,\lambda} F$. It is shown in [4] that the operator $CF$ can be used to give an alternate definition of $T^p_q$ when $0 < q < p < \infty$, since in this case
\[
\|A_q F\|_p \asymp \|C_q F\|_p.
\]

Roughly speaking, our first theorem investigates the relations between $A_q F, C_q F$ and $A^\#_{q,\lambda} F$, and, in particular, allows us to define the spaces $T^p_q$ in terms of $A^\#_{q,\lambda} F$ for all $0 < q < \infty$, $0 < p \leq \infty$. 

Theorem 2.1. Let $F$ be any measurable function defined on $\mathbb{R}^{n+1}$.

a) For all $0 < q < \infty$ and $x \in \mathbb{R}^n$,
\[ C_q F(x) \preceq M_q A_{\lambda,q}^\# F(x) \quad (0 < \lambda < 1), \]
where $M_q f = (M|f|^q)^{1/q}$ and $M$ is the Hardy-Littlewood maximal function.

b) For all $1 \leq q < \infty$ and $t > 0$,
\[ (A_q F)^\ast(t) \leq c_1 \int_{c_2 t}^{\infty} (A_{q,\lambda}^\# F)^\ast(\tau) \frac{d\tau}{\tau} \quad (0 < \lambda < 1), \]
where $c_1, c_2$ depend only on $\lambda$ and $n$.

Some comments about these results are in order. For any measurable function $f$ on $\mathbb{R}^n$ consider the maximal function $m_\lambda f$ defined by
\[ m_\lambda f(x) = \sup_{B \ni x} (f \chi_B)^\ast(\lambda |B|) \quad (0 < \lambda < 1). \]
Since \( \{m_\lambda f > \alpha\} \subset \{M \chi_{\{|f|>\alpha\}} \geq \lambda\} \), we have \((m_\lambda f)^\ast(t) \leq f^\ast(c_{p,n} t)\), and hence, for all $p > 0$,
\[ \|m_\lambda f\|_p \leq c_{p,\lambda,n} \|f\|_p. \]
Clearly, $A_{q,\lambda}^\# F(x) \leq m_\lambda (A_q F(x))$, and so, \(\|A_{q,\lambda}^\# F\|_p \leq c_{p,\lambda,n} \|A_q F\|_p\) for all $p, q > 0$. Using the duality argument, one can show that the converse also holds. However, the combining of (3) and classical Hardy’s inequality (see, e.g., [2, p. 124]) immediately gives a direct proof of this fact. Moreover, taking into account (2), we obtain that one can characterize the spaces $T^p_q$ in terms of $A_{\lambda,q}^\# F$:
\[ \|F\|_{T^p_q} \propto \|A_{\lambda,q}^\# F\|_p \quad (0 < q < \infty, 0 < p \leq \infty, 0 < \lambda < 1). \]
Inequality (1) is the other corollary of this theorem, which follows from (2), (3) and the Hardy-Littlewood maximal theorem.

Define the maximal function $M_\lambda^\# f$ by
\[ M_\lambda^\# f(x) = \sup_{Q \ni x} \inf_{c \in \mathbb{R}} ((f - c) \chi_Q)^\ast(\lambda |Q|) \quad (0 < \lambda < 1), \]
where the sup is taken over all cubes $Q$ containing $x$. In [11], the following rearrangement inequality was obtained for all $t > 0$ and any measurable function $f$ with $f^\ast(+\infty) = 0$:
\[ f^\ast(t) \leq \frac{2}{\log 2} \int_t^{\infty} (M_\lambda^\# f)^\ast(\tau) \frac{d\tau}{\tau} \quad (0 < \lambda < \lambda_n). \]
This will be a key tool in proving (3).

It is worth noting that the function $M_\lambda^\# f$ was also introduced in the above mentioned paper [14], where its definition was motivated by an alternate
characterization of the space $BMO$. Recall that $BMO$ consists of all locally integrable functions $f$ on $\mathbb{R}^n$ such that
\[
\|f\|_* = \sup_Q \frac{1}{|Q|} \int_Q |f(x) - f_Q|\,dx < \infty \quad (f_Q = \frac{1}{|Q|} \int_Q f).
\]

By Chebyshev’s inequality, it is clear that $\lambda\|M^\#_\lambda f\|_\infty \leq \|f\|_*$. However, F. John [10] for $0 < \lambda < 1/2$ and J.-O. Strömberg [14] for $\lambda = 1/2$ proved that the converse is also true:
\[
\|f\|_* \leq c_n\|M^\#_\lambda f\|_\infty \quad (0 < \lambda \leq 1/2).
\]

2.2. Real interpolation of tent spaces. It is proved in [4] that
\[
(T^\theta_{q_0}, T^\theta_{q_1})_\theta,p = T^p_q \quad (1 < q < \infty),
\]
where $0 < \theta < 1$, $1 \leq p_0 < p_1 \leq \infty$ and $1/p = (1 - \theta)/p_0 + \theta/p_1$, and $(\cdot, \cdot)_{\theta,p}$ is the real method of interpolation (cf. [2, p. 299]). A different proof of (7) is given by J. Alvarez and M. Milman [1]. Generally speaking, both proofs consist of two main steps: First (7) is proved under certain constraints on the interval for $p_0, p_1$ (without end-points), then the result is extended to the whole range of $p_0, p_1$ by the duality and Wolff’s reiteration theorem.

Here we use our approach to prove (7) via sharp estimates for the $K$-functional. Namely, in our next result we state that in the end-point case Peetre’s $K$-functional for the couple $(T^1_q, T^\infty_q)$ is explicitly characterized in terms of $A^\#_{q, \lambda} F$. Our proof works for all $0 < q < \infty$, so (7) can be extended to the case $0 < q \leq 1$. Recall that the $K$-functional is defined as
\[
K(F, t; T^1_q, T^\infty_q) = \inf_{F = F_1 + F_2} \left( \|F_1\|_{T^1_q} + t\|F_2\|_{T^\infty_q} \right)
\]
for all $F \in T^1_q + T^\infty_q$ and $t > 0$.

**Theorem 2.2.** Let $0 < q < \infty$. Then for all $F \in T^1_q + T^\infty_q$ and $t > 0$,
\[
K(F, t; T^1_q, T^\infty_q) \asymp \int_0^t (A^\#_{q, \lambda} F)^*(\tau)d\tau \quad (0 < \lambda < 1).
\]

This theorem along with (5) easily implies (7) in the case $p_0 = 1, p_1 = \infty$ for all $0 < q < \infty$. Now one can apply the Holmstedt reiteration theorem (see, e.g., [2, p. 307]) to describe the $K$-functional for any of the couples $(T^p_q, T^{p'}_q)$ and get (7) for all $1 < p_0 < p_1 < \infty$.

2.3. Factorization of tent spaces. For the tent spaces the following factorization holds:
\[
T^p_q = T^\infty \cdot T^p_q.
\]
This fact was proved in [4] in the case $p > 2$ and $q = 2$. Recently, W.S. Cohn and I.E. Verbitsky [3] have extended (8) to all $0 < p, q < \infty$. This result is partially based on the next inequality [3]:

$$\|FG\|_{T_p^q} \leq c_{p,q} \|F\|_{T_p^\infty} \|G\|_{T_q^\infty}. \quad (9)$$

We propose a different proof of (9). It follows immediately from (4), (5) and the next elementary estimate.

**Proposition 2.3.** For any functions $F, G$ defined on $\mathbb{R}^{n+1}_{++}$ and for all $x$,

$$A_{q,\lambda}^\#(FG)(x) \leq m_{\lambda/4}(A_\infty F)(x)A_{q,\lambda/4}^\# G(x) \quad (0 < \lambda < 1).$$

2.4. **Pointwise estimates for $S_\lambda^\# f$.** Let us return to the function $S_\lambda^\# f$ and discuss several pointwise estimates motivated by (2) and the well-known C. Fefferman’s duality theorem. We consider the definition of $S_\lambda^\# f$ in the following slightly generalized form. Let $\varphi$ be a real-valued differentiable function on $\mathbb{R}^n$ which satisfies:

(i) $|\varphi(x)| \leq c(1 + |x|)^{-n-1}$, $|\nabla \varphi(x)| \leq c(1 + |x|)^{-n-1}$,

(ii) $\int_{\mathbb{R}^n} \varphi(x)dx = 0$.

Write $\varphi_t(x) = \varphi(x/t)t^n, t > 0$. Given a function $f$ with

$$\int_{\mathbb{R}^n} |f(x)|(1 + |x|)^{-n-1}dx < \infty,$$

set $F(y, t) = |f \ast \varphi_t(y)|$, and define $S_\lambda^\# f$ by (cf. [14])

$$S_\lambda^\# f(x) = A_{2,\lambda}^\# F(x).$$

Denote by $\mathcal{S}(\mathbb{R}^n)$ the class of Schwartz functions on $\mathbb{R}^n$. Assume that, in addition to (i) and (ii), $\varphi$ satisfies

(iii) there exists a function $\psi \in \mathcal{S}(\mathbb{R}^n)$ such that

$$\int_{\mathbb{R}^n} \psi(x)dx = 0, \quad \text{supp} \psi \subset \{|x| \leq 1\} \quad \text{and}$$

$$\int_0^\infty \hat{\varphi}(t\xi)\hat{\psi}(t\xi)\frac{dt}{t} \equiv 1 \quad \text{for all} \ \xi \neq 0. \quad (11)$$

In particular, $\varphi$ satisfies (iii) whenever $\varphi$ is radial and $\hat{\varphi}(\xi) \geq 0$ [13, p. 186]. Define the maximal function $\mathcal{F}^\# f$ by

$$\mathcal{F}^\# f(x) = C_2 F(x),$$

where, as above, $F(y, t) = |f \ast \varphi_t(y)|$. If $\varphi$ satisfies (i)-(iii), then one of the equivalent formulations of C. Fefferman’s duality theorem (see [5] and [6] or [13, p. 159]) states that $f \in BMO \iff \mathcal{F}^\# f \in L^\infty$ and

$$\|f\|_* \asymp \|\mathcal{F}^\# f\|_\infty.$$
By (2), we see that $\mathcal{F}^# f(x) \asymp M_2 S^# f(x)$, and therefore $\|f\|_\ast \asymp \|S^# f\|_\infty$. In view of (3), we also obtain that for all $p > 0$,

$$
\|S f\|_p \asymp \|S^# f\|_p \quad (0 < \lambda < 1),
$$

where $S f$ is the Lusin area integral. Hence,

$$
\|f\|_{H^p} \asymp \|S^# f\|_{L^p} \quad (0 < \lambda < \lambda_n).
$$

This was proved by B. Jawerth [8]. His proof was based on atomic decomposition. Note that the following characterization:

$$
K(f, t; H^1, BMO) \asymp \int_0^t (S^# f)^*(\tau) d\tau
$$

was also established in [8]. It is interesting to compare this result with Theorem 2.2.

Observe that inequality (2) may be viewed as an analogue of the following two-sided estimate proved by B. Jawerth and A. Torchinsky [9]:

$$
(f)^#(x) \asymp M_q M^# f(x) \quad (0 < \lambda < \lambda_n),
$$

whenever $q > 0$ and $f \in L^q + BMO$, where

$$
(f)^#(x) = \sup_{Q \ni x} \left( \frac{1}{|Q|} \int_Q |f(y) - f_Q|^q dy \right)^{1/q}.
$$

A natural question arises from this: What is a pointwise connection between functions $M^# f$ and $S^# f$? The following estimate answers the question in one direction:

$$
S^# f(x) \leq c f^#(x).
$$

In essence, it was proved in [14]. It is clear that the reverse inequality fails (e.g., for $f \in H^p$, $p \leq 1$). Nevertheless, using the quasi-orthogonal decomposition of $f$ [13, p. 166], we prove:

**Theorem 2.4.** Suppose $f$ satisfies (10) and $\varphi$ satisfies (i)-(iii). Then for all $x \in \mathbb{R}^n$,

$$
M^# f(x) \leq c S^# f(x) \quad (0 < \lambda < \lambda_n),
$$

where $c$ depends on $\lambda$ and $n$.

This inequality also can not be reversed (e.g., for $f \in L^p \setminus H^p$, $p \leq 1$). However, using inequalities (2), (12-14), we obtain the following “pointwise” version of C. Fefferman’s theorem:

**Corollary 2.5.** Let $f$ satisfies (10) and $\varphi$ satisfies (i)-(iii). Then

$$
\mathcal{F}^# f(x) \asymp M_2 S^# f(x) \asymp M_2 M^# f(x) \asymp (f^#)^2(x) \quad (0 < \lambda < \lambda_n).
$$
3. Preliminaries.

We say that $f^*$ is the non-increasing rearrangement of $f$ if it is non-increasing on $(0, +\infty)$ and equimeasurable with $|f(x)|$. We shall assume that the rearrangement is left-continuous. Then it is uniquely determined and can be defined by the equality

$$f^*(t) = \sup_{|E|=t} \inf_{x \in E} |f(x)|.$$

Throughout the paper, we shall use the following simple inequality (see, e.g., [2, p. 41]):

$$(f + g)^*(t_1 + t_2) \leq f^*(t_1) + g^*(t_2).$$

We will prove one more property of rearrangements, though it is apparently known. For any measurable set $E \subset \mathbb{R}^n$ we shall denote its complement $\mathbb{R}^n \setminus E$ by $E^c$.

**Proposition 3.1.** Let $\alpha + \beta < t$. Then for any measurable functions $f, g$,

$$(fg)^*(t) \leq f^*(\alpha)g^*(\beta).$$

**Proof.** Let $E_1 = \{x : |f(x)| \leq f^*(\alpha)\}, E_2 = \{x : |g(x)| \leq g^*(\beta)\}$. Then $|E_1^c \cup E_2^c| \leq \alpha + \beta$. Thus, for any measurable set $E \subset \mathbb{R}^n$ with $|E| = t$ we have $|E \cap (E_1 \cap E_2)| > 0$, and so

$$\inf_{x \in E} |fg(x)| \leq f^*(\alpha)g^*(\beta),$$

which completes the proof. \(\square\)

Now, let us define

$$A'(F|h)(x) = \int_{\Gamma^2(x)} |F(y,t)| \frac{dydt}{t^{n+1}}.$$

We shall need two following lemmas:

**Lemma 3.2.** Let $\mathcal{F} \subset \mathbb{R}^n$ be an arbitrary closed set whose complement $\mathcal{F}^c$ has finite measure. There is a subset $\mathcal{F}^* \subset \mathcal{F}$ such that $|\mathcal{F}^*| \leq c_n|\mathcal{F}^c|$, and for any non-negative $F$ the following inequality holds:

$$\int_{\mathcal{F}^* \cap \Gamma^2(x)} F(y,t)t^n dy dt \leq c'_n \int_{\mathcal{F}^*} \left( \int_{\Gamma(x)} F(y,t) dy dt \right) dx.$$

This result is well-known (see [4] or [13, p. 126]).

**Lemma 3.3.** For any ball $B$ containing $x$, we have

$$
\left( (A'(F|r_B)) \chi_B \right)^* (\lambda|B|) \leq c_1 A^{s\#}_{c_2 \lambda} F(x) \quad (0 < \lambda < 1),
$$

where $c_1$ depends on $\lambda$ and $n$, while $c_2$ depends only on $n$. 
Proof. Set $D = \bigcup_{x \in B} \Gamma_{2B}^B(x)$ and

$$\mathfrak{F} = \{x : A(F\chi_D)(x) \leq (A(F\chi_D))^*(\lambda |B|/2c_n)\},$$

where $c_n$ is the same as in Lemma 3.2. A simple geometric argument shows that $A(F\chi_D)$ is supported in $8B$. Thus $(A(F\chi_D))^*(\lambda |B|/2c_n) \leq A_{c\lambda}^F(x)$. Let $E \subset B$ be an arbitrary measurable set with $|E| = \lambda |B|$. Choose $\mathfrak{F}^* \subset \mathfrak{F}$ as in Lemma 3.2. Then $|\mathfrak{F}^*| \leq c_n|\mathfrak{F}| \leq \lambda |B|/2$, and hence $|\mathfrak{F}^* \cap E| \geq \lambda |B|/2$. Applying Lemma 3.2 and Fubini’s theorem, we get

$$\inf_{\xi \in E} A'(F|_B)(\xi) \leq \inf_{\xi \in \mathfrak{F}^* \cap E} A'(F\chi_D)(\xi) \leq \frac{2}{\lambda |B|} \int_{\mathfrak{F}^* \cap E} A'(F\chi_D)(\xi) d\xi$$

$$\leq \frac{c}{|B|} \int_{\xi \in \mathfrak{F}^* \cap \Gamma_{2}(\xi)} |F\chi_D(y,t)| \frac{dydt}{t}$$

$$\leq \frac{c}{|B|} \int_{\mathfrak{F}} A(F\chi_D)(\xi) d\xi \leq c_1 A_{c_2\lambda}^F(x).$$

To complete the proof take the sup over all $E \subset B$ with $|E| = \lambda |B|$. \qed

Also in this section we recall some useful ideas when dealing with $S_{\lambda}^# f$. Suppose $f$ satisfies (10) and $\varphi$ satisfies (i)-(ii). Set

$$S_h f(x) = \left( \int_{\Gamma_h(x)} |f * \varphi_t(y)|^2 \frac{dydt}{t^{n+1}} \right)^{1/2}, \quad S f(x) = S_{\infty} f(x).$$

Let $x \in B$ and let $f_1 = f\chi_{4B}$, $f_2 = f\chi_{(4B)^c}$. Since $S$ is of weak type $(1,1)$ (see [12]), we have

$$((S_{rB} f_1)\chi_B)^*(\lambda |B|) \leq (Sf_1)^*(\lambda |B|) \leq \frac{c}{\lambda |B|} \int_{4B} |f|. \tag{15}$$

Further, standard arguments (see, e.g., [13, p. 160]) yield

$$|f_2 * \varphi_t(y)| \leq c \int_{\mathbb{R}^n \setminus 4B} |f(\xi)| \frac{t}{(r_B + |x - \xi|)^{n+1}} d\xi,$$

whenever $x \in B$ and $(y,t) \in \bigcup_{\eta \in B} \Gamma_{r_B}(\eta)$, and hence,

$$((S_{rB} f_2)\chi_B)^*(\lambda |B|) \leq c \left( \frac{1}{|B|} \int_{T(3B)} (t/r_B)^2 \frac{dydt}{t} \right)^{1/2} \int_{\mathbb{R}^n \setminus 4B} |f(\xi)| \frac{r_B}{(r_B + |x - \xi|)^{n+1}} d\xi$$

$$\leq c \int_{\mathbb{R}^n \setminus 4B} |f(\xi)| \frac{r_B}{(r_B + |x - \xi|)^{n+1}} d\xi. \tag{16}$$

To extend (14) from Schwartzian functions to those satisfying (10) in the proof of Theorem 2.4, we use that $\int_{\mathbb{R}^n} |f_1(\xi)|(1 + |\xi|)^{-n-1} d\xi \to 0$ implies $(S_{rB} f_j)^*(\lambda |B|) \to 0$ as $j \to \infty$. This readily follows from (15) and (16).
Observe also that taking \( f_1 = (f - f_{AB})\chi_{AB}, f_2 = (f - f_{AB})\chi_{(AB)^c} \) yields that the right sides of (15) and (16) are at most \( c_{\lambda,n}f^\#(x) \). Since \( \int \varphi = 0 \), we have \( S_{r_B}f \leq \sqrt{2}(S_{r_B}f_1 + S_{r_B}f_2) \), and, by (15) and (16),
\[
((S_{r_B}f)\chi_{B})^* (\lambda|B|) \leq \sqrt{2}((S_{r_B}f_1)\chi_{B})^* (\lambda|B|/2) \\
+ \sqrt{2}((S_{r_B}f_2)\chi_{B})^* (\lambda|B|/2) \leq c_{\lambda,n}f^\#(x),
\]
which gives (13).

4. Proofs of the main results.

First of all, let us show that, as we mentioned after (4), the inequality
\[
\|A_qF\|_p \leq c_{\lambda,p,q}\|A_q^\#F\|_p \quad (0 < p, q < \infty, 0 < \lambda < 1)
\]
can be derived by duality. Suppose first that \( 1 \leq p < \infty, \lambda < q < \infty \). We use the following standard argument: For \( q > 0 \) define the “stopping-time” \( h(x) \) by
\[
h(x) = \sup\{h > 0 : A_q(F|h)(x) \leq A_q^\#F(x)\}.
\]
Then \( A_q(F|h(x))(x) \leq A_q^\#F(x) \) and \( \{|x \in B : h(x) > r_B\}| \geq (1 - \lambda)|B| \) for each \( B \). By Fubini’s theorem and Hölder’s inequality we have
\[
\int_{R^n+} |F(y,t)G(y,t)| \frac{dydt}{t} \leq 1 - \lambda \int_{R^n} \int_{\Gamma^h(x)} |F(y,t)G(y,t)| \frac{dydt}{t^n+1} dx \\
\leq 1 - \lambda \int_{R^n} A_q(F|h(x))(x)A_qG(x) dx \\
\leq 1 - \lambda \int_{R^n} A_q^\#F(x)A_qG(x) dx
\]
(note that in [4] different variants of such inequality were obtained with \( C_qF \) or \( A_qF \) instead of \( A_q^\#F \) on the right side). Applying Hölder’s inequality again, we get
\[
\int_{R^n+} |F(y,t)G(y,t)| \frac{dydt}{t} \leq \frac{1}{1 - \lambda} \|G\|_{T_{p',q'}} A_q^\#F \|_{p},
\]
which gives (17) by duality. The restrictions on \( p, q \) are easily removed by replacing \( |F| \) by \( |F|^\delta \) with appropriate \( \delta > 0 \).

Proof of the first part of Theorem 2.1. It is clear that it suffices to consider the case \( q = 1 \). Let \( q = 1 \) and let \( h \) as above. Then for \( 0 < t < r_B, y \in B \) we get
\[
\{|x \in 3B : (y,t) \in \Gamma^h(x)\}| \geq c_{n}(1 - \lambda)t^n,
\]
where \( c_n \) denotes the volume of the unit ball in \( \mathbb{R}^n \). Applying Fubini’s theorem gives

\[
\int_{T(3B)} |F(y,t)| \frac{dy \, dt}{t} \leq \frac{1}{c_n(1-\lambda)} \int_{3B} A(F|h)(x) \, dx \\
\leq \frac{1}{c_n(1-\lambda)} \int_{3B} A^\lambda F(x) \, dx,
\]

and thus \( CF(x) \leq c_{\lambda,n} MA^\lambda F(x) \).

Let us prove the converse. Let \( x, \xi \in B \) and let \( B' \) be an arbitrary ball containing \( \xi \). If \( B \subset 3B' \), then

\[
\left( (A(F|_{r_{B'}})) \chi_{B'} \right)^*(\lambda|B') \leq \frac{1}{\lambda|B'|} \int_{B'} A(F|_{r_{B'}})(x) \, dx \\
\leq \frac{c}{|B'|} \int \int |F(y,t)| \frac{dy \, dt}{t} \leq c CF(x),
\]

since \( \bigcup_{x \in B'} \Gamma_{r_{B'}}(x) \subset T(3B') \). Assume now that \( B \not\subset 3B' \). Then \( B' \subset 3B \) and in this case

\[
\left( (A(F|_{r_{B'}})) \chi_{B'} \right)^*(\lambda|B') \leq m_{\lambda} \left( (A(F|_3r_{B})) \chi_{3B} \right)(\xi).
\]

Therefore, for all \( \xi \in B \),

\[
A^\# F(\xi) \leq c CF(x) + m_{\lambda} \left( (A(F|_3r_{B})) \chi_{3B} \right)(\xi).
\]

Using (4), we get

\[
\frac{1}{|B|} \int_B A^\# F(\xi) \, d\xi \leq c CF(x) + \frac{1}{|B|} \left\| m_{\lambda} \left( (A(F|_3r_{B})) \chi_{3B} \right) \right\|_1 \\
\leq c CF(x) + \frac{c}{|B|} \int_{3B} A(F|_{3r_{B}})(x) \, dx \leq c CF(x).
\]

Hence \( MA^\# F(x) \leq c_{\lambda,n} CF(x) \), and (2) is proved.

**Proof of the second part of Theorem 2.1.** Choose a function \( \Phi \in S(\mathbb{R}^n) \) such that \( \chi_{B(0,1)} \leq \Phi \leq \chi_{B(0,3/2)} \), and define

\[
\tilde{A}(F)(x) = \int_0^\infty \int_{\mathbb{R}^n} |F(y,t)| \Phi_t(x - y) \, dy \, dt.
\]

Clearly, \( A(F)(x) \leq \tilde{A}(F)(x) \). The crucial observation to prove (3) is that

\[
M^\#_\lambda (\tilde{A}(F))(x) \leq c A^\#_{\epsilon,\lambda} F(x) \quad (0 < \lambda < 1).
\]
Let \( x \in Q, d_Q \) and \( x_Q \) denote the diameter and center of \( Q \) respectively, and 
\[
c_Q = \int_{2d_Q}^{\infty} \int_{\mathbb{R}^n} |F(y,t)|\Phi_t(x_Q - y) \frac{dydt}{t}.
\]

Then, for \( z \in Q \) we have 
\[
\left| \int_{2d_Q}^{\infty} \int_{\mathbb{R}^n} |F(y,t)|\Phi_t(z - y) \frac{dydt}{t} - c_Q \right| 
\leq c d_Q \sum_{k=0}^{\infty} \int_{2^{k+1}d_Q}^{2^{k+2}d_Q} \int_{B(z,3t/2) \cup B(x_Q,3t/2)} |F(y,t)| \frac{dydt}{t^{n+2}}.
\]

Note that for all \( z \in Q, \xi \in 2^kQ \) and \( t \geq 2^{k+1}d_Q \) we get \( |z - \xi| \leq 2^k d_Q \leq t/2 \), and hence \( B(z,3t/2) \cup B(x_Q,3t/2) \subset B(\xi,2t) \). It follows from this and Lemma 3.3 that the right side of (19) is at most 
\[
c \sum_{k=0}^{\infty} \frac{1}{2^k} \inf_{\xi \in 2^kQ} \int_{2^{k+2}Q(\xi)} |F(y,t)| \frac{dydt}{t^{n+1}} 
\leq c \sum_{k=0}^{\infty} \frac{1}{2^k} \left( (A'(F|2^{k+2}d_Q))\chi_{2^kQ} \right)^{2^k} (2^k |Q|) \leq c A_{x_n}^{\#} F(x).
\]

Therefore, using Lemma 3.3 again, we get 
\[
\left( (\tilde{A}(F) - c_Q)\chi_Q \right)^{\#} (\lambda |Q|) 
\leq \left( (A'(F|2d_Q))\chi_Q \right)^{\#} (\lambda |Q|) + c A_{x_n}^{\#} F(x) \leq c A_{x_n}^{\#} F(x),
\]
which proves (18).

Let \( A_q F = (\tilde{A}(|F|^q))^{1/q}, q \geq 1 \). Applying (6), (18) and the fact that 
\( M_{\lambda}^{\#} (|f|^{1/q})(x) \leq (M_{\lambda}^{\#} f)^{1/q}(x) \), we obtain 
\[
(A_q F)^*(t) \leq (\tilde{A}_q F)^*(t) \leq c \int_t^{\infty} (A_{q,\lambda}^{\#} F)^*(\tau) \frac{d\tau}{\tau} \quad (0 < \lambda < \lambda_n)
\]
provided \( (\tilde{A}F)^*(+\infty) = 0 \). However, if \( F \) is compactly supported, then \( \tilde{A}_q F \) is also compactly supported, and so \( (\tilde{A}_q F)^*(+\infty) = 0 \). Hence this latter assumption can be removed by taking an increasing sequence of functions \( F_k \uparrow F \) with compact support, and using the fact that \( |f_k| \uparrow |f| \) implies \( f_k^*(t) \uparrow f^*(t) \) (see [2, p. 41]).

It remains to prove (3) for \( \lambda_n < \lambda < 1 \). This follows immediately from (21) and the next lemma, which will also be crucial in the Proof of Theorem 2.2.

Define the \( E \)-functional by 
\[
E_q(F, t) = \inf_{D} \|C_q(F_{X_D})\|_{\infty},
\]
where the infimum is taken over all sets \( D \subset \mathbb{R}^{n+1}_+ \) with \( \text{supp} A_q (F_{X_D}) \leq t \).
Lemma 4.1. For any $0 < \eta, \lambda < 1$ and $0 < q < \infty$ we have
\[
c_1 (A_{q,\eta}^\# F)^*(c't/\eta) \leq E_q(F, t) \leq c_2 (A_{q,\lambda}^\# F)^*(c''t) \quad (t > 0),
\]
with constants $c', c''$ depending only on $n$, and constants $c_1, c_2$ depending on $n, \eta, \lambda$ and $q$.

Proof. Since $(A_{q,\eta}^\# F)^*(t) \leq (m_q A_q F)^*(t) \leq (A_q F)^*(t\eta/\gamma_n)$ and, by (2),
\[
\|A_{q,\eta}^\# F\|_{\infty} \leq c_n \|\mathcal{C}_q F\|_{\infty},
\]
for any $D$ with $|\text{supp} A_q (F \chi_{D^c})| \leq t$ and $c' = 3\gamma_n$ we have
\[
(A_{q,\eta}^\# F)^*(c't/\eta) \leq \left( \left( A_{q/2}^\# (|F|^q \chi_{D^c}) + A_{q/2}^\# (|F|^q \chi_D) \right)^*(c't/\eta) \right)^{1/q} \\
\leq \left( (A(|F|^q \chi_{D^c}))^*(3t/2) + c \|\mathcal{C}(|F|^q \chi_D)\|_\infty \right)^{1/q} \\
= c^{1/q} \|\mathcal{C}_q (F \chi_D)\|_{\infty}.
\]

To prove the converse, define $\Omega = \{ x : A_{q,\lambda}^\# F(x) > (A_{q,\lambda}^\# F)^*(c_n t) \}$, where $c_n$ is chosen so that $|\Omega| = \{ x : M \chi_\Omega(x) > 1/3^n \}$ is at most $t$. Now set
\[
D = \bigcup_{x \in \Omega} \Gamma(x),
\]
and thus $\supp A_q (F \chi_{D^c}) = \Omega$.

Let $B$ be an arbitrary ball. If $B \cap \Omega^c \neq \emptyset$, then, obviously,
\[
(A_q (F \chi_D |r_B)) (\lambda |B)) \leq (A_{q,\lambda}^\# F)^*(c_n t).
\]
If $B \subset \Omega$, then $3B \subset \Omega$, and thus $T(3B) \subset D^c$. Since $\cup_{x \in B} \Gamma(r_B(x) \subset T(3B)$, we get that $(A_q (F \chi_D |r_B)) (\lambda |B)) = 0$. Hence, by (2),
\[
\|\mathcal{C}_q (F \chi_D)\|_{\infty} \leq c \|A_{q,\lambda}^\# (F \chi_D)\|_{\infty} \leq c (A_{q,\lambda}^\# F)^*(c_n t).
\]
The lemma is proved. \hfill $\square$

Clearly, this lemma along with (21) implies (3) for any $0 < \lambda < 1$. Thus, the proof of Theorem 2.1 is complete. \hfill $\square$

Remark 4.2. Note that in [7] a Fefferman-Stein sharp function estimate of $A(F)$ was obtained to give an alternate proof of (1).

Proof of Theorem 2.2. We have to prove only that
\[
K(F, t, T_q^1, T_q^\infty) \leq c \int_0^t (A_{q,\lambda}^\# F)^*(\tau) d\tau
\]
since the proof of the converse is easily follows from (5). Choose $D$ as in the proof of Lemma 4.1. Set $F_1 = F \chi_{D^c}, F_2 = F \chi_D$. Then we have
\[
\|F_2\|_{T_q^\infty} \leq c (A_{q,\lambda}^\# F)^*(c_n t) \quad \text{and} \quad |\text{supp} A_q (F_1)| \leq t.
\]
Thus $c \int_0^t (A_{q,\lambda}^\# F)^*(\tau) d\tau$. and hence (see [2, p. 53]),
\[
\|A_{q,\lambda}^\# (F_1)\|_1 \leq c' \int_0^t (A_{q,\lambda}^\# F)^*(\tau) d\tau.
\]
From this and (5) we get
\[ \| F_1 \|_{T^{\#}_q} + t \| F_2 \|_{T^{\#}_q} \leq c' \int_0^{ct} (A_{q,\lambda}^\# F)(\tau) d\tau, \]
as required. \( \square \)

**Proof of Proposition 2.3.** It is easy to see that for any \( h > 0 \),
\[ \left( \int_{\Gamma^h(\xi)} |FG(y, t)|^q \frac{dydt}{t^{n+1}} \right)^{1/q} \leq A_\infty(F)(\xi) \left( \int_{\Gamma^h(\xi)} |G(y, t)|^q \frac{dydt}{t^{n+1}} \right)^{1/q}. \]
From this estimate and Proposition 3.1 we obtain
\[ ((A_q(FG)_F^*\lambda^*|B|) \leq (A_\infty(F)\lambda^*|B|/4)((A_q(G)_F^*\lambda^*|B|/4), \]
which completes the proof. \( \square \)

**Proof of Theorem 2.4.** Since the function \( M^\#_\lambda f \) is non-increasing in \( \lambda \), it suffices to prove that
\[ M^\#_\lambda f(x) \leq c_{n,\lambda} S^\#_{n,\lambda} f(x) \quad (0 < \lambda < 1). \]
Then the statement of the theorem will follow for \( 0 < \lambda < \lambda_n \). Assume first that \( f \in S \). Then (11) is equivalent to the fact that
\[ f(\xi) = \int_0^\infty f * \varphi_t * \psi_t(\xi) \frac{dt}{t} \quad \text{for all} \quad \xi \in \mathbb{R}^n. \]
Let \( x, \xi \in Q \). Define \( F(y, t) = f * \varphi_t(y) \) and set
\[ f_{1,Q}(\xi) = \int_0^{2^dQ} \int_{\mathbb{R}^n} F(y, t) \psi_t(\xi - y) \frac{dydt}{t}, \]
\[ f_{2,Q}(\xi) = \int_0^{2^dQ} \int_{\mathbb{R}^n} F(y, t) \left( \psi_t(\xi - y) - \psi_t(Q - y) \right) \frac{dydt}{t}, \]
\[ C_Q(f) = \int_0^{2^dQ} \int_{\mathbb{R}^n} F(y, t) \psi_t(Q - y) \frac{dydt}{t}. \]
By (22), \( f - C_Q(f) = f_{1,Q} + f_{2,Q} \). Using (19), (20) and Hölder’s inequality, we get
\[ |f_{2,Q}(\xi)| \leq c \sum_{k=0}^{\infty} \frac{1}{2^k} \left( (A_1^*F|2^{k+2}d_Q^*\lambda_{2^kQ})^* (2^k|Q|) \right) \equiv T_2(f; Q). \]
We now estimate \( (f_{1,Q} \lambda|Q|). \) Set \( D = \bigcup_{x \in Q} \Gamma^{2dQ}(x) \) and
\[ \mathcal{F} = \{ x : A_2(F\lambda_D)(x) \leq (A_2(F\lambda_D))^* (\lambda|Q|/2c_n) \}, \]
where \( c_n \) is the same as in Lemma 3.2. Choose also \( \mathcal{F}^* \subset \mathcal{F} \) as in Lemma 3.2. Let \( E \subset Q \) be an arbitrary measurable set with \( |E| = \lambda|Q| \). Then, arguing
as in the proof of Lemma 3.3, we see that $|\mathfrak{F}^* \cap E| \geq \lambda |Q|/2$. Next, let $D_1 = \bigcup_{x \in \mathfrak{F}^* \cap E} T^{2d_Q}(x)$ and

$$\tilde{f}(x) = \int_0^{2d_Q} \int_{\mathbb{R}^n} F_{\chi_{D_1}}(y, t) \psi_t(x - y) \frac{dydt}{t}.$$

Note that $f_{1,Q}(\xi) = \tilde{f}(\xi)$ for all $\xi \in \mathfrak{F}^* \cap E$. Write $\tilde{f}(\xi) = \sum_{Q'} a_{Q'} \gamma_{Q'}(\xi)$, where

$$\gamma_{Q'}(\xi) = \frac{1}{a_{Q'}} \int_{\ell_{Q'/2}}^{\ell_{Q'}} \int_{Q'} F_{\chi_{D_1}}(y, t) \psi_t(\xi - y) \frac{dydt}{t},$$

$$a_{Q'} = c \left( \int_{\ell_{Q'/2}}^{\ell_{Q'}} \int_{Q'} |F_{\chi_{D_1}}|^2 \frac{dydt}{t} \right)^{1/2},$$

and the summation is carried over all dyadic cubes $Q'$ ($\ell_{Q'}$ denotes the side length of $Q'$). Using the quasi-orthogonal argument (see [13, p. 171-172]), Lemma 3.2, and the fact that $\text{supp} \ A_2(F_{\chi_{D}}) \subset 5\sqrt{n}Q$, we get

$$\inf_{x \in E} |f_{1,Q}(x)| \leq \inf_{x \in \mathfrak{F}^* \cap E} |f_{1,Q}(x)| \leq \left( \frac{c}{|Q|} \int |\tilde{f}|^2 \right)^{1/2} \leq c \left( \int_{D_1} |F(y, t)|^2 \frac{dydt}{t} \right)^{1/2}$$

$$\leq c \left( \int_{D_1} \int_{\mathfrak{F}(x)} |F_{\chi_{D}}(y, t)|^2 \frac{dydt}{t^{n+1}} \right)^{1/2} \leq c((A_2(F^{2d_Q})_{\chi_{5\sqrt{n}Q}})^*(\lambda |Q|/2c_n)).$$

Taking the supremum over all $E \subset Q$ with $|E| = \lambda |Q|$ gives

$$(f_{1,Q}\chi_{Q})^*(\lambda |Q|) \leq c((A_2(F^{2d_Q})_{\chi_{5\sqrt{n}Q}})^*(\lambda |Q|/2c_n)) = T_2(f; Q).$$

(24) $$(f_{1,Q}\chi_{Q})^*(\lambda |Q|) \leq c((A_2(F^{2d_Q})_{\chi_{5\sqrt{n}Q}})^*(\lambda |Q|/2c_n)) \equiv T_2(f; Q).$$

Now let $f$ be an arbitrary function satisfying (10). Choose a sequence $f_j \in S$ such that $(f_j)^\#(x) \leq c f^\#(x)$ and

$$\int_{\mathbb{R}^n} \frac{|f(x) - f_j(x)|}{(1 + |x|)^{n+1}} = 0$$

as $j \to \infty$. Using (13), (15) and (16), we obtain that

$$T_1(f - f_j; Q) + T_2(f - f_j; Q) \to 0 \quad \text{as} \quad j \to \infty.$$
Applying also (23), (24) and Lemma 3.3, we get
\[
\inf_c ((f - c) \chi_Q)^*(\lambda|Q|) \\
\leq \inf_j \left( \left( (f - f_j) \chi_Q \right)^*(\lambda|Q|/2) + \left( (f_j - c_Q) \chi_Q \right)^*(\lambda|Q|/2) \right) \\
\leq \inf_j \left\{ \left( (f - f_j) \chi_Q \right)^*(\lambda|Q|/2) + T_1(f - f_j; Q) \\
+ T_2(f - f_j; Q) \right\} + cS^\#_{\lambda_n, \lambda} f(x) \\
= cS^\#_{\lambda_n, \lambda} f(x).
\]
The theorem is proved. \[\Box\]

**Proof of Corollary 2.5.** Taking \(\lambda = \lambda_n\) and applying then (14) along with (12) and (2), we obtain
\[
(f)^2_2(x) \asymp M_2(M_\lambda^# f)(x) \leq cM_2(S_\lambda^# f)(x) \asymp T^# f(x).
\]
To prove the converse, we need also the following simple estimate:
(25) \[
M_2(M f(x)) \leq cM_2 f(x).
\]
In fact, this follows from elementary geometry of balls and from the \(L^2\) boundedness of the Hardy-Littlewood maximal operator. We leave the details to the reader. Now, using (2), (13) and (25), we get
\[
T^# f(x) \asymp M_2(S_\lambda^# f)(x) \leq cM_2(f^#)(x) \asymp M_2(MM_\lambda^# f)(x) \\
\leq cM_2(M_\lambda^# f)(x) \asymp (f)^2_2(x),
\]
and this completes the proof. \[\Box\]

**Acknowledgements.** The author thanks E. Liflyand for his interest in this work and helpful discussions. I am also grateful to the referee for his useful comments and suggestions.

**References**


Received December 24, 2001 and revised January 27, 2002.

Department of Mathematics
Bar-Ilan University
52900 Ramat Gan
Israel
E-mail address: aklerner@netvision.net.il
OBSERVATIONS ON LICKORISH KNOTTING OF CONTRACTIBLE 4-MANIFOLDS

CHARLES LIVINGSTON

Lickorish has constructed large families of contractible 4-manifolds that have knotted embeddings in the 4-sphere and has also shown that every finitely presented perfect group with balanced presentation occurs as the fundamental group of the complement of a knotted contractible manifold. Here we make a few observations regarding Lickorish’s construction, showing how to extend it to construct contractible 4-manifolds which have an infinite number of knotted embeddings and also to construct knotted embeddings of the Mazur manifold for which the complement has trivial fundamental group.

In his recent paper, Lickorish [Li] describes a clever construction yielding for each finitely presented perfect group $G$ with balanced presentation a compact contractible 4-manifold $M_G$ with two embeddings in $S^4$, one for which the complement is diffeomorphic to $M_G$ and the other with complement having fundamental group $G$. Here we make several observations based on Lickorish’s work. With minor modifications, we use the notation of [Li] throughout.

**Observation 1.** The construction can be modified to assure that:

1. For each group $G$ there is an infinite family of $M_G$ having the desired pair of embeddings and
2. For different groups $G$ the constructed infinite families have no elements in common.

**Proof.** We let $M_0$ be the Mazur manifold [Ma] with Kirby diagram as shown in Figure 1. (The curves $\gamma$ and $\gamma'$ are extraneous for now.) Since $M_0$ embeds in $S^4$ with simply connected complement (as in [Li], $M_0 \times I \cong B^5$), the manifold $M_G$ of the construction can be replaced with the boundary connected sum $M_G \# \partial M_0$. This manifold will still have two embedding into $S^4$, one with contractible complement and the other with complement having fundamental group $G$. Note that this manifold is not diffeomorphic to $M$ since the boundary has changed by forming the connected sum with the boundary of the Mazur manifold, which Mazur showed is not $S^3$. By repeating this process one can easily build the desired families of examples;
for instance, arrange that for the first group $G$ all the boundaries have a prime number of summands, for the second group $G$ arrange that all have a prime squared number of summands, etcetera.

\[\square\]

\textbf{Figure 1.}

Following Observation 2 we will give examples showing that groups with the desired properties exist.

\textbf{Observation 2.} If the finitely generated perfect group $G$ with balanced presentation maps onto a nontrivial finite quotient of a 2-knot group then there exists an infinite family of embeddings, $\{\phi_i\}$, of the manifold constructed by Lickorish, $M_G$, into $S^4$ distinguished by $\pi_1(S^4 - \phi_i(M_G))$.

\textit{Proof.} Let $K$ be the 2-knot and let $H = \pi_1(S^4 - K)/N$ be the finite quotient. Let $\rho : G \to H$ be the given surjective homomorphism. Also, let $x$ be a meridian of $K$ and let $\overline{x}$ denote its image in $H$.

Pick an element $g \in G$ such that $\rho(g) = \overline{x}$. It can be arranged in the construction of $M_G$ that $g$ is among the generators of $G$ in the balanced presentation: Simply add a generator $z$ to the balanced presentation along with the relation $z = g$, with $g$ written in terms of the generators of the initial balanced presentation.

It now follows from the initial construction of $M_G$ as the complement of $X_G$ that the meridian of the 2-handle of $M_G$ corresponding to the generator $z$ represents $g \in G = \pi_1(X_G)$. To construct a new embedding of $M_G$ into $S^4$, tie the knot $K$ as a local knot in the given 2-handle of $M_G$. This does not change $M_G$ but changes the fundamental group of $X_G$; the new group is constructed from the free product $G \ast \pi_1(S^4 - K)$ by identifying $g \in G$ with $x \in \pi_1(S^4 - K)$. We denote this group by $G_1$ and also write it as $G_1 \ast_{\pi_1(S^4 - K)} \pi_1(S^4 - K)$, though it is not an amalgamated product in the case that $g$ has finite order in $G$. (It is not clear that this new group is not isomorphic to $G$.)

There are two homomorphisms of $\pi_1(S^4 - K)$ to $H$: The first is the projection, $p$, the second factors through the cyclic group $Z$, mapping the meridian to $x$. Denote this second map by $q$. The maps $\rho \ast p$ and $\rho \ast q$ each determine homomorphisms of $G_1$ to $H$. These homomorphisms are distinct as one is surjective when restricted to the image of $\pi_1(S^4 - K)$ and the other is not surjective when restricted to this subgroup. (Note that $H$ is perfect since it is a quotient of $G$ and hence is not cyclic.)

We first observe that these two embeddings cannot be isotopic; if they were there would be an isomorphism from $G$ to $G \ast Z \pi_1(S^4 - K)$ carrying the meridian representing to $g$ to the meridian $m'$ representing $g = x$. Thus there would be a group isomorphism from $G$ to $G \ast Z \pi_1(S^4 - K)$ sending $g$ to $g = x$. However that cannot be as, by the above argument, $G$ and $G \ast Z \pi_1(S^4 - K)$ have different numbers of homomorphisms onto $H$ sending these preferred meridians to $x$. (Notice that since $H$ is finite there is a finite number of such homomorphisms.)

By repeating the construction of locally knotting the 2-handle of $M$ using $K$, a sequence of nonisotopic embeddings of $M$ into $S^4$ is constructed.

We cannot show that the sequence of fundamental groups of the complements are all distinct, but by counting homomorphisms to $H$ it follows that some subsequence must be distinct; the number of homomorphisms onto $H$ goes to infinity since after adding $n$ knots to the band there are at least $2^n$ homomorphisms onto $H$. □

**Examples.** This simplest example of Observation 2 occurs with the binary icosahedral group, $H(2,3,5)$, the perfect group with 120 elements representing the fundamental group of +1 surgery on the trefoil knot. This group clearly has a balanced presentation and is also a quotient of the trefoil group, which is isomorphic to the fundamental group of the 0-twist spin of the trefoil. More generally, consider the group of the $r$-fold cyclic branched cover of the $(p,q)$-torus knot, denoted $H(p,q,r)$. If $p$, $q$, and $r$, are pairwise relatively prime, then $H(p,q,r)$ is perfect. Furthermore, according to Gordon, [Go], the $r$-twist spin of the $(p,q)$-torus knot has fundamental group $H(p,q,r) \times Z$, and hence maps onto $H(p,q,r)$. It remains to show that $H(p,q,r)$ has a finite quotient. A presentation of $H(p,q,r)$ is given by $\langle x,y \mid x^p = y^q = (xy)^r \rangle$. For such groups a nontrivial representation to an alternating group can be constructed. (This was done by Fox in [Fo]; a more accessible reference is Milnor [Mi].)

**Observation 3.** There exist contractible manifolds built with a single 1-handle that possess two embeddings in $S^4$, one with simply connected complement and one with nontrivial complementary fundamental group.

**Proof.** This fact follows from the result of Neuzil [Ne] showing that the Dunce Cap embeds in $S^4$ with nonsimply connected complement. The following approach gives us more control over the contractible manifold as well.
Suppose that $L = L_1 \cup L_2$ is a 2-component link in $S^3$ with $L_1$ unknotted bounding a trivial slice disk $D_1$ in $B^4$ and $L_2$ slice, bounding a slice disk $D_2$ in $B^4$. Assume the linking number is 1. Let $S^3$ separate $S^4$ into two components, $B_1$ and $B_2$ and view $D_1 \subset B_1$ and $D_2 \subset B_2$.

As in Lickorish’s construction, we let $M = (B_1 - N(D_1)) \cup N(D_2)$. This is clearly a contractible 4-manifold that doubles to give $S^4$. However, its complement is $X = (B_2 - N(D_2)) \cup N(D_1)$. Its group is given by the group of the complement of the slice disk with an added relations coming from adding the 2-handle, $N(D_1)$.

**Examples.** For any knot $K$, the knot $L_2 = K \# - K$ is slice with fundamental group of the complement of the slice disk being $\pi_1(S^3 - K)$. Any element of this group can be represented by an unknot $L_1$ in the complement of $L_2$. Hence, the groups that arise in this construction include all perfect groups constructed by adding a single relation to a classical knot group. For instance: The fundamental group of a homology 3-sphere built by surgery on a classical knot is the fundamental group of the complement of the embedding of a contractible 4-manifold with one 1-handle into $S^4$.

In the previous construction it is not clear that we are constructing distinct embeddings if $L_2$ is unknotted; this case is perhaps the most perplexing. We have the following example:

**Observation 4.** The Mazur Manifold illustrated in Figure 1 has two nonisotopic embeddings into $S^4$.

**Proof.** As described for instance in [Ak], the boundary of the Mazur manifold $M$ has an involution $F$ carrying $\gamma$ to $\gamma'$. A handlebody picture of the manifold $M \cup_{id} M$ is formed from Figure 1 by adding a 2-handle with 0 framing to $\gamma'$. Similarly, a handlebody picture of the manifold $M \cup_F M$ is formed from Figure 1 by adding a 2-handle with 0 framing to $\gamma$. (In each case a 3- and 4-handle must be added as well.) It is an easy exercise in Kirby calculus [AK] to see that both are $S^4$. Hence, we have two embeddings of $M$ into $S^4$.

Clearly, in the first case the curve $\gamma'$ is slice in the complement – the 2-handle is added to $\gamma'$. In the second case $\gamma'$ is not slice in the complement – this is a result of Akbulut, [Ak].

**Questions.** In the above construction, if $L_2$ is unknotted is there an embedding of the constructed Mazur-like manifold into $S^4$ with nonsimply connected complement? If the crossing that is not part of the clasp of the attaching map of the 2-handle in Figure 1 is changed the previous argument does not apply – Akbulut has shown that in this case $\gamma$ will be slice. Does this manifold knot in $S^4$? Does there exist a contractible 4-manifold that does not knot in $S^4$?
Acknowledgements. Thanks go to Ray Lickorish, first for identifying this interesting topic, secondly for pointing out some of the issues resolved here, and finally for his reflections on my first attempts at extending his results.

References


Received November 24, 2001 and revised January 6, 2002.

Department of Mathematics
Indiana University
Bloomington, IN 47405
E-mail address: livingst@indiana.edu
We give an explicit description of all 16-dimensional locally compact translation planes admitting the unimodular quaternion group $SL_2 \mathbb{H}$ as a group of collineations. Moreover, we shall also determine the full collineation groups of these planes.

1. Introduction.

In this paper, all 16-dimensional locally compact translation planes admitting the unimodular quaternion group $SL_2 \mathbb{H}$ as a group of collineations will be determined explicitly. Besides the classical plane over the octonions there are a vast number of planes having this property, cf. the Classification Theorem (2.8). Indeed, the class of these planes covers an interesting borderline case: Among all 16-dimensional locally compact translation plane, only the classical plane admits the action of a noncompact almost simple Lie group of dimension larger than $\dim SL_2 \mathbb{H} = 15$, cf. [7, Theorem A].

The connected component $G^e$ of the automorphism group $G$ of a non-classical example is composed of the translation group, the group of homotheties, the group $SL_2 \mathbb{H}$, and a compact group $\Delta$ isomorphic to $\{e\}, SO_2 \mathbb{R}$, $SO_2 \mathbb{R} \times SO_2 \mathbb{R}$, or $SU_2 \mathbb{C}$, cf. Theorem 3.8. Thus, $\dim G$ is at most 35.

It is worth mentioning that $\Gamma = G^e$ leaves precisely one projective line (namely the translation axis) invariant, but does not fix any projective points. In general, a 16-dimensional compact projective plane whose automorphism group contains a closed connected subgroup $\Gamma$ having this property and satisfying $\dim \Gamma \geq 35$ is necessarily a translation plane, thanks to a theorem of H. Salzmann [10]. Recently, H. Hähl has shown in [4] that there are precisely three families of such planes: A subfamily of the planes considered here, and the planes admitting $SU_4 \mathbb{C} \cdot SU_2 \mathbb{C}$ or $SU_4 \mathbb{C} \cdot SL_2 \mathbb{R}$ as a group of collineations, determined in Hähl [5]. In particular, $\dim \Gamma \geq 36$ implies that the plane is isomorphic to the octonion plane.

\footnote{More precisely: The planes for which the group $\Delta$ mentioned above equals $SU_2 \mathbb{C}$; see 3.8(2) for further details.}
Organization. The second section is devoted to the proof of the Classification Theorem (2.8) which is based on the general theory of noncompact semisimple groups acting on locally compact translation planes. (See [7] and [8], and compare 2.2, 2.3 and 2.5 for the particular applications.)

In 2.8, we shall assign to each continuous function \( \sigma : \text{Spin}(3) \to \mathbb{R}_{\text{pos}} \) a 16-dimensional translation plane \( \mathcal{P}^\sigma \) admitting the action of \( \text{SL}_2 \mathbb{H} \). One of the quasifields belonging to such a plane \( \mathcal{P}^\sigma \) will be obtained in 2.11.

In 3.7 we shall give a necessary and sufficient condition for two functions to define isomorphic planes. Finally, we determine the automorphism groups in 3.8 by computing the reduced stabilizer \( S_{G_0} \) of each plane \( \mathcal{P}^\sigma \). With the exception of the octonion plane, the automorphism groups of the planes under consideration have dimension at most 35.

1.1. Notation. Let \( \text{Spin}(3) \) denote the group of quaternions of length 1. For \( \vec{x}, \vec{y} \in \mathbb{H}^2 \), we put \( \langle \vec{x}, \vec{y} \rangle := x_1 \overline{y_1} + x_2 \overline{y_2} \). For the orthogonal complement of a subspace \( X \) with respect to this scalar product we shall write \( X^\perp \).

If \( A \) is an element of \( \text{SL}_2 \mathbb{H} \), then \( A^* \) denotes the inverse of the adjoint map of \( A \) with respect to \( \langle \cdot, \cdot \rangle \), i.e., \( A^* := (A^\dagger)^{-1} \). We emphasize that we have
\[
\langle \vec{x} \cdot A^*, \vec{y} \cdot A \rangle = \langle \vec{x}, \vec{y} \rangle \quad \text{for all} \quad \vec{x}, \vec{y} \in \mathbb{H}^2, A \in \text{SL}_2 \mathbb{H}.
\]
Let \( \text{SH}_2^+ \mathbb{H} \) be the set of positive definite Hermitian \((2 \times 2)\)-matrices over \( \mathbb{H} \) with determinant 1. Notice that \( \text{SH}_2^+ \mathbb{H} \) coincides with the set of all \( (A^*)^{-1}A = A^\dagger A \), \( A \in \text{SL}_2 \mathbb{H} \). (Recall the polar decomposition of unimodular matrices.)

Finally, \( \text{diag}(x_1, \ldots, x_n) \) denotes a diagonal matrix with the given entries.

2. The classification.

2.1. The general situation. We consider a 16-dimensional locally compact affine translation plane \((P, \mathcal{L})\) which is represented in the usual way: The point space \( P \) is a 16-dimensional real vector space, the line pencil \( \mathcal{L}_0 \) through the origin consists of 8-dimensional vector subspaces of \( P \), and the other lines are the affine cosets of the elements of \( \mathcal{L}_0 \). Moreover, the spread \( \mathcal{L}_0 \) is a compact subset of the Grassmannian manifold of all 8-dimensional vector subspaces of \( P \). In fact, \( \mathcal{L}_0 \) is homeomorphic to the 8-sphere.

The group \( G \) of all automorphisms (i.e., continuous collineations) is a semidirect product \( G = G_0 \rtimes T \) of the translation group \( T \) (which coincides with the group of all vector translations of \( P \)) and the stabilizer \( G_0 \) of the origin. The latter group is a closed subgroup of \( \text{GL}(P) \) and, hence, is a Lie group.

\[2\]Basic facts concerning 16-dimensional locally compact translation planes are collected in Chapter 8 of [11]; results used without a reference can be found there.
2.2. The group action of $\text{SL}_2 \mathbb{H}$ on $(P, \mathcal{L})$. Throughout this paper we suppose that $\text{SL}_2 \mathbb{H}$ acts on the translation plane $(P, \mathcal{L})$ as a group of collineations, i.e., we have a Lie homomorphism $\Phi : \text{SL}_2 \mathbb{H} \to G$ with discrete kernel. Since $G$ is an almost direct product of $G_0$ and the abelian translation group, we may assume that the image

$$\Lambda = \Phi(\text{SL}_2 \mathbb{H})$$

is contained in $G_0$. In fact, $\Phi$ is a representation of $\text{SL}_2 \mathbb{H}$ on $P$.

According to [7, 6.8], $\Phi$ is a direct product of the obvious representation of $\text{SL}_2 \mathbb{H}$ on $\mathbb{H}^2$ and the contragredient representation on $\mathbb{H}^2$. Therefore, $P$ and the left quaternion vector space $\mathbb{H}^4$ can be identified in such a way that the representation $\Phi$ of $\text{SL}_2 \mathbb{H}$ on $P = \mathbb{H}^4$ is given by right multiplication with the matrices

$$\Phi(A) = \begin{pmatrix} A^* & 0 \\ 0 & A \end{pmatrix} \quad \text{for } A \in \text{SL}_2 \mathbb{H}.$$ 

We emphasize that the $\Lambda$-invariant subspaces $\mathbb{H}^2 \times \{0\}$ and $\{0\} \times \mathbb{H}^2$ are not elements of $\mathcal{L}_0$, since the noncompact almost simple group $\Lambda$ does not fix any affine line, cf. [7, Theorem B].

2.3. The weight sphere. We apply the general theory of noncompact almost simple subgroups of $G_0$ (for which [7] contains the details) to our particular case: Being the image of $\text{diag}(-1,1) \in \text{sl}_2 \mathbb{H}$ under the derivative of $\Phi$, the real diagonal matrix $d := \text{diag}(1,-1,-1,1)$ is an element of the Lie algebra $\mathfrak{L}\Lambda$.

Since $d$ has precisely two eigenvalues, [7, 5.3] implies that both eigenspaces of $d$ are elements of the spread $\mathcal{L}_0$. Collecting all the eigenspaces of all real diagonalizable elements of $\mathfrak{L}\Lambda$ yields the so-called weight sphere $S \subseteq \mathcal{L}_0$ of $\Lambda$, see [7] for details. The main result [7, Theorem B] concerning the weight sphere asserts that $\Lambda$ acts transitively on it. Therefore, $S$ is the $\Lambda$-orbit of the eigenspace $E := \mathbb{H} \cdot (1,0) \times \mathbb{H} \cdot (0,1)$ of $d$ with respect to 1.

Lemma 2.4.

(a) The weight sphere $S$ of $\Lambda$ consists precisely of the subspaces $X \times X^\perp$, where $X$ is a 1-dimensional $\mathbb{H}$-linear subspace of $\mathbb{H}^2$.

(b) A vector $(\vec{x}, \vec{y}) \in \mathbb{H}^2 \times \mathbb{H}^2$ is contained in some element of $S$ if and only if $\vec{x}$ is perpendicular to $\vec{y}$.

Proof. (a) Let $G_1 \mathbb{H}^2$ be the set of 1-dimensional $\mathbb{H}$-linear subspaces of $\mathbb{H}^2$. We have to show that the sets $S$ and $S' := \{X \times X^\perp | X \in G_1 \mathbb{H}^2\}$ coincide. For this, let $X \in G_1 \mathbb{H}^2$ and $A \in \text{SL}_2 \mathbb{H}$. Derive $(XA^*)^\perp = X^\perp A$ from Equation (1) in 1.1. This shows that

$$(X \times X^\perp) \Phi(A) = X A^* \times X^\perp A = X A^* \times (X A^*)^\perp,$$
whence $S'$ is $\Lambda$-invariant. In fact, $S'$ is a $\Lambda$-orbit, because $\text{SL}_2\mathbb{H}$ acts transitively on $G_1\mathbb{H}^2$. We already know that the weight sphere $S$ is a $\Lambda$-orbit, too. Since $E = \mathbb{H} \cdot (1, 0) \times \mathbb{H} \cdot (0, 1)$ is an element of $S \cap S'$, we conclude that $S = S'$. Part (b) is an easy consequence of (a).

2.5. Stabilizers. Let $O$ be an orbit of $\Lambda$ in $L_0$ different from $S$. From [8, 6.2.b] we infer that the stabilizer of an element of $O$ is a compact group. Thus, there exists a line $L \in O$ such that $\Lambda_L$ is contained in the particular maximal compact subgroup $\Delta := \Phi(U_2\mathbb{H})$ of $\Lambda$.

Being a subset of the 8-sphere $L_0$, the orbit $O$ has dimension at most 8. Applying Halder’s dimension formula yields

$$\dim \Lambda_L = \dim \Lambda - \dim \Lambda^A \geq 15 - 8 = 7.$$  

This implies that $\Lambda_L$ equals $\Delta$, because $U_2\mathbb{H}$ does not contain proper subgroups of dimension at least 7: The rank of a subgroup $\Sigma$ of $U_2\mathbb{H}$ is at most 2. Moreover, $\Sigma$ is not isomorphic to the group $\text{SU}_3\mathbb{C}$ which does not have a representation of the quaternion vector space $\mathbb{H}^2$, cf. [11, 95.10]. Checking all compact groups of rank at most 2 yields the assertion.

2.6. $\Delta$-invariant subspaces and their orbits. Note that a matrix $B \in \text{SL}_2\mathbb{H}$ is an element of $U_2\mathbb{H}$ if and only if $B^* = B$, whence we obtain

$$\Phi(B) = \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix}$$

for all $B \in U_2\mathbb{H}$.

Thus, the restriction of $\Phi$ to $U_2\mathbb{H}$ is a direct sum of two copies of the irreducible representation of $U_2\mathbb{H}$ on $\mathbb{H}^2$. By [2, p. 43, Prop. 6], precisely the following proper $\mathbb{R}$-linear subspaces of $P$ are invariant under $\Delta = \Phi(U_2\mathbb{H})$:

$$U_h := \{(\bar{x}, h\bar{x}) \mid \bar{x} \in \mathbb{H}^2\} \text{ for } h \in \mathbb{H} \text{ and } U_\infty := \{\bar{0}\} \times \mathbb{H}^2.$$  

We compute the image of $U_h$, $h \in \mathbb{H}$, under $\Phi(A)$, $A \in \text{SL}_2\mathbb{H}$:

$$U_h^{\Phi(A)} = \{(\bar{x}A^*, h\bar{x}A) \mid \bar{x} \in \mathbb{H}^2\} = \{(\bar{y}, h\bar{y}\bar{A}A) \mid \bar{y} \in \mathbb{H}^2\}.$$  

Recall that $\bar{A}A$ is an element of $\text{SH}_2^+\mathbb{H}$ and that every element of $\text{SH}_2^+\mathbb{H}$ has this form. Therefore, the $\Lambda$-orbit $U_h^\Lambda$ consists precisely of the subspaces

$$\{(\bar{x}, h\bar{x}S) \mid \bar{x} \in \mathbb{H}^2\} \text{ where } S \in \text{SH}_2^+\mathbb{H}.$$  

Lemma 2.7.

(a) A nonzero vector $(\bar{x}, \bar{y}) \in \mathbb{H}^2 \times \mathbb{H}^2$ is contained in an element of the $\Lambda$-orbit $U_h^\Lambda$ if and only if $\langle \bar{x}, \bar{y} \rangle = r\bar{h}$ holds for some $r \in \mathbb{R}_{\text{pos}}$.
(b) For every $h \in \mathbb{H}^\times$ the set $S \cup U_h^\Lambda$ is a partial spread.
(c) If $h$ and $l$ are distinct nonzero quaternions, then $U_h^\Lambda \cup U_l^\Lambda$ is a partial spread if and only if $h/|h| \neq l/|l|$.
Proof. (a) Let $A \in \text{SL}_2 \mathbb{H}$. Then every nonzero vector belonging to the element $\{ \langle \vec{x} A^*, h \vec{x} A \rangle \mid \vec{x} \in \mathbb{H}^2 \}$ of $U^\Lambda_h$ has the desired property, because $\langle \vec{x} A^*, h \vec{x} A \rangle = \langle \vec{x}, h \vec{x} \rangle = \| \vec{x} \|^2 \cdot h$. Conversely, let $(\vec{x}, \vec{y})$ be an element of $\mathbb{H}^2 \times \mathbb{H}^2$ such that $\langle \vec{x}, \vec{y} \rangle = r h$ holds for some $r \in \mathbb{R}_{\text{pos}}$. Without loss of generality we may assume $\vec{x} = (1,0)$ (otherwise, replace $(\vec{x}, \vec{y})$ by $(\vec{x} B^*, \vec{y} B)$, where $B \in \text{SL}_2 \mathbb{H}$ satisfies $\vec{x} B^* = (1,0)$). Then $\langle \vec{x}, \vec{y} \rangle = r h$ implies that $\vec{y} = (r h, l)$ holds for some $l \in \mathbb{H}$. Put

$$
A := \left( \begin{array}{cc} \sqrt{r^{-1}} & -\sqrt{r^{-1}} h^{-1} l \\ 0 & \sqrt{r} \end{array} \right)
$$

and observe that $\Phi(A)$ maps $(\vec{x}, \vec{y}) = (1,0, r h, l)$ to the element $(\sqrt{r}, 0, \sqrt{r} h, 0)$ of $U_h$. This proves the claim.

(b) It is easy to see that the weight sphere $S$ is a partial spread. If $U_h$ and $U_{h}^{\Phi(A)}$, $A \in \text{SL}_2 \mathbb{H}$, have some nonzero vector $(\vec{x}, \vec{y})$ in common, then $\vec{y} = h \vec{x} = h \vec{x} A^t A$ implies that $1$ is an eigenvalue of the positive definite Hermitian $(2 \times 2)$-matrix $\vec{x}^t A$. Since $\text{det} \ A^t A = 1$, we derive that $A^t A$ is an element of $U_2 \mathbb{H}$, whence $U_h$ and $U_{h}^{\Phi(A)}$ coincide. Consequently, $U^\Lambda_h$ is a partial spread. Moreover, the sets of nonzero vectors covered by $S$ and $U^\Lambda_h$, respectively, are disjoint, cf. (a) and 2.4(b). This proves the assertion.

(c) If $h/|h| \neq l/|l|$, then (a) shows that the sets of nonzero vectors covered by the partial spreads $U^\Lambda_h$ and $U^\Lambda_l$, respectively, are disjoint. Thus, $U^\Lambda_h \cup U^\Lambda_l$ is a partial spread, too. For the converse direction we suppose that $h/|h| = l/|l|$. Then $U^\Lambda_h$ and $U^\Lambda_l$ are partial spreads covering the same set of vectors, thanks to (a), hence their union fails to be a partial spread unless $U^\Lambda_h = U^\Lambda_l$. The latter condition implies that $U_l$ equals $U^\Phi(A)$ for some $A \in \text{SL}_2 \mathbb{H}$, and consequently $l \vec{x} = h \vec{x} A^t A$ holds for all $\vec{x} \in \mathbb{H}^2$. This implies that every vector is an eigenvector of $\vec{x} A^t A$ with respect to the eigenvalue $h^{-1} l$. Since $A$ is unimodular, we derive that $l = h$. \hfill \Box

Classification Theorem 2.8. Let $\sigma : \text{Spin}(3) \to \mathbb{R}_{\text{pos}}$ be a continuous function. Then the set

$$
\mathcal{L}^\sigma_0 := S \cup \left\{ U^\Lambda_{\sigma(p)} \mid p \in \text{Spin}(3), \lambda \in \Lambda \right\}
$$

is a $\Lambda$-invariant compact spread on $P = \mathbb{H}^2 \times \mathbb{H}^2$ and, hence, defines a 16-dimensional locally compact translation plane $\mathcal{P}^\sigma$ whose automorphism group contains the group $\Lambda \cong \text{SL}_2 \mathbb{H}$. (Recall the definition of $\Lambda$, $S$ and $U_h$ in 2.2, 2.4 and 2.6, respectively.)

Conversely, if $\mathcal{P}$ is a 16-dimensional locally compact translation plane admitting the group $\text{SL}_2 \mathbb{H}$ as a group of collineations, then $\mathcal{P}$ is isomorphic to $\mathcal{P}^\sigma$ for some continuous function $\sigma : \text{Spin}(3) \to \mathbb{R}_{\text{pos}}$. 


we can identify \((\vec{x}, \vec{y})\) be an element of \(P\) with \((\vec{x}, \vec{y}) = h\). If \(h\) vanishes, then \((\vec{x}, \vec{y})\) is covered by \(S\), cf. 2.4. For \(h \neq 0\) put \(l = \sigma(\vec{h}/|h|) \cdot \vec{h}/|h|\) and use 2.7(a) to infer that \((\vec{x}, \vec{y})\) is contained in an element of \(U^L_1\). Thus, \(L^L_0\) covers \(P\) and we have shown that \(L^L_0\) is a spread.

(2) If we can prove that \(L_0\) is closed in the Grassmannian manifold of all 8-dimensional subspaces of \(P\), then \(L_0\) is a compact spread and, hence, \(P^\gamma\) is a locally compact translation plane. Therefore, we consider a sequence \((L_i)_i\), \(L_i \in L_0\), which is convergent to some 8-dimensional vector subspace \(L \leq P\).

If \(L_i\) is an element of the compact weight sphere \(S\) for infinitely many \(i\), then also \(L\) is an element of \(S\). Thus, we may assume that \(L_i \in L^L_0 \setminus S\) holds for all \(i\). By the definition of \(L^L_0\), we have that \(L_i = U^{h_i}_i\), where \(\lambda_i \in \Lambda\) and where \(h_i = \sigma(p_i)p_i\) for some \(p_i \in \text{Spin}(3)\).

For \(r \in \mathbb{R}_{\text{pos}}\) we put \(\rho(r) := \Phi(\text{diag}(r, r^{-1})) = \text{diag}(r^{-1}, r, r, r^{-1})\). By the KAK-decomposition [6, 7.39] of \(\text{SL}_2 \mathbb{H}\), there are \(\gamma_i, \delta_i \in \Delta\), \(r_i \in \mathbb{R}_{\text{pos}}\) such that \(\lambda_i = \gamma_i \rho(r_i) \delta_i\). Note that \(\Delta\) and \(\text{Spin}(3)\) are compact and that \(\sigma\) is continuous. By passing to a subsequence we may achieve the following:

(a) \(p_i\) is convergent to \(p \in \text{Spin}(3)\), whence \(\sigma(p_i)p_i\) is convergent to \(h := \sigma(p)p\),
(b) \(\delta_i\) is convergent to \(\delta \in \Delta\),
(c) \(r_i\) is convergent to \(r \in \mathbb{R}_{\text{pos}} \cup \{0, \infty\}\), and
(d) \(U^{\rho(r_i)}_{h_i}\) is convergent to some 8-dimensional vector subspace \(K\) of \(P\).

We claim that \(K\) is an element of \(L^L_0\) — then \(L = \lim_{i \to \infty} U^{\lambda_i}_{h_i} = \lim_{i \to \infty} (U^{\rho(r_i)}_{h_i})^{\delta_i}\) is a spread, thanks to 2.3. Combine 2.7(a) and 2.7(c) to infer the following: For every \(p \in \text{Spin}(3)\) there exists precisely one \(r \in \mathbb{R}_{\text{pos}}\) such that \(U^\gamma_{\rho(p)}\) is a subset of \(L_0\). Putting \(\sigma(p) := r\) we obtain a
function $\sigma : \text{Spin}(3) \to \mathbb{R}_{\text{pos}}$. It remains to show the continuity of $\sigma$. For this, consider a sequence $(p_i)_i$ in $\text{Spin}(3)$ which converges to $p$. In order to check $\lim_{i \to \infty} \sigma(p_i) = \sigma(p)$ we prove that $\sigma(p)$ is the only accumulation point of $(\sigma(p_i))_i$ in the interval $[0, \infty]$: Let $r$ be such an accumulation point. It is easy to see that $(U_{\sigma(p_i)p_i})_i$ is convergent to $U_{rp}$ in the Grassmannian topology. Since $\mathcal{L}_0$ is compact, we infer that $U_{rp}$ is an element of $\mathcal{L}_0$ and, hence, that $r = \sigma(p)$. □

**Remark 2.9.** The projective closures of the planes specified in 2.8 yield all 16-dimensional compact projective translation planes admitting the group $\text{SL}_2 \mathbb{H}$ as a group of collineations: According to [11, 64.4.c], such a plane either is classical, or the translation axis is invariant under all automorphisms, whence the group $\text{SL}_2 \mathbb{H}$ acts on the affine part as well.

**Remark 2.10.** The classification of 4-dimensional translation planes admitting the group $\text{SL}_2 \mathbb{R}$ as a group of collineations is due to D. Betten, see [11, 73.13 and 73.19] for the results. Note that there is an example with an irreducible $\text{SL}_2 \mathbb{R}$-action. The 8-dimensional translation planes admitting an $\text{SL}_2 \mathbb{C}$-action were completely determined by H. Hähl, see [3].

2.11. Coordinatizing quasifields of $P^\sigma$. We consider a function $\sigma : \text{Spin}(3) \to \mathbb{R}_{\text{pos}}$. Our aim is to introduce coordinates for the affine translation plane $P^\sigma$ with respect to the triangle $o = (0, 0, 0, 0)$, $w = (1, 0, 0, 0)$, $s = (0, 1, 0, 0)$. We claim that the resulting quasifield $Q^\sigma$ is obtained as follows: For $h \in \mathbb{H}$, we put

$$\zeta(h) := \begin{cases} 
0 & \text{if } h = 0, \\
\sigma(-h/|h|)^{-2}h & \text{if } h \neq 0.
\end{cases}$$

Then the quasifield in question is $Q^\sigma = \mathbb{H}^2$ with its natural addition, while the multiplication is given by

$$(h, l) \circ_\sigma (x, y) := (xh - \zeta(l)y, lx + yh) \text{ for } (h, l), (x, y) \in \mathbb{H}^2.$$ 

The line $G_{(h, l)} \in \mathcal{L}_0$ with slope $(h, l) \in \mathbb{H}^2$ is given by

$$G_{(h, l)} = \{(x, xh - \zeta(l)y, -lx - yh, y) \mid x, y \in \mathbb{H}^2\};$$

notice that $o \vee (w + s) = G_{(1, 0)} = \{(x, y, -x, y) \mid x, y \in \mathbb{H}\}$. Moreover, the vertical axis equals

$$G_\infty = o \vee s = \{0\} \times \mathbb{H} \times \mathbb{H} \times \{0\}.$$ 

We have to show that $\mathcal{L}_0^\sigma = \{G_z \mid z \in \mathbb{H}^2 \cup \{\infty\}\}$. To this end it suffices to prove that:

(1) $\mathcal{S} = \{G_{(h, 0)} \mid h \in \mathbb{H}\} \cup \{G_\infty\}$, and

(2) $U_{-\sigma(p)p} = \{G_{(h, rp)} \mid r \in \mathbb{R}_{\text{pos}}, h \in \mathbb{H}\}$ for all $p \in \text{Spin}(3)$.

3For details on how to coordinatize translation planes by quasifields we refer to [1].
Property (1) can be easily derived from the following equations:

\[ G_\infty = (\mathbb{H} \cdot (1, 0)) \times (\mathbb{H} \cdot (1, 0))^\perp \]

\[ G_{(h,0)} = \{(x, xh, -yh, y) \mid x, y \in \mathbb{H}\} = (\mathbb{H} \cdot (1, h)) \times (\mathbb{H} \cdot (1, h))^\perp, \]

recall the description of the elements of \( S \) in 2.3. We turn to (2): Directly from the definition of \( U_\sigma(p) \) we see that

\[ G_{(0,\sigma(-p)p)} = \{(x, -\sigma(-p)^{-1}py, -\sigma(-p)px, y) \mid x, y \in \mathbb{H}^2\} = U_{-\sigma(-p)p}. \]

Consider an element \( \lambda \) in \( \Lambda \). By the Iwasawa decomposition [6, 6.46] of \( \text{SL}_2 \mathbb{H} \), there are elements \( B \in U_{2\mathbb{H}}, s \in \mathbb{R}_{\text{pos}}, \) and \( h \in \mathbb{H} \) such that

\[ \Phi^{-1}(\lambda) = B \cdot \text{diag}(s, s^{-1}) \cdot \begin{pmatrix} 1 & 0 \\ -h & 1 \end{pmatrix}. \]

A short computation shows that \( U_{\lambda}^{\tau_{\sigma(p)p}} = G_{(0,\sigma(-p)p)} = G_{(h,s^2\sigma(-p)p)} \), and Property (2) follows easily. This finishes the proof.

**Corollary 2.12.** We consider the constant map \( \sigma : \text{Spin}(3) \to \mathbb{R}_{\text{pos}}; p \mapsto 1 \). Then \( \mathcal{P}^{\sigma} \) is isomorphic to the affine plane over the octonions.

**Proof.** The multiplication of the quasifield of \( \mathcal{P}^{\sigma} \) determined in 2.11 is \((h,l) \circ (x,y) = (xh - ly, lx + yh)\). Indeed, this is the multiplication of the division algebra \( \mathbb{O} \). \( \square \)

### 3. Isomorphisms and automorphisms.

**3.1. General remarks.** Let \( \mathcal{P} \) be a 16-dimensional locally compact translation plane whose group \( \mathcal{G} \) contains a subgroup \( \Lambda \) which is locally isomorphic to \( \text{SL}_2 \mathbb{H} \). Following the previous section, we identify \( \mathcal{P} \) and \( \mathbb{H}^4 \) such that \( \Lambda \) is the group specified in 2.2.

Moreover, let \( T \) be the group of vector translations of \( \mathcal{P} \) and let \( Y \) be the group of homotheties of \( \mathcal{P} \) with a positive real scalar. Putting \( \text{SG}_0 := \mathcal{G}_0 \cap \text{SL}(P) \), we infer that \( \mathcal{G}^{e} = (Y \times \text{SG}_0) \rtimes T \). (The exponent \( e \) refers to the connected component of a Lie group.) The group \( \text{SG}_0 \) is called the “reduced stabilizer” of \( \mathcal{P} \), see [11, 81.0] for details. In particular, we have that

\[ \dim \mathcal{G} = \dim \text{SG}_0 + \dim Y + \dim T = \dim \text{SG}_0 + 17. \]

**Proposition 3.2.** Retain the notation above. If \( \Lambda \) is not normal in the reduced stabilizer \( \text{SG}_0 \), then \( \mathcal{P} \) is isomorphic to the affine plane over the octonions. In every other case, \( \text{SG}_0 \) is an almost direct product \( \text{SG}_0 = \Lambda \cdot \Psi \) of \( \Lambda \) and a compact connected subgroup \( \Psi \) of the centralizer of \( \Lambda \) in \( \text{GL}(P) \).

**Proof.** Observe that \( \text{SG}_0 \) is a noncompact group which fixes no affine lines of \( \mathcal{P} \), since its subgroup \( \Lambda \) has this property. According to [8, 1.1], \( \text{SG}_0 \) is an almost direct product of an almost simple Lie group \( S \) of real rank 1 and
a compact group $\Psi$. From [7, Theorem B] we conclude that $\mathcal{P}$ is isomorphic to the octonion plane, or that $S = \Lambda$ is a normal subgroup of $SG_0$. In the latter case, $\Psi$ indeed is a subgroup of the centralizer $\Xi$ of $\Lambda$. □

**Remark 3.3.** The reduced stabilizer of the octonion plane is isomorphic to the almost simple Lie group Spin$_{10}(\mathbb{R}, 1)$. Thus, the group of affine automorphisms of the octonion plane has dimension $45 + 17 = 62$.

### 3.4. The normalizer of $\Lambda$

We shall determine the normalizer $\Gamma$ of $\Lambda$ in $GL(P)$. To this end we consider the automorphism group $A$ of the Lie algebra $L_\Lambda \cong sl_2 H$. Notice that the adjoint representation $Ad$ is a Lie homomorphism from $\Gamma$ to $A$ whose kernel coincides with the centralizer $\Xi$ of $\Lambda$ in $GL(P)$. Observe that the map $
iota : H^2 \times H^2 \rightarrow H^2 \times H^2; (\vec{x}, \vec{y}) \mapsto (\vec{y}, \vec{x})$

is an element of $\Gamma$ and that $Ad \niota$ equals the automorphism $X \mapsto -\overline{X}$ of $sl_2 H$. From [9, §4(c)] we infer that $A = Ad(\Lambda \cdot \langle \niota \rangle)$. (The group of inner automorphism has index 2 in $A$ and $Ad \niota$ is an outer automorphism.) Indeed, we have that

\[(2) \quad \Gamma = \langle \niota \rangle \cdot \Lambda \cdot \Xi.\]

The subrepresentations of $\Phi$ on $H^2 \times \{\vec{0}\}$ and on $\{\vec{0}\} \times H^2$ are inequivalent, irreducible quaternion representations. Thus, the centralizer $\Xi$ of $\Lambda$ consists precisely of the following maps:

$\xi_{a,b} : H^2 \times H^2 \rightarrow H^2 \times H^2; (\vec{x}, \vec{y}) \mapsto (a \cdot \vec{x}, b \cdot \vec{y})$ with $a, b \in H^\times$.

**Proposition 3.5.** Let $\sigma, \tau : Spin(3) \rightarrow \mathbb{R}_{\text{pos}}$ be continuous maps and let $a, b \in Spin(3), r, s \in \mathbb{R}_{\text{pos}}$. Then the following holds:

(a) The map $\xi_{ra, sb} : (\vec{x}, \vec{y}) \mapsto (ra \vec{x}, sb \vec{y})$ is an isomorphism from $P^\sigma$ onto $P^\tau$ if and only if $\tau(h) = r^{-1}s(\sigma(b^{-1}ha))$ holds for every $h \in Spin(3)$.

(b) The map $\xi_{ra, sb} : (\vec{x}, \vec{y}) \mapsto (sb \vec{y}, ra \vec{x})$ is an isomorphism from $P^\sigma$ onto $P^\tau$ if and only if $\tau(h) = r^{-1}s(\sigma(a^{-1}h^{-1}b))^{-1}$ holds for all $h \in Spin(3)$.

**Proof.** Let $f$ be an element of $\langle \niota \rangle \cdot \Xi$. Observe that $f$ leaves the weight sphere $S$ invariant. Moreover, notice that $f$ centralizes the maximal compact subgroup $\Delta$ of $\Lambda$. This implies that $f$ is an isomorphism from $P^\sigma$ onto $P^\tau$ if and only if $f$ maps every $\Delta$-invariant line $U_{\sigma(h) \ast h}, h \in Spin(3)$, of $L^\sigma_0$ to a $\Delta$-invariant line $U_{\tau(h) \ast h}$ of $L^\tau_0$. A short computation shows:

\[
\{(\vec{x}, \sigma(h) \ast \vec{x}) | \vec{x} \in H^2 \}^{\xi_{ra, sb}} = \{(\vec{y}, r^{-1}s(\sigma(h)bha^{-1} \vec{y})) | \vec{y} \in H^2 \}
\]

\[
\{(\vec{x}, \sigma(h) \ast \vec{x}) | \vec{x} \in H^2 \}^{\xi_{ra, sb}} = \{(\vec{y}, r^{-1}s(\sigma(h)bha^{-1} \vec{y})) | \vec{y} \in H^2 \}.
\]

From these equations we infer easily the assertions of the proposition. □
Proposition 3.6. We consider a continuous function $\sigma : \text{Spin}(3) \to \mathbb{R}_{\text{pos}}$. Let $\mathcal{G}_0$ the stabilizer of the connected component of the automorphism group of $\mathcal{P}^\sigma$. Then the following statements are equivalent:

1. $\mathcal{P}^\sigma$ is isomorphic to the affine plane over the octonions.
2. $\sigma$ is a constant map.
3. $\mathcal{G}_0$ contains the group $\{\xi_{a,1} \mid a \in \text{Spin}(3)\}$.
4. $\mathcal{G}_0$ contains the group $\{\xi_{1,b} \mid b \in \text{Spin}(3)\}$.

Proof. Use 3.5(a) to derive $(2 \iff 3 \iff 4)$.

$(1 \Rightarrow 3)$: If $\mathcal{P}^\sigma$ is isomorphic to the affine plane over the octonions, then $\mathcal{G}_0 \cap \text{SL}(\mathcal{P})$ is isomorphic to $\text{Spin}_{10}(\mathbb{R}, 1)$. Moreover, the centralizer of $\Lambda \cong \text{SL}_2 \mathbb{H}$ in $\text{Spin}_{10}(\mathbb{R}, 1)$ is locally isomorphic to $\text{SU}_2 \mathbb{C} \cdot \text{SU}_2 \mathbb{C}$. (Up to conjugation, the Lie algebra $\mathfrak{so}_{10}(\mathbb{R}, 1)$ contains only one subalgebra which is isomorphic to $\mathfrak{so}_6(\mathbb{R}, 1) \cong \mathfrak{sl}_2 \mathbb{H}$, see [7, 6.9].) Therefore, the maximal compact subgroup $\{\xi_{a,b} \mid a, b \in \text{Spin}(3)\}$ of the centralizer of $\Lambda$ in $\text{GL}(\mathcal{P})$ consists of automorphisms of $\mathcal{P}^\sigma$.

$(2 \Rightarrow 1)$: Since the automorphism group of the octonion plane $\mathcal{P}$ contains a subgroup isomorphic to $\text{SL}_2 \mathbb{H}$ (see above), we infer that $\mathcal{P}$ is isomorphic to $\mathcal{P}^\sigma$ for some $\sigma$ by the Classification Theorem (2.8). By “1 $\Rightarrow$ 2”, $\sigma$ is a constant map, i.e., $\sigma \equiv r$ holds for some $r \in \mathbb{R}_{\text{pos}}$. If $\tau \equiv s$, $s \in \mathbb{R}_{\text{pos}}$, is an arbitrary constant map, then $\xi_{1,r/s}$ is an isomorphism between $\mathcal{P}^\tau$ and $\mathcal{P}^\sigma$, whence $\mathcal{P}^\tau$ is isomorphic to the octonion plane. □

Theorem 3.7. Let $\sigma, \tau : \text{Spin}(3) \to \mathbb{R}_{\text{pos}}$ be continuous functions. Then $\mathcal{P}^\sigma$ and $\mathcal{P}^\tau$ are isomorphic if and only if there exists $a, b \in \text{Spin}(3)$ and $r \in \mathbb{R}_{\text{pos}}$ such that one of the following two properties is satisfied:

\[ \tau(h) = r\sigma(ahb) \text{ for all } h \in \text{Spin}(3) \text{ or} \]

\[ \tau(h) = r|\sigma(ahb)|^{-1} \text{ for all } h \in \text{Spin}(3). \]

Proof. If one of the two properties above holds, then use 3.5 to obtain an isomorphism between $\mathcal{P}^\sigma$ and $\mathcal{P}^\tau$.

Conversely, suppose that $\mathcal{P}^\sigma$ and $\mathcal{P}^\tau$ are isomorphic. Then there exists an $\mathbb{R}$-linear map $f : \mathbb{H}^4 \to \mathbb{H}^4$ which maps $\mathcal{L}_0^\sigma$ onto $\mathcal{L}_0^\tau$.

If $\sigma \equiv t$, $t \in \mathbb{R}_{\text{pos}}$, is a constant map, then $\mathcal{P}^\sigma$ and, hence, $\mathcal{P}^\tau$ are isomorphic to the octonion plane (3.6). This implies that $\tau \equiv t'$, $t' \in \mathbb{R}_{\text{pos}}$ is a constant map (3.6). Thus, $\tau(h) = t'/t \cdot \sigma(h)$ holds for all $h \in \text{Spin}(3)$.

Finally, suppose that neither $\sigma$ nor $\tau$ is a constant map. Then the reduced stabilizers of $\mathcal{P}^\sigma$ is an almost direct product of $\Lambda$ and some compact group, see 3.2. Since this assertion holds for $\mathcal{P}^\tau$ as well, $f$ is an element of the normalizer of $\Lambda$. Modifying $f$ with elements of $\Lambda$, we may achieve that $f$ is an element of $\langle l \rangle \cdot \Xi$, and the desired property follows from 3.5. □
Theorem 3.8. Let $\mathcal{P}$ be a 16-dimensional translation plane with automorphism group $G$ and reduced stabilizer $SG_0$. If $G$ contains a subgroup locally isomorphic to $SL_2 \mathbb{H}$, then only the following (mutually exclusive) possibilities can occur:

(1) $SG_0$ is isomorphic to $Spin_{10}(\mathbb{R},1)$ and $\mathcal{P}$ is isomorphic to the octonion plane.

(2) $\mathcal{P}$ is isomorphic to $\mathcal{P}^\sigma$, where $\sigma : Spin(3) \to \mathbb{R}_{pos}$ is a continuous function which is not constant and depends only on the real part of its argument. The reduced stabilizer of $\mathcal{P}^\sigma$ is the almost direct product of $\Lambda \cong SL_2 \mathbb{H}$ and the group

$$\Psi = \{(x,y) \mapsto (a\vec{x}, a\vec{y}) | a \in Spin(3)\} \cong SU_2 \mathbb{C}.$$

In particular, the dimension of $G$ equals 35.

(3) $\mathcal{P}$ is isomorphic to $\mathcal{P}^\sigma$, where $\sigma$ is derived from a continuous, not constant function $\rho : [0;1] \to \mathbb{R}_{pos}$, as follows:

$$\sigma : Spin(3) = \{u+jv | u,v \in \mathbb{C}, |u|^2 + |v|^2 = 1\} \to \mathbb{R}_{pos}; u+jv \mapsto \rho(|u|).$$

In this case, the reduced stabilizer of $\mathcal{P}^\sigma$ is the almost direct product of $\Lambda \cong SL_2 \mathbb{H}$ and the group

$$\Psi = \{(x,y) \mapsto (a\vec{x}, b\vec{y}) | a,b \in Spin(3) \cap \mathbb{C}\} \cong SO_2 \mathbb{R} \times SO_2 \mathbb{R}.$$

In particular, the dimension of $G$ equals 34.

(4) The reduced stabilizer of $\mathcal{P}^\sigma$ an almost direct product of $\Lambda$ and an at most 1-dimensional compact group, and $dim G \in \{32,33\}$.

Proof. By the Classification Theorem (2.8), there exists a continuous function $\sigma : Spin(3) \to \mathbb{R}_{pos}$ such that $\mathcal{P}$ is isomorphic to $\mathcal{P}^\sigma$.

We suppose that $\mathcal{P}^\sigma$ is not isomorphic to the octonion plane. Then $\sigma$ is not constant (3.6) and the reduced stabilizer of $\mathcal{P}^\sigma$ is an almost direct product of $\Lambda$ and a connected compact group $\Psi$, cf. 3.2. Indeed, $\Psi$ is a subgroup of the centralizer $\Xi$ of $\Lambda$ in $GL(P)$ and, hence, $\Psi$ is contained in the maximal compact subgroup $\Xi' = \{\xi_{a,b} | a,b \in Spin(3)\} \subseteq SU_2 \mathbb{C} \times SU_2 \mathbb{C}$.

By 3.5(a), we infer that

$$\Psi = \{\xi_{a,b} | a,b \in Spin(3), \sigma(b^{-1}ha) = \sigma(h) \text{ for all } h \in Spin(3)\}.$$

We emphasize that we are allowed to replace $\Psi$ by $\xi_{a,b}^{-1}\Psi\xi_{a,b}$ for arbitrary $a,b \in Spin(3)$: This corresponds to the replacement of $\mathcal{P}^\sigma$ by the isomorphic plane $\mathcal{P}^{\tau}$, $\tau(h) = \sigma(b^{-1}ha)$, see 3.7. Checking the connected subgroups of $\Xi' \cong SU_2 \mathbb{C} \times SU_2 \mathbb{C}$ yields the following fact: Up to conjugation, there are precisely the following possibilities for $\Psi$:

(i) $\Psi$ has dimension 0 or 1.
(ii) $\Psi = \{\xi_{a,a} | a \in Spin(3)\}$.
(iii) $\Psi = \{\xi_{a,b} | a,b \in Spin(3) \cap \mathbb{C}\}$.
(iv) $\Psi$ contains the group $\{\xi_{a,1} | a \in Spin(3)\}$ or the group $\{\xi_{1,b} | b \in Spin(3)\}$. 
Since \( \mathcal{P}^\sigma \) is not isomorphic to the octonion plane, case (iv) can not occur, see 3.6. Using Equation (3), it is not hard to see that \( \Psi \) equals \( \{ \xi_{a,b} | a \in \text{Spin}(3) \} \) if and only if \( \sigma(h) \) depends only on the real part of \( h \).

If \( \sigma \) is one of the functions specified in Part (3) of the theorem, then we derive that \( \Psi = \{ \xi_{a,b} | a,b \in \text{Spin}(3) \cap \mathbb{C} \} \) from Equation (3). Conversely, suppose that \( \xi_{a,b} \) is an automorphism of \( \mathcal{P}^\sigma \) for every \( a,b \in \text{Spin}(3) \). Let \( u + jv \) be an arbitrary element of \( \text{Spin}(3) \) with \( u,v \in \mathbb{C}, |u|^2 + |v|^2 = 1 \). If \( u = |u|e^{ir} \) and \( v = |v|e^{is} \) are the polar decompositions, then we put \( a = e^{-i(r+s)/2} \) and \( b = e^{i(r-s)/2} \). Then \( \xi_{a,b} \) is an automorphism of \( \mathcal{P}^\sigma \) and we infer from 3.5(a) that

\[
\sigma(u + jv) = \sigma(e^{-i(r-s)/2}(|u|e^{ir} + j|v|e^{is})e^{-i(r+s)/2}) = \sigma(|u| + j|v|)
= \sigma(|u| + j\sqrt{1 - |u|^2}),
\]

whence \( \sigma \) depends only on \( |u| \), as asserted in Part (3). This finishes the proof. \( \square \)

**Corollary 3.9.** Let \( \mathcal{P} \) be a 16-dimensional locally compact translation plane admitting \( \text{SL}_2 \mathbb{H} \) as a group of collineations. If the dimension of the automorphism group of \( \mathcal{P} \) strictly exceeds 35, then \( \mathcal{P} \) is isomorphic to the octonion plane. \( \square \)

**References**


[5] , *Sixteen-dimensional locally compact translation planes admitting \( \text{SU}_4 \mathbb{C} \cdot \text{SU}_2 \mathbb{C} \) or \( \text{SU}_4 \mathbb{C} \cdot \text{SL}_2 \mathbb{R} \) as a group of collineations*, Abh. Math. Sem. Univ. Hamburg, 70 (2000), 137-163, CMP 1 809 542, Zbl 0992.51007.


Received April 9, 2002.

Technische Universität Braunschweig
Institut für Analysis, Abt. Topologie
Pockelsstrasse 14
38 106 Braunschweig
Germany
E-mail address: h.loewe@tu-bs.de
We prove that an operator space is completely isometric to a ternary ring of operators if and only if the open unit balls of all of its matrix spaces are bounded symmetric domains. From this we obtain an operator space characterization of C*-algebras.

1. Introduction.

In the category of operator spaces, that is, subspaces of the bounded linear operators $B(H)$ on a complex Hilbert space $H$ together with the induced matricial operator norm structure, objects are equivalent if they are completely isometric, i.e., if there is a linear isomorphism between the spaces which preserves this matricial norm structure. Since operator algebras, that is, subalgebras of $B(H)$, are motivating examples for much of operator space theory, it is natural to ask if one can characterize which operator spaces are operator algebras. One satisfying answer was given by Blecher, Ruan and Sinclair in [10], where it was shown that among operator spaces $A$ with a (unital but not necessarily associative) Banach algebra product, those which are completely isometric to operator algebras are precisely the ones whose multiplication is completely contractive with respect to the Haagerup norm on $A \otimes A$. (For a completely bounded version of this result, see [7].)

A natural object to characterize in this context are the so called ternary rings of operators (TRO’s). These are subspaces of $B(H)$ which are closed under the ternary product $xy^*z$. This class includes C*-algebras. TRO’s, like C*-algebras, carry a natural operator space structure. In fact, every TRO is (completely) isometric to a corner $pA(1-p)$ of a C*-algebra $A$. TRO’s are important because, as shown by Ruan [35], the injectives in the category of operator spaces are TRO’s (corners of injective C*-algebras) and not, in general, operator algebras. (For the dual version of this result see [15].) Injective envelopes of operator systems and of operator spaces ([23] and [35]) have proven to be important tools, see for example [9]. The characterization of TRO’s among operator spaces is the subject of this paper. (See Theorem 5.3.)

Closely related to TRO’s are the so called JC*-triples, norm closed subspaces of $B(H)$ which are closed under the triple product $(xy^*z + zy^*x)/2$. 

339
These generalize the class of TRO’s and have the property, as shown by Harris in [25], that isometries coincide with algebraic isomorphisms. It is not hard to see this implies that the algebraic isomorphisms in the class of TRO’s are complete isometries, since for each TRO $A$, $\mathcal{M}_n(A)$ is a JC*-triple. (For the converse of this, see [24, Proposition 2.1].) As a consequence, if an operator space $X$ is completely isometric to a TRO, then the induced ternary product on $X$ is unique, i.e., independent of the TRO.

Building on the pioneering work of Arveson ([3] and [4]) on noncommutative analogs of the Choquet and Shilov boundaries, Hamana (see [24]) proved that every operator space $A$ has a unique enveloping TRO $T(A)$ which is an invariant of complete isometry and has the property that for any TRO $B$ generated by a realization of $A$, there exists a homomorphism of $B$ onto $T(A)$. The space $T(A)$ is also called the *Hilbert C*-envelope of $A$. The work in [8] suggests that the *Hilbert C*-envelope is an appropriate noncommutative generalization to operator spaces of the classical theory of Shilov boundary of function spaces.

It is also true that a commutative TRO ($xy^*z = zy^*x$) is an associative JC*-triple and hence by [19, Theorem 2], is isometric (actually completely isometric) to a complex $C_{\text{hom}}$-space, that is, the space of weak*-continuous functions on the set of extreme points of the unit ball of the dual of a Banach space which are homogeneous with respect to the natural action of the circle group, see [19]. Hence, if one views operator spaces as noncommutative Banach spaces, and C*-algebras as noncommutative $C(\Omega)$’s, then TRO’s and JC*-triples can be viewed as noncommutative $C_{\text{hom}}$-spaces.

As noted above, injective operator spaces, i.e., those which are the range of a completely contractive projection on some $B(H)$, are completely isometrically TRO’s; the so called *mixed injective* operator spaces, those which are the range of a contrative projection on some $B(H)$, are isometrically JC*-triples. The operator space classification of mixed injectives was begun by the authors in [32] and [33] and is ongoing.

 Relevant to this paper is another property shared by all JC*-triples (and hence all TRO’s). For any Banach space $X$, we denote by $X_0$ its open unit ball: $\{x \in X : \|x\| < 1\}$. The open unit ball of every JC*-triple is a *bounded symmetric domain*. This is equivalent to saying that it has a transitive group of biholomorphic automorphisms. It was shown by Koecher in finite dimensions (see [31]) and Kaup [28] in the general case that this is a defining property for the slightly larger class of JB*-triples. The only illustrative basic examples of JB*-triples which are not JC*-triples are the space $H_3(\mathcal{O})$ of $3 \times 3$ Hermitian matrices over the octonians and a certain subtriple of $H_3(\mathcal{O})$. These are called *exceptional* triples, and they cannot be represented as a JC*-triple. This holomorphic characterization has been useful as it gives an elegant proof, due to Kaup [29], that the range of a contractive projection on a JB*-triple is isometric to another JB*-triple. The same
statement holds for JC*-triples, as proven earlier by Friedman and Russo in [21]. Youngson proved in [38] that the range of a completely contractive projection on a C*-algebra is completely isometric to a TRO. These results, as well as those of [2] and [17], are rooted in the fundamental result of Choi-Effros [12] for completely positive projections on C*-algebras and the classical result ([30] and [18, Theorem 5]) that the range of a contractive projection on C(Ω) is isometric to a Cσ-space, hence a C_hom-space.

Motivated by this characterization for JB*-triples, we will give a holomorphic characterization of TRO’s up to complete isometry. We will prove in Theorem 5.3 that an operator space \( A \) is completely isometric to a TRO if and only if the open unit balls \( M_n(A)_0 \) are bounded symmetric domains for all \( n \geq 2 \). As a consequence, we obtain in Theorem 5.7 a holomorphic operator space characterization of C*-algebras as well. It should be mentioned that Upmeier (for the category of Banach spaces) in [37] and El Amin-Campoy-Palacios (for the category of Banach algebras) in [1], gave different but still holomorphic characterizations of C*-algebras up to isometry. We note in passing that injective operator spaces satisfy the hypothesis of Theorem 5.3, so we obtain that they are (completely isometrically) TRO’s based on deep results about JB*-triples rather than the deep result of Choi-Effros. (See Corollaries 5.5 and 5.6.)

We now describe the organization of this paper. Section 2 contains the necessary background and some preliminary results on contractive projections. In Section 3, three auxiliary ternary products are introduced and are shown to yield the original JB*-triple product upon symmetrization. Section 4 is devoted to proving that these three ternary products all coincide. Section 5 contains the statement and proof of the main result and its consequences.

2. Preliminaries.

An operator space will be defined as a normed space \( A \) together with a linearly isometric representation as a subspace of some \( B(H) \). This gives \( A \) a family of operator norms \( \| \cdot \|_n \) on \( M_n(A) \subset B(H^n) \). As proved in [34], an operator space can also be defined abstractly as a normed space \( A \) having a norm on \( M_n(A) \) \( (n \geq 2) \) satisfying certain properties. Each such family of norms is regarded as a “quantization” of the underlying Banach space. These properties give rise to an isometric representation of the operator space as a subspace of \( B(H) \) where the natural amplification maps preserve the matricial norm structure. This is analogous to (and generalizes) the way an abstract Banach space \( B \) can be isometrically embedded as a subspace of \( C(\Omega) \). The resulting operator space structure in this case is called \( MIN(B) \) and is seen as a commutative quantization of \( B \).
Two operator spaces $A$ and $B$ are $n$-isometric if there exists an isometry $\phi$ from $A$ onto $B$ such that the amplification mapping $\phi_n : M_n(A) \to M_n(B)$ defined by $\phi_n([a_{ij}]) = [\phi(a_{ij})]$ is an isometry. $A$ and $B$ are completely isometric if there exists a mapping $\phi$ from $A$ onto $B$ which is an $n$-isometry for all $n$. For other basic results about operator spaces, see [16].

The following definition is a Hilbert space-free generalization of the TRO’s mentioned in the introduction:

**Definition 2.1 (Zettl [39]).** A C*-ternary ring is a Banach space $A$ with ternary product $[x,y,z] : A \times A \times A \to A$ which is linear in the outer variables, conjugate linear in the middle variable, is associative:

$$[ab[cde]] = [a[dcb]e] = [ab[cde]],$$

and satisfies $\|xyz\| \leq \|x\|\|y\|\|z\|$ and $\|xxx\| = \|x\|^3$.

A TRO is a C*-ternary ring under any of the products $[xyz]_{\lambda} = \lambda xy^*z$, for any complex number $\lambda$ with $|\lambda| = 1$.

A linear map $\varphi$ between C*-ternary rings is a homomorphism if $\varphi([xyz]) = [\varphi(x), \varphi(y), \varphi(z)]$ and an anti-homomorphism if $\varphi([xyz]) = -[\varphi(x), \varphi(y), \varphi(z)]$.

The following is a Gelfand-Naimark representation theorem for C*-ternary rings:

**Theorem 2.2 ([39]).** For any C*-ternary ring $A$, $A = A_1 \oplus A_{-1}$, where $A_1$ and $A_{-1}$ are sub-C*-ternary rings, $A_1$ is isometrically isomorphic to a TRO $B_1$ and $A_{-1}$ is isometrically anti-isomorphic to a TRO $B_{-1}$.

It follows that $A_{-1} = 0$ if and only if $A$ is ternary isomorphic to a TRO. In Theorem 5.3, we shall show that under suitable assumptions on an operator space $A$, it becomes a C*-ternary ring with $A_{-1} = 0$ and the above ternary isomorphism is a complete isometry from $A$ with its original operator space structure to a TRO with its natural operator space structure.

An immediate consequence of our proof of Theorem 5.3 is an answer to a question posed by Zettl [39, p. 136]: For a C*-ternary ring $A$, $A_{-1} = 0$ if and only if $A$ is a JB*-triple (see the next definition) under the triple product

$$\{abc\} = \frac{1}{2}([abc] + [cba]).$$

The following definition generalizes the JC*-triples defined in the introduction:

**Definition 2.3 ([28]).** A JB*-triple is a Banach space $A$ with a product $D(x,y,z) = \{x \ y \ z\}$ which is linear in the outer variables, conjugate linear in the middle variable, is commutative: $\{x \ y \ z\} = \{z \ y \ x\}$, satisfies an
associativity condition:

\[(1) \quad [D(x,y), D(a,b)] = D(\{x \circ y \circ a\}, b) - D(a, \{b \circ x \circ y\})\]

and has the topological properties that:

(i) \(\|D(x,x)\| = \|x\|^2\),

(ii) \(D(x,x)\) is Hermitian (in the sense that \(\|e^{itD(x,x)}\| = 1\)) and has positive spectrum in the Banach algebra \(B(A)\).

We abbreviate \(D(x,x)\) to \(D(x)\).

As noted in the introduction, JC*-triples (and hence TRO’s and C*-algebras) are examples of JB*-triples. Other examples include any Hilbert space, and the spaces of symmetric and anti-symmetric elements of \(B(H)\) under a transpose map defined by a conjugation.

If one ignores the norm and the topological properties in Definition 2.3, the algebraic structure which results, called a Jordan triple system, or Jordan pair, has a life of its own, \([31]\). Note that \((1)\) can be written as

\[(2) \quad \{x, y, \{abz\}\} - \{a, b, \{xyz\}\} = \{\{xya\}, b, z\} - \{a, \{yxb\}, z\} .\]

For easy reference we record here two identities for Jordan triple systems which can be derived from \((1)\) \([31, JP8, JP16]\).

\[(3) \quad 2D(x, \{yxz\}) = D(\{xyx\}, z) + D(\{xzx\}, y)\]

\[(4) \quad \{\{xya\}, b, z\} - \{a, \{yxb\}, z\} = \{x, \{bay\}, z\} - \{\{abx\}, y, z\} .\]

We will now list some facts about JB*-triples that are relevant to our paper. A survey of the basic theory can be found in \([36]\). As proved by Kaup \([28]\), JB*-triples are in 1-1 isometric correspondence with Banach spaces whose open unit ball is a bounded symmetric domain. The triple product here arises from the Lie algebra of the group of biholomorphic automorphisms. This Lie algebra is the space of complete vector fields on the open unit ball and consists of certain polynomials of degree at most 2. The quadratic term in each of these polynomials is determined by the constant term. For a bounded symmetric domain, the constant terms which occur exhaust \(A\). Thus, linearizing the quadratic term for every element \(a \in A\) leads to a triple product on \(A\).

It is this correspondence which motivates the study of the more general JB*-triples. Indeed, the proofs of two important facts follow naturally from the holomorphic point of view \([29]\). Firstly, the isometries between JB*-triples are precisely the algebraic isomorphisms. From this follows the important fact, used several times in this paper, that, unlike the case for binary products, the triple product of a JB*-triple is unique. Secondly, the
range of a contractive projection \( P \) on a JB*-triple \( Z \) is isometric to a JB*-triple. More precisely, \( P(Z) \) is a JB*-triple under the norm and linear operations it inherits from \( Z \) and the triple product \( \{xyz\}_{P(Z)} := P(\{xyz\}_Z) \), for \( x, y, z \in P(Z) \).

In the context of JC*-triples, these facts were proven by functional analytic methods in \([25]\) and \([21]\) respectively. These facts show that JB*-triples are a natural category in which to study isometries and contractive projections. Recently, in \([13]\) the authors with C.-H. Chu have shown that \( w^* \)-continuous contractive projections on dual JB*-triples (called JBW*-triples) preserve the Jordan triple generalization of the Murray-von-Neumann type decomposition established in \([26]\) and \([27]\). Two other properties of contractive projections were used in that work and will be needed in the present paper. They consist of two conditional expectation formulas for contractive projections on JC*-triples (\([20]\), Corollary 1), namely

\[
P \{Px,Py,Pz\} = P \{Px,Py,z\} = P \{Px,y,Pz\};
\]

and the fact that the range of a bicontractive projection on a JC*-triple is a subtriple \([20]\), Proposition 1). Recall that a projection \( P \) is said to be bicontractive if \( \|P\| \leq 1 \) and \( \|I - P\| \leq 1 \).

Let \( A \) be a JB*-triple. For any \( a \in A \), there is a triple functional calculus, that is, a triple isomorphism of the closed subtriple \( C(a) \) generated by \( a \) onto the commutative C*-algebra \( C_0(\mathrm{Sp} D(a,a) \cup \{0\}) \) of continuous functions vanishing at zero, with the triple product \( fgh \). Any JBW*-triple (defined in the previous paragraph) has the property that it is the norm closure of the linear span of its tripotents, that is, elements \( e \) with \( e = \{eee\} \). A unitary tripotent is a tripotent \( v \) such that \( D(v,v) = Id \). For a C*-algebra, tripotents are the partial isometries and for unital C*-algebras, unitary tripotents are precisely the unitaries. For tripotents \( u \) and \( v \), algebraic orthogonality, i.e., \( D(u,v) = 0 \), coincides with Banach space orthogonality: \( \|u \pm v\| = 1 \). For \( a \) and \( b \) in \( A \), we will denote the property \( D(a,b) = 0 \) by \( a \perp b \).

As proved in \([14]\), the second dual \( A^{**} \) of a JB*-triple \( A \) is a JBW*-triple containing \( A \) as a subtriple. Multiplication in a JBW*-triple is norm continuous and, as proved in \([5]\), separately \( w^* \)-continuous.

We close this section of preliminaries with an elementary proposition showing that certain concrete projections are contractive.

**Proposition 2.4.** Let \( A \) be an operator space in \( B(H) \).

(a) Define a projection \( P \) on \( M_2(A) \) by

\[
P \left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = \frac{1}{2} \begin{bmatrix} a + b & a + b \\ 0 & 0 \end{bmatrix}.
\]

Then \( \|P\| \leq 1 \). Moreover, the restriction of \( P \) to \( \left\{ \begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix} : a, b \in A \right\} \) is bicontractive.
(b) Let $P_{11} : M_2(A) \to M_2(A)$ be the map
\[
\begin{bmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{bmatrix}
\to
\begin{bmatrix}
a_{11} & 0 \\
0 & 0
\end{bmatrix},
\]
and similarly for $P_{12}, P_{21}, P_{22}$. Then $P_{ij}$ is contractive and $P_{11} + P_{21}$, $P_{11} + P_{12}$, and $P_{11} + P_{22}$ are bicontractive. More generally, the $P_{ij}$ : $M_n(A) \to M_n(A)$ are contractive and for any subset $S \subset \{1, 2, \ldots, n\}$,
\[
\sum_{i \in S} \sum_{j=1}^n P_{ij}
\]
and
\[
\sum_{j \in S} \sum_{i=1}^n P_{ij}
\]
are bicontractive.

(c) The projections $P : M_2(A) \to M_2(A)$ and $Q : P(M_2(A)) \to P(M_2(A))$ defined by
\[
P \left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = \frac{1}{2} \begin{bmatrix} a + d & b + c \\ b + c & a + d \end{bmatrix},
\]
and
\[
Q \left( \begin{bmatrix} a & b \\ b & a \end{bmatrix} \right) = \frac{1}{2} \begin{bmatrix} a + b & a + b \\ a + b & a + b \end{bmatrix}
\]
are bicontractive.

Proof. We omit the proofs of (a) and (b). To prove (c), since for example $I - P = (I - (2P - I))/2$ and $P = (I + (2P - I))/2$, it suffices to show that $2P - I$ and $2Q - I$ are contractive. But
\[
(2P - I) \left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = \begin{bmatrix} d & c \\ b & a \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},
\]
and
\[
(2Q - I) \left( \begin{bmatrix} a & b \\ b & a \end{bmatrix} \right) = \begin{bmatrix} b & a \\ a & b \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ a & b \end{bmatrix}.
\]

3. Additivity of the ternary products.

Throughout this section, $A \subset B(H)$ will be an operator space such that the open unit ball $M_2(A)_0$ is a bounded symmetric domain. Let $\{\cdot, \cdot, \cdot\}_{M_2(A)}$ denote the associated JB*-triple product on $M_2(A)$. Note that although $M_2(A)$ inherits the norm and linear structure of $M_2(B(H)) = B(H \oplus H)$, its triple product $\{\cdot, \cdot, \cdot\}_{M_2(A)}$ in general differs from the concrete triple product $(XYZ + ZXY)/2$ of $B(H \oplus H)$. In fact, the results of this section would become trivial if these two triple products were the same.

By properties of contractive projections and the uniqueness of the triple product, $A$, being linearly isometric to $P_{ij}(M_2(A))$ becomes a JB*-triple whose triple product $\{xyz\}_A$ is given, for example, by
\[
\left( \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z & 0 \\ 0 & 0 \end{bmatrix} \right)_A = P_{11} \left( \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z & 0 \\ 0 & 0 \end{bmatrix} \right)_{M_2(A)}.
\]
and similarly using the other $P_{ij}$. Usually we shall just use the notation \{\cdots\} for either of the triple products \{xyz\}_A and \{\cdot,\cdot,\cdot\}_{M_2(A)}$. Lemma 3.6 shows that the projection $P_{11}$ could be removed in this definition.

We assume $A$ is as above and proceed to define (in Definition 3.7) three auxiliary ternary products, denoted $[\cdot,\cdot,\cdot]$, $(\cdot,\cdot,\cdot)$, and $\langle\cdot,\cdot,\cdot\rangle$ and show their relation to \{\cdot,\cdot,\cdot\}. We begin with a sequence of lemmas which establish some properties of the terms in the following identity, where $a,b,c \in A$:

\[
(6) \quad \begin{bmatrix} a & a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & c \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}
\]

\[
+ \begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & c \\ 0 & 0 \\ 0 & 0 \end{bmatrix}
\]

\[
+ \begin{bmatrix} 0 & a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & c \\ 0 & 0 \\ 0 & 0 \end{bmatrix}
\]

\[
+ \begin{bmatrix} 0 & a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} c & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.
\]

It will be shown in Lemma 3.2 that the left side of (6) has the form

\[
\begin{bmatrix} x & y \\ z & w \end{bmatrix},
\]

where $(x + y)/2 = \{abc\}$. In Lemmas 3.4-3.6, each term on the right side of (6) will be analyzed.

**Remark 3.1.** The space

\[\tilde{A} = \{ \tilde{a} = \begin{bmatrix} a & a \\ 0 & 0 \end{bmatrix} : a \in A \}\]

with the triple product

\[
(7) \quad \{\tilde{a}\tilde{b}\tilde{c}\}_{\tilde{A}} := \begin{bmatrix} 2 \{abc\} & 2 \{abc\} \\ 0 & 0 \end{bmatrix}
\]

and the norm of $M_2(A)$, is a JB$^*$-triple.

Note that by Proposition 2.4(a), $\tilde{A}$ is a subtriple of $M_2(A)$, but we do not know a priori that its triple product is given by (7).

**Proof.** The proposed triple product, which we denote by $\{\tilde{a}\tilde{b}\tilde{c}\}$, is obviously linear and symmetric in $\tilde{a}$ and $\tilde{c}$, and conjugate linear in $\tilde{b}$. Since, for example,

\[
\{\tilde{a}\tilde{b}\{\tilde{c}\tilde{d}\}\} = \begin{bmatrix} 2 \{ab\{cde\}\} & 2 \{ab\{cde\}\} \\ 0 & 0 \end{bmatrix},
\]

the main identity (2) is satisfied.
From $\left\| \begin{bmatrix} a & a \\ 0 & 0 \end{bmatrix} \right\| = \sqrt{2}\|a\|$ one obtains $\|\{\bar{a}a\}\| = \|\bar{a}\|^2$, $\|\{\bar{abc}\}\| \leq \|\bar{a}\|\|\bar{b}\|\|\bar{c}\|$ and hence $\|D(a)\| = \|\bar{a}\|^2$.

Since $e^{i\theta D(x)}y = (e^{i\theta D(x)}y)$, $\|e^{i\theta D(x)}y\| = \sqrt{2}\|e^{2i\theta D(x)}y\| = \sqrt{2}\|y\| = \|\bar{y}\|$, so $D(x)$ is Hermitian.

Finally, for $\lambda < 0$, the inverse of $\lambda - D(x)$ is given by

$\lambda \mapsto \left[ \begin{array}{cc} -2D(x)^{-1}y & (\lambda - 2D(x))^{-1}y \\ 0 & 0 \end{array} \right]$.

Hence, $Sp_{B(\bar{A})}(D(x)) \subset [0, \infty)$.

Lemma 3.2. For $a, b, c \in A$, there exist $x, y, z, w \in A$ such that

$\left\{ \begin{bmatrix} a & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & c \\ 0 & 0 \end{bmatrix} \right\} = \begin{bmatrix} x & y \\ z & w \end{bmatrix}$,

and $(x + y)/2 = \{abc\}$.

Proof. Consider the projection $P$ defined in Proposition 2.4(a). By (5),

$P\left( \left\{ \begin{bmatrix} a & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & c \\ 0 & 0 \end{bmatrix} \right\} \right) = P\left( \left\{ \begin{bmatrix} a & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b/2 & b/2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & c \\ 0 & 0 \end{bmatrix} \right\} \right)$.

By Remark 3.1 and the uniqueness of the triple product in a JB$^*$-triple,

$P\left( \left\{ \begin{bmatrix} a & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & c \\ 0 & 0 \end{bmatrix} \right\} \right) = 2\begin{bmatrix} \{abc\} & \{abc\} \\ 0 & 0 \end{bmatrix}$.

Thus, if

$\left\{ \begin{bmatrix} a & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & c \\ 0 & 0 \end{bmatrix} \right\} = \begin{bmatrix} x & y \\ z & w \end{bmatrix}$,

then

$\begin{bmatrix} \{abc\} & \{abc\} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} (x + y)/2 & (x + y)/2 \\ 0 & 0 \end{bmatrix}$.

□

It will be shown below in the proof of Lemma 3.8 that $x = y = \{abc\}$ and that each $z = w = 0$.

Lemma 3.3. For each $a, b \in A$,

$\begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \perp \begin{bmatrix} 0 & 0 \\ 0 & b \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix} \perp \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix}$.
Suppose first that $a = \sum \lambda_i u_i$ where $\lambda_i > 0$ and the $u_i$ are tripotents in $A$, and similarly for $b = \sum \mu_j v_j$. Because the image of a bicontractive projection is a subtriple ([20, Proposition 1]), $U_i := \begin{bmatrix} u_i & 0 \\ 0 & 0 \end{bmatrix}$ and $V_j := \begin{bmatrix} 0 & 0 \\ 0 & v_j \end{bmatrix}$ are tripotents, and since they are orthogonal in $B(H \oplus H)$, $\|U_i \pm V_j\| = 1$. Hence $D(U_i, V_j) = 0$ in (the abstract triple product of) $M_2(A)$ and so for all $x, y, z, w \in A$,

$$\left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x & y \\ z & w \end{bmatrix} \right\} = \sum_{i,j} \lambda_i \mu_j \left\{ \begin{bmatrix} u_i & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & v_j \end{bmatrix} \begin{bmatrix} x & y \\ z & w \end{bmatrix} \right\} = 0.$$}

For the general case, note that, by [16, 3.2.1], there is an operator space structure on the dual of any operator space $A$ such that the canonical inclusion of $A$ into $A^{**}$ is a complete isometry. Moreover, by [6, Theorem 2.5] the norm structure on $M_n(A^{**})$ coincides with that obtained from the identification $M_n(A^{**}) = M_n(A)^{**}$. Hence, for all $n$, $M_n(A^{**})$ is a JBW*-triple containing $M_n(A)$ as subtriple. Since each element of $A$ can be approximated in norm by finite linear combinations of tripotents in $A^{**}$, the first statement in the lemma follows from the norm continuity of the triple product.

Since interchanging rows is an isometry, hence an isomorphism, the second statement follows. \hfill $\Box$

**Lemma 3.4.** Let $a, b, c \in A$. Then

$$\left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix} \right\} = 0,$$

(8)

$$\left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix} \right\} = 0.$$}

**Proof.** To prove the first statement, let $X$ denote

$$\left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix} \right\}.$$}

By (5),

$$P_{11}(X) = P_{11} \left( \left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix} \right\} \right) = 0.$$}

Similarly, $(P_{11} + P_{21})(X) = (P_{21} + P_{22})(X) = 0$, so that $X = \begin{bmatrix} 0 & x \\ 0 & 0 \end{bmatrix}$. 
Let $X' = \begin{bmatrix} 0 & 0 \\ 0 & x \end{bmatrix}$. We claim that for any $Y \in M_2(A)$, $\{XX'Y\} = 0$.

Indeed, with $A = \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix}$, $C = \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix}$, we have $A \perp X'$, $C \perp X'$ and by (4),

$$\{XX'Y\} = \{ABC\}X'Y = \{C\{BAX'\}Y + A\{X'CB\}Y - \{CX'A\}BY\} = 0.$$

Thus $D(X, X') = 0$, which, by [22, Lemma 1.3(a)], implies that $X$ and $X'$ are orthogonal in the Banach space sense: $\|X \pm X'\| = \max(\|X\|, \|X'\|)$. Since $\|X + X'\| = \sqrt{2}\|x\|$, it follows that $x = 0$. The second assertion is proved similarly, using $X = \begin{bmatrix} 0 & 0 \\ x & 0 \end{bmatrix}$, $X' = \begin{bmatrix} 0 & 0 \\ 0 & x \end{bmatrix}$.

By interchanging rows and columns, it follows that the following triple products all vanish (the last three by orthogonality):

(9) $\left\{ \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & c \end{bmatrix} \right\} = \left\{ \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\}$,

(10) $\left\{ \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} 0 & 0 \\ c & 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & c \end{bmatrix} \right\}$,

(11) $\left\{ \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & c \end{bmatrix} \right\} = \left\{ \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ c & 0 \end{bmatrix} \right\}$.

For use in Lemma 5.2, we adjoin

$\left\{ \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ c & 0 \end{bmatrix} \right\} = \left\{ \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix} \right\} = 0$,

and

$\left\{ \begin{bmatrix} 0 & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & c \end{bmatrix} \right\} = 0$.

**Lemma 3.5.** For $a, b, c \in A$, there exists $z \in A$ such that

$$\left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix} \right\} = \begin{bmatrix} z & 0 \\ 0 & 0 \end{bmatrix}.$$

**Proof.** Let $X$ denote $\left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & c \\ 0 & 0 \end{bmatrix} \right\}$. By (5), $(P_{12} + P_{22})(X) = 0$ and $(P_{12} + P_{21})(X) = 0$. \qed
Lemma 3.6. For \(a,b,c \in A\),
\[
\begin{bmatrix}
0 & a \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & b \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & c \\
0 & 0
\end{bmatrix}
= \begin{bmatrix}
0 & \{abc\} \\
0 & 0
\end{bmatrix}.
\]

Proof. Since \(P_{11} + P_{12}\) and \(P_{12} + P_{22}\) are bicontractive, the intersection of their ranges is a subtriple. Since \(A\) is a JB\(^*\)-triple under the product induced by \(P_{12}\), and triple products are unique, the result follows. \(\square\)

As noted in the proof of Lemma 3.3, interchanging rows or columns is an isometry, hence an isomorphism. Therefore we also have, for example,
\[
\begin{bmatrix}
a & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
b & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
c & 0 \\
0 & 0
\end{bmatrix}
= \begin{bmatrix}
\{abc\} & 0 \\
0 & 0
\end{bmatrix},
\]
and so forth.

In Proposition 2.4(b) we have defined projections \(P_{ij} : M_n(A) \to M_n(A)\) as follows: If \(X = [x_{ij}] \in M_n(A)\), then \(P_{ij}(X)\) is the element of \(M_n(A)\) with \(x_{ij}\) in the \((i,j)\) entry and zeros elsewhere. In what follows, we shall use the maps \(p_{ij} : M_n(A) \to A\) defined by \(p_{ij}(X) = x_{ij}\) for \(X = [x_{ij}] \in M_n(A)\).

Definition 3.7. Define a ternary product \([a,b,c]\) or \([abc]\) on \(A\) by
\[
[a,b,c] = 2p_{11} \left(\begin{bmatrix}
0 & a & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & b & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & c & 0 \\
0 & 0 & 0
\end{bmatrix}\right).
\]
Similarly, define two more ternary products \((abc)\) and \(\langle abc \rangle\) as follows:
\[
(abc) = 2p_{11} \left(\begin{bmatrix}
a & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
b & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
c & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}\right)
\]
and
\[
\langle abc \rangle = 2p_{11} \left(\begin{bmatrix}
0 & 0 & a \\
c & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & b \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & c \\
0 & 0 & 0
\end{bmatrix}\right).
\]

We treat first the ternary product \([a,b,c]\). Note that, by Lemma 3.5,
\[
\frac{1}{2} \begin{bmatrix}
[a,b,c] & 0 \\
0 & 0
\end{bmatrix} = \begin{bmatrix}
0 & a & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & b & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & c & 0 \\
0 & 0 & 0
\end{bmatrix},
\]
and that by interchanging suitable rows and columns,
\[
[a,b,c] = 2p_{21} \left(\begin{bmatrix}
0 & 0 & 0 \\
0 & a & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 \\
0 & b & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 \\
0 & c & 0
\end{bmatrix}\right)
\]
\[
= 2p_{12} \left(\begin{bmatrix}
a & 0 & 0 \\
0 & b & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}\right)
\]
\[
= 2p_{22} \left(\begin{bmatrix}
0 & 0 & 0 \\
a & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 \\
0 & b & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 \\
0 & c & 0
\end{bmatrix}\right).
\]
Lemma 3.8. For \( a, b, c \in A \),
\[
[a, b, c] + [c, b, a] = 2 \{abc\},
\]
and hence
\[
\|a, a, a\| = \|a\|^3.
\]

Proof. Given \( a, b, c \in A \), it follows from Lemma 3.2, Lemmas 3.4-3.6, Definition 3.7 and (6) that there are elements \( x, y, z, w \in A \) such that \( x + y = 2 \{abc\} \) and
\[
\begin{bmatrix}
x & y \\
z & w
\end{bmatrix} = \begin{bmatrix}
0 & 0 \\
0 & 0
\end{bmatrix} + \begin{bmatrix}
[abc]/2 & 0 \\
0 & 0
\end{bmatrix} + \begin{bmatrix}
0 & \{abc\} \\
0 & 0
\end{bmatrix} + \begin{bmatrix}
[cba]/2 & 0 \\
0 & 0
\end{bmatrix}.
\]
Hence \([abc]/2 + [cba]/2 = x = y = \{abc\}\) (and \(z = w = 0\)). \(\square\)

We shall see later in Proposition 4.5 that in fact \([abc] = (abc) = \langle abc \rangle\). First we shall show the analog of Lemma 3.8 for each of the ternary products \((abc)\) and \(\langle abc \rangle\). We note that, as above,
\[
\begin{bmatrix}
(abc)/2 & 0 \\
0 & 0
\end{bmatrix} = \begin{bmatrix}
a & 0 \\
0 & 0
\end{bmatrix} \begin{bmatrix}
0 & 0 \\
b & 0
\end{bmatrix} \begin{bmatrix}
c & 0
\end{bmatrix}
\]
and
\[
\begin{bmatrix}
\langle abc \rangle/2 & 0 \\
0 & 0
\end{bmatrix} = \begin{bmatrix}
0 & 0 \\
c & 0
\end{bmatrix} \begin{bmatrix}
0 & 0 \\
0 & b
\end{bmatrix} \begin{bmatrix}
0 & a
\end{bmatrix}.
\]
Moreover, by interchanging rows and/or columns,
\[
(abc) = 2p_{22} \left( \begin{bmatrix}
0 & 0 \\
0 & a
\end{bmatrix} \begin{bmatrix}
0 & b \\
0 & 0
\end{bmatrix} \begin{bmatrix}
0 & c
\end{bmatrix} \right)
\]
\[
= 2p_{12} \left( \begin{bmatrix}
0 & a \\
0 & 0
\end{bmatrix} \begin{bmatrix}
0 & 0 \\
0 & b
\end{bmatrix} \begin{bmatrix}
0 & c
\end{bmatrix} \right)
\]
\[
= 2p_{21} \left( \begin{bmatrix}
0 & 0 \\
a & 0
\end{bmatrix} \begin{bmatrix}
0 & b \\
0 & 0
\end{bmatrix} \begin{bmatrix}
c & 0
\end{bmatrix} \right)
\]
and
\[
\langle abc \rangle = 2p_{22} \left( \begin{bmatrix}
0 & c \\
0 & 0
\end{bmatrix} \begin{bmatrix}
b & 0 \\
0 & 0
\end{bmatrix} \begin{bmatrix}
0 & 0
\end{bmatrix} \right)
\]
\[
= 2p_{12} \left( \begin{bmatrix}
0 & 0 \\
c & 0
\end{bmatrix} \begin{bmatrix}
0 & 0 \\
0 & b
\end{bmatrix} \begin{bmatrix}
a & 0
\end{bmatrix} \right)
\]
\[
= 2p_{21} \left( \begin{bmatrix}
0 & c \\
0 & 0
\end{bmatrix} \begin{bmatrix}
0 & b \\
0 & 0
\end{bmatrix} \begin{bmatrix}
0 & a
\end{bmatrix} \right).
\]

Proposition 3.9. If \( A \) is an operator space such that \( M_2(A) \) is a bounded symmetric domain (and consequently \( M_2(A) \) and \( A \) are JB*-triples), then \( \langle abc \rangle + \langle cba \rangle = 2 \{abc\}_A \) and \( (abc) + (cba) = 2 \{abc\}_A \).
Proof. The proof for $(\cdot,\cdot,\cdot)$ is similar to the proof for $[\cdot,\cdot,\cdot]$, using instead the identity
\[
\begin{bmatrix} a & 0 \\ a & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \begin{bmatrix} c & 0 \\ c & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ c & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \begin{bmatrix} a & 0 \\ a & 0 \end{bmatrix} + \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \begin{bmatrix} c & 0 \\ c & 0 \end{bmatrix} + \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ c & 0 \end{bmatrix}
\]
and the projection
\[
P\left( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right) = \frac{1}{2} \begin{bmatrix} a + c & 0 \\ a + c & 0 \end{bmatrix}.
\]
To prove the statement for $\langle \cdot, \cdot, \cdot \rangle$ consider (cf. Remark 3.1) the space
\[\tilde{A} = \{ a = \begin{bmatrix} a & a \\ a & a \end{bmatrix} : a \in A \},\]
which is a subtriple of $M_2(A)$ since it is the range of a product $QP$ of the bicontractive projections $Q,P$ of Proposition 2.4(c). It follows as in the proof of Remark 3.1 that $\tilde{A}$ is a JB$^*$-triple under the triple product $\{ \cdot \cdot \cdot \}'$ defined by $\{ \tilde{abc} \}' = 4(\{abc\})$. To see this, let $D'(\tilde{x})\tilde{a} = \{\tilde{a}x\tilde{a}\}'$ and note that $\|\tilde{x}\| = 2\|x\|$, $D'(\tilde{x})\tilde{a} = 4(D(x)a)$, $e^{itD'(\tilde{x})\tilde{y}} = (e^{itD(x)}y)$ and that $(\lambda - D(\tilde{x}))^{-1}\tilde{y} = ((\lambda - D(x))^{-1}y)$. By the uniqueness of the triple product on $M_2(A)$, $\{\tilde{abc}\}' = \{\tilde{abc}\}'$. Hence, by expanding $\{\tilde{x}\tilde{y}\tilde{z}\} = \{ \begin{bmatrix} x & x \\ x & x \end{bmatrix} \begin{bmatrix} y & y \\ y & y \end{bmatrix} \begin{bmatrix} z & z \\ z & z \end{bmatrix} \}$ into computable terms,
\[
\begin{align*}
4 \{xyz\}
&= \{\tilde{x}\tilde{y}\tilde{z}\} \\
&= (\{xyz\} + (xyz)/2 + (zyx)/2 + [xyz]/2 + [xyz]/2 + (xyz)/2) \\
&= (3 \{xyz\} + \langle xyz \rangle/2 + \langle yxz \rangle/2).
\end{align*}
\]
This proves the statement for $\langle \cdot, \cdot, \cdot \rangle$. \qed

4. Equality of the ternary products.

In this section, we continue to assume that $A \subset B(H)$ is an operator space such that the open unit ball $M_2(A)_0$ is a bounded symmetric domain. We shall prove the equality of the three ternary products defined in Section 3.
Even though they agree, all three products are needed in the proof of the crucial Proposition 5.1.

In the following we shall let \( a \in M_2(A) \) denote \[
\begin{bmatrix}
a & 0 \\
0 & a
\end{bmatrix}
\] and \( \overline{a} \in M_2(A) \) denote \[
\begin{bmatrix}
0 & a \\
a & 0
\end{bmatrix}.
\]
By Lemmas 3.3 and 3.6, the ranges of \( P_{12} + P_{21} \) and \( P_{11} + P_{22} \) are invariant under the continuous functional calculus in a JB*-triple. In particular, for any \( \lambda > 0 \),

\[
a^\lambda = \begin{bmatrix} a^\lambda & 0 \\ 0 & a^\lambda \end{bmatrix} \quad \text{and} \quad \overline{a}^\lambda = \begin{bmatrix} 0 & a^\lambda \\ a^\lambda & 0 \end{bmatrix}.
\]

Here, \( a^\lambda \) is defined by the triple functional calculus in the JB*-triple \( M_2(A) \) and \( a^\lambda \) is defined by the triple functional calculus in the JB*-triple \( A \).

**Lemma 4.1.** Let \( \lambda, \mu, \nu \) be positive numbers and let \( a \in A \). Then

\[
a^{\lambda+\mu+\nu} = \left\{ a^\lambda a^\mu a^\nu \right\} = \left\{ \overline{a}^\lambda \overline{a}^\mu \overline{a}^\nu \right\} = \left\{ a^\lambda \overline{a}^\mu a^\nu \right\}
\]

and

\[
\overline{a}^{\lambda+\mu+\nu} = \left\{ \overline{a}^\lambda \overline{a}^\mu \overline{a}^\nu \right\} = \left\{ a^\lambda \overline{a}^\mu a^\nu \right\} = \left\{ \overline{a}^\lambda a^\mu a^\nu \right\}.
\]

**Proof.** \( a^{\lambda+\mu+\nu} = \left\{ a^\lambda a^\mu a^\nu \right\} \) is immediate from the functional calculus. The proofs of the other statements are all proved in the same way, for example,

\[
\left\{ \overline{a}^\lambda \overline{a}^\mu \overline{a}^\nu \right\} = \left\{ \begin{bmatrix}
0 & a^\lambda \\
a^\lambda & 0
\end{bmatrix} \begin{bmatrix}
a^\mu & 0 \\
0 & a^\mu
\end{bmatrix} \begin{bmatrix}
0 & a^\nu \\
a^\nu & 0
\end{bmatrix} \right\} = \left\{ \begin{bmatrix}
0 & a^\lambda \\
0 & 0
\end{bmatrix} \begin{bmatrix}
a^\mu & 0 \\
0 & a^\mu
\end{bmatrix} \begin{bmatrix}
0 & a^\nu \\
0 & 0
\end{bmatrix} \right\} + \left\{ \begin{bmatrix}
0 & a^\lambda \\
0 & 0
\end{bmatrix} \begin{bmatrix}
0 & 0 \\
0 & a^\mu
\end{bmatrix} \begin{bmatrix}
0 & a^\nu \\
0 & 0
\end{bmatrix} \right\}
\]

\[
+ \left\{ \begin{bmatrix}
0 & a^\nu \\
a^\lambda & 0
\end{bmatrix} \begin{bmatrix}
0 & 0 \\
0 & a^\mu
\end{bmatrix} \begin{bmatrix}
0 & a^\nu \\
0 & 0
\end{bmatrix} \right\} + \left\{ \begin{bmatrix}
0 & 0 \\
a^\lambda & 0
\end{bmatrix} \begin{bmatrix}
0 & 0 \\
0 & a^\mu
\end{bmatrix} \begin{bmatrix}
0 & a^\nu \\
0 & 0
\end{bmatrix} \right\}.
\]
which further expands, using (8)-(11) into

\[
\begin{align*}
&\left\{ \begin{bmatrix} 0 & a^\lambda \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} a^\mu & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & a^\nu \end{bmatrix} \right\} \\
+ &\left\{ \begin{bmatrix} 0 & a^\lambda \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & a^\mu \end{bmatrix}, \begin{bmatrix} a^\nu & 0 \\ 0 & 0 \end{bmatrix} \right\} \\
+ &\left\{ \begin{bmatrix} 0 & 0 \\ a^\lambda & 0 \end{bmatrix}, \begin{bmatrix} a^\mu & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ a^\nu & 0 \end{bmatrix} \right\} \\
+ &\left\{ \begin{bmatrix} 0 & 0 \\ a^\lambda & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ a^\mu & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ a^\nu & 0 \end{bmatrix} \right\} \\
= &\left[ \begin{array}{cc} 0 & 0 \\ \langle a^\nu a^\mu a^\lambda \rangle / 2 & 0 \end{array} \right] + \left[ \begin{array}{cc} \langle a^\lambda a^\mu a^\nu \rangle / 2 & 0 \\ 0 & 0 \end{array} \right] \\
+ &\left[ \begin{array}{cc} 0 & 0 \\ \langle a^\nu a^\mu a^\lambda \rangle / 2 & 0 \end{array} \right] + \left[ \begin{array}{cc} \langle a^\lambda a^\mu a^\nu \rangle / 2 & 0 \\ 0 & 0 \end{array} \right] = a^{\lambda+\mu+\nu}.
\end{align*}
\]

\[\square\]

**Lemma 4.2.** \(D(a, a) = D(\overline{a}, \overline{a}).\)

**Proof.** We shall use (3) with \(z = \{xxy\},\) which states that

\[(17) \quad D(\{xyx\}, \{xxy\}) = 2D(x, \{yx \{xxy\}\}) - D(\{x \{xxy\} x\}, y).\]

We have, by (17) and Lemma 4.1,

\[
D(a, a) = D\left(\left\{ \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}} \right\} \right) \\
= 2D\left(\left\{ \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}} \right\} \right) \\
\quad - D\left(\left\{ \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}} \right\} \right) \\
= 2D\left(\left\{ \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}} \right\} \right) - D\left(\left\{ \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}}, \frac{a^{1/3}}{\overline{a}^{1/3}} \right\} \right) \\
= 2D\left(\overline{a}, \overline{a}\right) - D(a, a),
\]

which proves the lemma. \[\square\]

**Lemma 4.3.** \(D(a, \overline{a}) = D(\overline{a}, a).\)
Proof. By Lemma 4.1 and two applications of (1),
\[
D(a, \bar{a}) = D \left( \left\{ a^{1/4}, \bar{a}^{1/4}, \bar{a}^{1/2} \right\} , \bar{a} \right)
\]
\[
= D(\bar{a}^{1/2}, \left\{ a^{1/4}, \bar{a}^{1/4} \right\}) + [D(a^{1/4}, \bar{a}^{1/4}), D(\bar{a}^{1/2}, \bar{a})]
\]
\[
= D(\bar{a}^{1/2}, a^{3/2}) + [D(a^{1/4}, \bar{a}^{1/4}), D(a^{1/2}, a)]
\]
(by Lemma 4.2 since \( D(\bar{a}^{1/2}, \bar{a}) = D(\bar{a}^{3/4}, \bar{a}) \))
\[
= D(\bar{a}^{1/2}, a^{3/2}) + D \left( \left\{ a^{1/4}A^{1/4}a^{1/2} \right\} , a \right) - D \left( a^{1/2}, \left\{ aa^{1/4}A^{1/4} \right\} \right)
\]
\[
= D(\bar{a}^{1/2}, a^{3/2}) + D(\bar{a}, a) - D(a^{1/2}, \bar{a}^{3/2}).
\]

Hence \( D(a, \bar{a}) - D(\bar{a}, a) = D(\bar{a}^{1/2}, a^{3/2}) - D(a^{1/2}, \bar{a}^{3/2}) \).

It remains to show that \( D(\bar{a}, a^3) - D(a, \bar{a}^3) = 0 \) for every \( a \in A \). Now by (3) and Lemma 4.1,
\[
D(\bar{a}, a^3) = D(\bar{a}, \{a, \bar{a}, \bar{a}\})
\]
\[
= D(\{\bar{a}aa\}, \bar{a})/2 + D(\bar{a}^3, a)/2
\]
\[
= D(\bar{a}^3, \bar{a})/2 + D(\bar{a}^3, a)/2
\]
\[
= D(a, \bar{a}^3) \text{ (by interchanging } a \text{ and } \bar{a}).
\]

This proves the lemma. \( \square \)

By linearization from the preceding two lemmas we obtain:

Lemma 4.4. \( D(a, b) = D(\bar{a}, \bar{b}) ; D(\bar{a}, b) = D(a, \bar{b}) \).

Proof. From \( D(a + b, \bar{a} + \bar{b}) = D(\bar{a} + \bar{b}, a + b) \) follows \( D(b, \bar{a}) + D(a, \bar{b}) = D(\bar{a}, b) + D(\bar{b}, a) \). Now replace \( a \) by \( ia \) and add to obtain \( D(a, \bar{b}) = D(\bar{a}, b) \).

The second statement follows similarly from \( D(a + b, a + b) = D(\bar{a} + \bar{b}, \bar{a} + \bar{b}) \). \( \square \)

Proposition 4.5. If \( A \) is an operator space such that \( M_2(A)_0 \) is a bounded symmetric domain, then \( [abc] = (abc) = (abc) \).

Proof. By expanding as in the second part of the proof of Lemma 4.3,
\[
D(a, b) \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix}
\]
\[
= \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix} \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix}
\]
\[
= \begin{bmatrix} \{abx\} & 0 \\ 0 & 0 \end{bmatrix}
\]
and
\[
D(\bar{a}, \bar{b}) \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} = \left\{ \begin{bmatrix} 0 & \alpha \\ \alpha & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ b & 0 \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right\}
\]
\[
= \left\{ \begin{bmatrix} 0 & \alpha \\ \alpha & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ b & 0 \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right\}
\]
\[
+ \left\{ \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ b & 0 \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right\}
\]
\[
= \begin{bmatrix} \frac{[\alpha]}{2} + \frac{[\beta]}{2} & 0 \\ 0 & 0 \end{bmatrix},
\]
so that \([\beta] = (\beta).

Similarly,
\[
\begin{bmatrix} 0 & \langle \beta \rangle / 2 \\ \langle \alpha \rangle / 2 & 0 \end{bmatrix} = \left\{ \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} 0 & b \\ b & 0 \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right\}
\]
\[
= D(\bar{a}, \bar{b}) \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix}
\]
\[
= D(\bar{a}, \bar{b}) \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix}
\]
\[
= \left\{ \begin{bmatrix} 0 & \alpha \\ \alpha & 0 \end{bmatrix} \begin{bmatrix} 0 & b \\ b & 0 \end{bmatrix} \begin{bmatrix} x & 0 \\ 0 & 0 \end{bmatrix} \right\}
\]
\[
= \begin{bmatrix} 0 & \langle \beta \rangle / 2 \\ \langle \alpha \rangle / 2 & 0 \end{bmatrix},
\]
so that \(\langle \beta \rangle = [\beta].

\section{Main result.}

\begin{proposition}
Let \(X\) be an operator space such that \(M_2(X)\) is a bounded symmetric domain. Then \((X, [\cdot], \| \cdot \|)\) is a C*-ternary ring in the sense of Zettl \cite{Zettl} (see Definition 2.3) and its JB*-triple product (see the beginning of Section 3) satisfies \(\{\alpha\beta\gamma\} = (\alpha\beta\gamma + \gamma\alpha\beta)/2\).
\end{proposition}

\begin{proof}
It was already shown in Lemma 3.8 that \(\{\alpha\beta\gamma\} = (\alpha\beta\gamma + \gamma\alpha\beta)/2\) and that \(\|\alpha\beta\gamma\| = \|\alpha\|^3\) and it is clear that \(\|[\alpha\beta\gamma]\| \leq \|\alpha\\|\|\beta\\|\|\gamma\|\). It remains to show associativity. To prove this we will use Lemma 3.3 and Proposition 4.5. For \(a, b, c, d, e \in X\), let
\[
A = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix}, C = \begin{bmatrix} c & 0 \\ 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 0 \\ d & 0 \end{bmatrix}, E = \begin{bmatrix} 0 & 0 \\ e & 0 \end{bmatrix}.
\]
Then

\[ [[abc]de] = ([abc]de) = 2p_{11} \left( \left\{ \left[ \begin{array}{ccc} abc & 0 \\ 0 & 0 \\ 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 \\ d & 0 \\ 0 & e \end{array} \right] \right\} \right) \quad \text{(by (12))} \\
= 4p_{11} \left( \left\{ \left[ \begin{array}{ccc} a & 0 & b \\ 0 & c & 0 \\ 0 & d & e \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 \\ 0 & 0 \\ d & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 \\ e & 0 \\ 0 & 0 \end{array} \right] \right\} \right) \quad \text{(by (14))} \\
= 4p_{11} \{ED \{CBA\}\} \quad \text{(by commutativity of the triple product)} \\
= 4p_{11}\{CB \{EDA\}\} + 4p_{11}\{EDC \{BA\}\} \quad 4p_{11}\{C \{BED\} A\} \quad \text{(by (2))} \\
= 0 + 4p_{11} \left( \left\{ \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ e & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ d & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ c & 0 & 0 \end{array} \right] \right\} \right) + 0 \\
= 2p_{11} \left( \left\{ \left[ \begin{array}{ccc} 0 & a & 0 \\ 0 & 0 & 0 \\ c & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & b & 0 \\ 0 & 0 & 0 \\ d & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ e & 0 & 0 \end{array} \right] \right\} \right) \quad \text{(by (15))} \\
= [[ab(cde)] = [ab[cde]].

To complete the proof of associativity, consider

\[ [a\{dcb\}e] = \langle a\{dcb\}e \rangle \]
\[ = 2p_{11} \left( \left\{ \left[ \begin{array}{ccc} 0 & a & 0 \\ 0 & 0 & 0 \\ d & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ e & 0 & 0 \end{array} \right] \right\} \right) \quad \text{(by (13))} \\
= 4p_{11} \left( \left\{ \left[ \begin{array}{ccc} 0 & a & 0 \\ 0 & 0 & 0 \\ d & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ d & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ e & 0 & 0 \end{array} \right] \right\} \right) \quad \text{(by (16))} \\
= 4p_{11}\{A \{DCB\} E\} \\
= 4p_{11}\{\{ABC\} DE\} + 4p_{11}\{EB \{ADC\}\} - 4p_{11}\{C \{BAD\} E\} \quad \text{(by (4))} \\
= 4p_{11}\{\{ABC\} DE\} \quad \text{(since } A \perp D) \\
= 4p_{11} \left( \left\{ \left[ \begin{array}{ccc} 0 & a & 0 \\ 0 & 0 & 0 \\ d & 0 & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ e & 0 & 0 \end{array} \right] \right\} \right) \\
= 2p_{11} \left( \left\{ \left[ \begin{array}{ccc} 0 & 0 & 0 \\ abc & 0 & 0 \\ 0 & d & 0 \end{array} \right], \left[ \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right] \right\} \right) \quad \text{(by (15))} \\
= [[abc]de].

\Box
Lemma 5.2. Let $A$ be an operator space such that $M_2(A)_0$ is a bounded symmetric domain, so that by Proposition 5.1, $A$ is a $C^*$-ternary ring. Suppose that the $C^*$-ternary ring $A$ is isomorphic to a TRO, that is, $A_{-1} = 0$ in Theorem 2.2. Form the ternary product $[\cdot \cdot \cdot]_{M_2(A)}$ induced by the ternary product on $A$ as if it was ordinary matrix multiplication, that is, if $X = [x_{ij}], Y = [y_{kl}], Z = [z_{pq}] \in M_2(A)$, then $[XYZ]_{M_2(A)}$ is the matrix whose $(i,j)$-entry is $\sum_{p,q} [x_{ip}y_{qp}z_{qj}]$. Then

$$2 \{XYZ\}_{M_2(A)} = [XYZ]_{M_2(A)} + [ZYX]_{M_2(A)}.$$ 

Proof. It suffices to prove that $\{XXX\}_{M_2(A)} = [XXX]_{M_2(A)}$. In the first place,

$$[XXX]_{M_2(A)} = \begin{bmatrix}
[x_{11}x_{11}x_{11}] + [x_{12}x_{12}x_{11}] & [x_{11}x_{11}x_{12}] + [x_{12}x_{12}x_{12}] \\
+[x_{11}x_{21}x_{21}] + [x_{12}x_{22}x_{21}] & +[x_{11}x_{21}x_{22}] + [x_{12}x_{22}x_{22}]
\end{bmatrix}.$$ 

On the other hand, by using Lemmas 3.3, 3.4, 3.6 and 3.8, and Proposition 3.9,

$$\sum_{k,l,p,q} \{P_{11}(X)P_{kl}(X)P_{pq}(X)\} = \begin{bmatrix}
{x_{11}x_{11}x_{11}} + [x_{12}x_{12}x_{11}] / 2 & [x_{11}x_{11}x_{12}] / 2 \\
+[x_{11}x_{21}x_{21}] / 2 & +[x_{11}x_{21}x_{22}] / 2
\end{bmatrix},$$

$$= \begin{bmatrix}
[x_{21}x_{11}x_{11}] / 2 + [x_{22}x_{12}x_{11}] / 2 & 0
\end{bmatrix}. $$
\[ \sum_{k,l,p,q} \{ P_{12}(X)P_{kl}(X)P_{pq}(X) \} = \begin{bmatrix} [x_{12}x_{22}x_{21}]/2 & \{ x_{12}x_{12}x_{12} \} + [x_{11}x_{11}x_{12}]/2 \\ +[x_{12}x_{12}x_{11}]/2 & +[x_{12}x_{22}x_{22}]/2 \\ 0 & [x_{21}x_{11}x_{12}]/2 + [x_{22}x_{12}x_{12}]/2 \end{bmatrix}, \]

\[ \sum_{k,l,p,q} \{ P_{21}(X)P_{kl}(X)P_{pq}(X) \} = \begin{bmatrix} [x_{11}x_{21}x_{21}]/2 + [x_{12}x_{22}x_{21}]/2 & 0 \\ \{ x_{21}x_{21}x_{21} \} + [x_{21}x_{11}x_{11}]/2 & [x_{21}x_{11}x_{12}]/2 \\ +[x_{22}x_{22}x_{21}]/2 & +[x_{21}x_{21}x_{22}]/2 \end{bmatrix}, \]

and

\[ \sum_{k,l,p,q} \{ P_{22}(X)P_{kl}(X)P_{pq}(X) \} = \begin{bmatrix} 0 & [x_{11}x_{21}x_{22}]/2 + [x_{12}x_{22}x_{22}]/2 \\ [x_{22}x_{12}x_{11}]/2 & \{ x_{22}x_{22}x_{22} \} + [x_{21}x_{21}x_{22}]/2 \\ +[x_{22}x_{22}x_{21}]/2 & [x_{22}x_{12}x_{12}]/2 \end{bmatrix}. \]

Since \( \{XXX\}_{M_2(A)} = \sum_{i,j} \sum_{k,l,p,q} \{ P_{ij}(X)P_{kl}(X)P_{pq}(X) \} \) and \( \{xxx\} = [xxx] \), the lemma follows. \[ \square \]

We now state and prove the main result of this paper.

**Theorem 5.3.** Let \( A \subset B(H) \) be an operator space and suppose that \( M_n(A)_0 \) is a bounded symmetric domain for some \( n \geq 2 \). Then \( A \) is \( n \)-isometric to a ternary ring of operators (TRO). If \( M_n(A)_0 \) is a bounded symmetric domain
for all $n \geq 2$, then $A$ is ternary isomorphic and completely isometric to a TRO.

Proof. The second statement follows from the first one. Suppose $n = 2$. From Theorem 2.2 and Proposition 5.1, we know that $A = A_1 \oplus A_{-1}$ where $A_1$ is ternary isomorphic to a TRO $B$ and $A_{-1}$ is anti-isomorphic to a TRO $C$. Let $\varphi : A_{-1} \to C$ be an anti-isomorphism. Since $C$ is a JB*-triple under the product $\{x \ y \ z\} = (1/2)(xy^*z + zy^*x)$ and $\varphi$ is an isometry, hence a triple isomorphism, it follows that
\[ \varphi(x)\varphi(x)^*\varphi(x) = \varphi\{xxx\} = \varphi[xxx] = -\varphi(x)\varphi(x)^*\varphi(x) \]
so that $\varphi(x)\varphi(x)^*\varphi(x) = 0$ and $x = 0$. Thus $A_{-1} = 0$ and $A$ is ternary isomorphic to a TRO $B$. Let $\psi : A \to B$ be a surjective ternary isomorphism. Then by Lemma 5.2, the amplification $\psi_2$ is a triple isomorphism of the JB*-triple $M_2(A)$ onto the JB*-triple $M_2(B)$, with the triple product
\[ \{RST\}_{M_2(B)} := (RS^*T + TS^*R)/2, \]
implicating that $\psi_2$ is a triple isomorphism, hence an isometry. Thus, $A$ is 2-isometric to $B$, proving the theorem for $n = 2$. The general case for $M_n(A)$ is now not difficult to obtain. We require only one short lemma.

Lemma 5.4. Let $A$ be an operator space such that for some $n \geq 3$, $M_n(A)$ has a JB*-triple structure. Then for $X, Y, Z \in M_n(A)$, the following products all vanish:

- $\{P_{ij}(X) \ P_{kj}(Y) \ P_{lj}(Z)\}$ (for distinct $i, k, l$)
- $\{P_{ij}(X) \ P_{ik}(Y) \ P_{il}(Z)\}$ (for distinct $j, k, l$)
- $\{P_{ij}(X) \ P_{kl}(Y) \ P_{pq}(Z)\}$ (for $i \neq k, j \neq l$ and either $p \notin \{i, k\}$ or $q \notin \{j, l\}$).

Proof. Two applications of the fact that the range of a bicontractive projection on a JB*-triple is a subtriple yield that $\{P_{ij}(X) \ P_{kj}(Y) \ P_{lj}(Z)\}$ lies in $(P_{ij} + P_{kj} + P_{lj})M_n(A)$. However, by a conditional expectation property,

\[ (P_{ij} + P_{kj})\{P_{ij}(X) \ P_{kj}(Y) \ P_{lj}(Z)\} = (P_{ij} + P_{kj})\{P_{ij}(X) \ P_{kj}(Y) \ 0\} = 0. \]

A similar calculation shows $(P_{kj} + P_{lj})\{P_{ij}(X) \ P_{kj}(Y) \ P_{lj}(Z)\} = 0$, proving the first statement. A similar argument proves the second statement. The proof of the last statement is the same as the proof of Lemma 3.3. For $n = 3$, one needs to prove, for example, that
\[ D \left( \begin{bmatrix} a & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) = 0. \]
Returning to the proof of Theorem 5.3, if \( M_n(A) \) is a JB*-triple, then \( M_2(A) \), which is isometric to the range of a contractive projection on \( M_n(A) \), is also a JB*-triple. Hence, by the \( n = 2 \) case, \( A \) is a C*-ternary ring which is ternary isomorphic and isometric under a map \( \phi \) to a TRO \( B \) and \( M_2(A) \) is triple isomorphic and isometric to \( M_2(B) \) under the amplification \( \phi_2 \). Every triple product \( \{XYZ\} \) in \( M_n(A) \) is the sum of products of the form \( \{P_{ij}(X) P_{kl}(Y) P_{pq}(Z)\} \). By Lemma 5.4, every such product of matrix elements in \( M_n(A) \) is either zero or takes place in the intersection of two rows with two columns. The subspace of \( M_n(A) \) defined by one such intersection is a subtriple of \( M_n(A) \) since it is the range of the product of two bicontractive projections. It is isometric, via

\[
P_{ij}(X) + P_{il}(Y) + P_{kj}(Z) + P_{kl}(W) \mapsto \begin{bmatrix} P_{ij}(X) & P_{il}(Y) \\ P_{kj}(Z) & P_{kl}(W) \end{bmatrix},
\]

hence triple isomorphic, to \( M_2(A) \). Hence, by the proof of the \( n = 2 \) case, all triple products in \( M_n(A) \) are the natural ones obtained from the ternary structure on \( A \) as in Lemma 5.2. It follows that \( M_n(A) \) is triple isomorphic to \( M_n(B) \) via the amplification map \( \phi_n \) which is thus an isometry. \( \square \)

As application, we offer two corollaries.

**Corollary 5.5.** Let \( A \subset \mathcal{B}(H,K) \) be a TRO and let \( P \) be a completely contractive projection on \( A \). Then the range of \( P \) is completely isometric to another TRO.

*Proof.* Since \( A \) is a TRO, \( M_n(A) \) is a JB*-triple. Therefore \( M_n(P(A)) = P_n(M_n(A)) \) is a JB*-triple, and its unit ball is a bounded symmetric domain. \( \square \)

Another way to obtain this corollary is to note that every TRO is a corner of a C*-algebra and hence the range of a completely contractive projection on that algebra. By composing these two projections, the corollary is reduced to [38].

Our second corollary is a variant of the fundamental Choi-Effros result.

**Corollary 5.6.** Let \( P \) be a unital 2-positive projection on a unital C*-algebra \( A \). Then \( P(A) \) is 2-isometric to a C*-algebra. If \( P \) is completely positive and unital, then \( P(A) \) is completely isometric to a C*-algebra.

In order to state our second theorem, we recall that a complex Banach space \( A \) is linearly isometric to a unital JB*-algebra if and only if its open unit ball \( A_0 \) is a bounded symmetric domain of tube type [11]. In [37], a necessary and sufficient condition, involving the Lie algebra of all complete holomorphic vector fields on \( A_0 \), is given for such \( A \) to be obtained from a C*-algebra with the anticommutator product. Our next theorem gives a holomorphic characterization of C*-algebras up to complete isometry.
Theorem 5.7. Let $A \subset B(H)$ be an operator space and suppose that $M_n(A)_0$ is a bounded symmetric domain for some $n \geq 2$. If the induced bounded symmetric domain structure on $A_0$ is of tube type, then $A$ is $n$-isometric to a $C^*$-algebra. If $M_n(A)_0$ is a bounded symmetric domain for all $n \geq 2$ and $A_0$ is of tube type, then $A$ is completely isometric to a $C^*$-algebra.

Proof. By Theorem 5.3, we may assume that $A$ is a TRO. Since $A$ has the structure of a unital JB*-algebra, there is a partial isometry $u \in A$ such that $au^*u = uu^*a = a$ for every $a \in A$. Then $A$ becomes a C*-algebra with product $a \cdot b = au^*b$ and involution $a^* = ua^*u$. Since $ab^*c = a \cdot b^* \cdot c$, and ternary isomorphisms of TRO’s are complete isometries, the result follows. □

Remark 5.8. One can construct operator spaces that are 2-isometric to a C*-algebra $A$ which are not completely isometric to $A$. Hence, if $M_2(A)_0$ is a bounded symmetric domain it does not follow that $M_n(A)_0$ is a bounded symmetric domain for every $n \geq 2$. It would be interesting to see if this were true under some further condition on $A$. The proof of Theorem 5.3 seems to require a bounded symmetric domain structure on $M_2(A)_0$, not simply on $M_{1,2}(A)_0$ for example. It would be interesting to see what could be said if it is assumed that $M_{1,n}(A)_0$ were a bounded symmetric domain for every $n \geq 2$.

Acknowledgements. The authors wish to thank Professors Zhong-Jin Ruan and David Blecher for their advice and encouragement at the beginning stages of this work.

References


Received November 22, 2001 and revised April 25, 2002. Both authors acknowledge the support of NSF Grant DMS-0101153.

Department of Mathematics
Denison University
Granville, OH 43023
E-mail address: nealm@denison.edu

Department of Mathematics
University of California
Irvine, CA 92697-3875
E-mail address: brusso@math.uci.edu
ON THE VERE–JONES CLASSIFICATION AND EXISTENCE OF MAXIMAL MEASURES FOR COUNTABLE TOPOLOGICAL MARKOV CHAINS

SYLVIE RUETTE

We consider topological Markov chains (also called Markov shifts) on countable graphs. We show that a transient graph can be extended to a recurrent graph of equal entropy which is either positive recurrent of null recurrent, and we give an example of each type. We extend the notion of local entropy to topological Markov chains and prove that a transitive Markov chain admits a measure of maximal entropy (or maximal measure) whenever its local entropy is less than its (global) entropy.

Introduction.

In this article we are interested in connected oriented graphs and topological Markov chains. All the graphs we consider have a countable set of vertices. If $G$ is an oriented graph, let $\Gamma_G$ be the set of two-sided infinite sequences of vertices that form a path in $G$ and let $\sigma$ denote the shift transformation. The Markov chain associated to $G$ is the (noncompact) dynamical system $(\Gamma_G, \sigma)$. The entropy $h(G)$ of the Markov chain $\Gamma_G$ was defined by Gurevich; it can be computed by several ways and satisfies the Variational Principle [7] and [8].

In [16] Vere-Jones classifies connected oriented graphs as transient, null recurrent or positive recurrent according to the properties of the series associated with the number of loops, by analogy with probabilistic Markov chains. To a certain extent, positive recurrent graphs resemble finite graphs. In [7] Gurevich shows that a Markov chain on a connected graph admits a measure of maximal entropy (also called maximal measure) if and only if the graph is positive recurrent. In this case, this measure is unique and it is an ergodic Markov measure.

In [13] and [14] Salama gives a geometric approach to the Vere-Jones classification. The fact that a graph can (or cannot) be “extended” or “contracted” without changing its entropy is closely related to its class. In particular a graph with no proper subgraph of equal entropy is positive recurrent. The converse is not true [14] (see also [6] for an example of a positive recurrent graph with a finite valency at every vertex that has
no proper subgraph of equal entropy). This result shows that the positive recurrent class splits into two subclasses: A graph is called strongly positive recurrent if it has no proper subgraph of equal entropy; it is equivalent to a combinatorial condition (a finite connected graph is always strongly positive recurrent). In [13] and [14] Salama also states that a graph is transient if and only if it can be extended to a bigger transient graph of equal entropy. We show that any transient graph $G$ is contained in a recurrent graph of equal entropy, which is positive or null recurrent depending on the properties of $G$. We illustrate the two possibilities — a transient graph with a positive or null recurrent extension — by an example.

The result of Gurevich entirely solves the question of existence of a maximal measure in term of graph classification. Nevertheless it is not so easy to prove that a graph is positive recurrent and one may wish to have more efficient criteria. In [10] Gurevich and Zargaryan give a sufficient condition for existence of a maximal measure; it is formulated in terms of exponential growth of the number of paths inside and outside a finite subgraph. We give a new sufficient criterion based on local entropy.

Why consider local entropy? For a compact dynamical system, it is known that a null local entropy implies the existence of a maximal measure ([11], see also [1] for a similar but different result). This result may be strengthened in some cases: It is conjectured that, if $f$ is a map of the interval which is $C^r$, $r > 1$, and satisfies $h_{\text{top}}(f) > h_{\text{loc}}(f)$, then there exists a maximal measure [2]. Our initial motivation comes from the conjecture above because smooth interval maps and Markov chains are closely related. If $f: [0, 1] \to [0, 1]$ is $C^{1+\alpha}$ (i.e., $f$ is $C^1$ and $f'$ is $\alpha$-Hölder with $\alpha > 0$) with $h_{\text{top}}(f) > 0$ then an oriented graph $G$ can be associated to $f$, $G$ is connected if $f$ is transitive, and there is a bijection between the maximal measures of $f$ and those of $\Gamma_G$ [2] and [3]. We show that a Markov chain is strongly positive recurrent, thus admits a maximal measure, if its local entropy is strictly less than its Gurevich entropy. However this result does not apply directly to interval maps since the “isomorphism” between $f$ and its Markov extension is not continuous so it may not preserve local entropy (which depends on the distance).

The article is organized as follows. Section 1 contains definitions and basic properties on oriented graphs and Markov chains. In Section 2, after recalling the definitions of transient, null recurrent and positive recurrent graphs and some related properties, we show that any transient graph is contained in a recurrent graph of equal entropy (Proposition 2.8) and we give an example of a transient graph which extends to a positive recurrent (resp. null recurrent) graph. Section 3 is devoted to the problem of existence of maximal measures: Theorem 3.8 gives a sufficient condition for the existence of a maximal measure, based on local entropy.
1. Background.

1.1. Graphs and paths. Let $G$ be an oriented graph with a countable set of vertices $V(G)$. If $u, v$ are two vertices, there is at most one arrow $u \to v$. A path of length $n$ is a sequence of vertices $(u_0, \ldots, u_n)$ such that $u_i \to u_{i+1}$ in $G$ for $0 \leq i < n$. This path is called a loop if $u_0 = u_n$. We say that the graph $G$ is connected if for all vertices $u, v$ there exists a path from $u$ to $v$; in the literature, such a graph is also called strongly connected.

If $H$ is a subgraph of $G$, we write $H \subset G$; if in addition $H \neq G$, we write $H \nsubseteq G$ and say that $H$ is a proper subgraph. If $W$ is a subset of $V(G)$, the set $V(G) \setminus W$ is denoted by $\overline{W}$. We also denote by $W$ the subgraph of $G$ whose vertices are $W$ and whose edges are all edges of $G$ between two vertices in $W$.

Let $u, v$ be two vertices. We define the following quantities:

- $p_{uv}^G(n)$ is the number of paths $(u_0, \ldots, u_n)$ such that $u_0 = u$ and $u_n = v$.
- $R_{uv}(G)$ is the radius of convergence of the series $\sum p_{uv}^G(n)z^n$.
- $f_{uv}^G(n)$ is the number of paths $(u_0, \ldots, u_n)$ such that $u_0 = u$, $u_n = v$, and $u_i \neq v$ for $0 < i < n$.
- $L_{uv}(G)$ is the radius of convergence of the series $\sum f_{uv}^G(n)z^n$.

Proposition 1.1 (Vere-Jones [16]). Let $G$ be an oriented graph. If $G$ is connected, $R_{uv}(G)$ does not depend on $u$ and $v$; it is denoted by $R(G)$.

If there is no confusion, $R(G)$ and $L_{uv}(G)$ will be written $R$ and $L_{uv}$. For a graph $G'$ these two radii will be written $R'$ and $L'_{uv}$.

1.2. Markov chains. Let $G$ be an oriented graph. $\Gamma_G$ is the set of two-sided infinite paths in $G$, that is,

$$\Gamma_G = \{(v_n)_{n \in \mathbb{Z}} \mid \forall n \in \mathbb{Z}, v_n \to v_{n+1} \text{ in } G\} \subset (V(G))^{\mathbb{Z}}.$$ 

$\sigma$ is the shift on $\Gamma_G$. The (topological) Markov chain on the graph $G$ is the system $\Gamma_G, \sigma$.

The set $V(G)$ is endowed with the discrete topology and $\Gamma_G$ is endowed with the induced topology of $(V(G))^{\mathbb{Z}}$. The space $\Gamma_G$ is not compact unless $G$ is finite. A compatible distance on $\Gamma_G$ is given by $d$, defined as follows:

$V(G)$ is identified with $\mathbb{N}$ and the distance $D$ on $V(G)$ is given by $D(n, m) = \frac{1}{2^n} - \frac{1}{2^m}$. If $\bar{u} = (u_n)_{n \in \mathbb{Z}}$ and $\bar{v} = (v_n)_{n \in \mathbb{Z}}$ are two elements of $\Gamma_G$,

$$d(\bar{u}, \bar{v}) = \sum_{n \in \mathbb{Z}} \frac{D(u_n, v_n)}{2^{|n|}} \leq 3.$$ 

The Markov chain $(\Gamma_G, \sigma)$ is transitive if for any nonempty open sets $A, B \subset \Gamma_G$ there exists $n > 0$ such that $\sigma^n(A) \cap B \neq \emptyset$. Equivalently, $\Gamma_G$ is transitive if and only if the graph $G$ is connected. In the sequel we will be interested in connected graphs only.
1.3. Entropy. If $G$ is a finite graph, $\Gamma_G$ is compact and the topological entropy $h_{\text{top}}(\Gamma_G, \sigma)$ is well-defined (see e.g., [5] for the definition of the topological entropy). If $G$ is a countable graph, the Gurevich entropy [7] of $G$ is given by

$$h(G) = \sup \{ h_{\text{top}}(\Gamma_H, \sigma) \mid H \subset G, H \text{ finite} \}.$$ 

This entropy can also be computed in a combinatorial way, as the exponential growth of the number of paths with fixed endpoints [8].

**Proposition 1.2** (Gurevich). Let $G$ be a connected oriented graph. Then for any vertices $u, v$

$$h(G) = \lim_{n \to +\infty} \frac{1}{n} \log p^G_{uv}(n) = -\log R(G).$$

Another way to compute the entropy is to compactify the space $\Gamma_G$ and then use the definition of topological entropy for compact metric spaces. If $G$ is an oriented graph, denote the one-point compactification of $V(G)$ by $V(G) \cup \{\infty\}$ and define $\Gamma_G$ as the closure of $\Gamma_G$ in $(V(G) \cup \{\infty\})^2$. The distance $d$ naturally extends to $\Gamma_G$. In [7] Gurevich shows that this gives the same entropy; this means that there is only very little dynamics added in this compactification. Moreover, the Variational Principle is still valid for Markov chains [7].

**Theorem 1.3** (Gurevich). Let $G$ be an oriented graph. Then

$$h(G) = h_{\text{top}}(\Gamma_G, \sigma) = \sup \{ h_{\mu}(\Gamma_G) \mid \mu \text{ } \sigma\text{-invariant probability measure} \}.$$ 

2. On the classification of connected graphs.

2.1. Transient, null recurrent, positive recurrent graphs. In [16] Vere-Jones gives a classification of connected graphs as transient, null recurrent or positive recurrent. The definitions are given in Table 1 (lines 1 and 2) as well as properties of the series $\sum p^G_{uu}(n) z^n$ which give an alternative definition.

In [13] and [14] Salama studies the links between the classification and the possibility to extend or contract a graph without changing its entropy. It follows that a connected graph is transient if and only if it is strictly included in a connected graph of equal entropy, and that a graph with no proper subgraph of equal entropy is positive recurrent.

**Remark 2.1.** In [13] Salama claims that $L_{uu}$ is independent of $u$, which is not true; in [14] he uses the quantity $L = \inf u L_{uu}$ and he states that if $R = L$ then $R = L_{uu}$ for all vertices $u$, which is wrong too (see Proposition 3.2 in [9]). It follows that in [13] and [14] the statement “$R=L$” must be interpreted either as “$R = L_{uu}$ for some $u$” or “$R = L_{uu}$ for all $u$” depending on the context. This encouraged us to give the proofs of Salama’s results in this article.
transient
null
positive

\[ \sum_{n>0} f_{uu}^G(n)R^n < 1 \quad \sum_{n>0} nf_{uu}^G(n)R^n \leq +\infty \quad \sum_{n>0} p_{uu}^G(n)R^n < +\infty \]

\[ \lim_{n \to +\infty} p_{uu}^G(n)R^n = 0 \quad \lambda_{uv} > 0 \]

\[ R = L_{uu} \quad R = L_{uu} \quad R \leq L_{uu} \]

**Table 1.** Properties of the series associated to a transient, null recurrent or positive recurrent graph \( G \); these properties do not depend on the vertices \( u,v \) (\( G \) is connected).

In [14] Salama shows that a transient or null recurrent graph satisfies \( R = L_{uu} \) for all vertices \( u \); we give the unpublished proof due to U. Fiebig [6].

**Proposition 2.2** (Salama). Let \( G \) be a connected oriented graph. If \( G \) is transient or null recurrent then \( R = L_{uu} \) for all vertices \( u \). Equivalently, if there exists a vertex \( u \) such that \( R < L_{uu} \) then \( G \) is positive recurrent.

**Proof.** For a connected oriented graph, it is obvious that \( R \leq L_{uu} \) for all \( u \), thus the two claims of the Proposition are equivalent. We prove the second one.

Let \( u \) be a vertex of \( G \) such that \( R < L_{uu} \). Let \( F(x) = \sum_{n \geq 1} f_{uu}^G(n)x^n \) for all \( x \geq 0 \). If we break a loop based in \( u \) into first return loops, we get the following formula:

\[ \sum_{n \geq 0} p_{uu}^G(n)x^n = \sum_{k \geq 0} (F(x))^k. \]

Suppose that \( G \) is transient, that is, \( F(R) < 1 \). The map \( F \) is analytic on \([0, L_{uu}]\) and \( R < L_{uu} \) thus there exists \( R < x < L_{uu} \) such that \( F(x) < 1 \). According to Equation (1) one gets that \( \sum_{n \geq 0} p_{uu}^G(n)x^n < +\infty \), which contradicts the definition of \( R \). Therefore \( G \) is recurrent. Moreover \( R < L_{uu} \) by assumption, thus \( \sum_{n \geq 1} n f_{uu}^G(n)R^n < +\infty \), which implies that \( G \) is positive recurrent. \( \square \)

**Definition 2.3.** A connected oriented graph is called **strongly positive recurrent** if \( R < L_{uu} \) for all vertices \( u \).

**Lemma 2.4.** Let \( G \) be a connected oriented graph and \( u \) a vertex.

i) \( R < L_{uu} \) if and only if \( \sum_{n \geq 1} f_{uu}^G(n)L_{uu}^n > 1 \).
ii) If $G$ is recurrent then $R$ is the unique positive number $x$ such that
\[ \sum_{n \geq 1} f^G_{uu}(n)x^n = 1. \]

Proof. Use the fact that $F(x) = \sum_{n \geq 1} f^G_{uu}(n)x^n$ is increasing. \qed

The following result deals with transient graphs [13]:

**Theorem 2.5** (Salama). Let $G$ be a connected oriented graph of finite positive entropy. Then $G$ is transient if and only if there exists a connected oriented graph $G' \supseteq G$ such that $h(G') = h(G)$. If $G$ is transient then $G'$ can be chosen transient.

Proof. The assumption on the entropy implies that $0 < R < 1$. Suppose first that there exists a connected graph $G' \supseteq G$ such that $h(G') = h(G)$, that is, $R' = R$. Fix a vertex $u$ in $G$. The graph $G$ is a proper subgraph of $G'$ thus there exists $n$ such that $f^G_{uu}(n) < f^{G'}_{uu}(n)$, which implies that
\[ \sum_{n \geq 1} f^G_{uu}(n) R^n < \sum_{n \geq 1} f^{G'}_{uu}(n) R^n \leq 1. \]

Therefore $G$ is transient.

Now suppose that $G$ is transient and fix a vertex $u$ in $G$. One has $\sum_{n \geq 1} f^G_{uu}(n) R^n < 1$. Let $k \geq 2$ be an integer such that
\[ \sum_{n \geq 1} f^G_{uu}(n) R^n + R^k < 1. \]

Define the graph $G'$ by adding a loop of length $k$ based at the vertex $u$; one has $R' \leq R$ and
\[ (2) \quad \sum_{n \geq 1} f^{G'}_{uu}(n) R'^n \leq \sum_{n \geq 1} f^G_{uu}(n) R^n = \sum_{n \geq 1} f^G_{uu}(n) R^n + R^k < 1. \]

Equation (2) implies that $R \leq L'_{uu}$ and also that the graph $G'$ is transient, so $R' = L'_{uu}$ by Proposition 2.2. Then one has $L'_{uu} = R' \leq R \leq L'_{uu}$ thus $R = R'$. \qed

In [14] Salama proves that if $R = L_{uu}$ for all vertices $u$ then there exists a proper subgraph of equal entropy. We show that the same conclusion holds if one supposes that $R = L_{uu}$ for some $u$. The proof below is a variant of the one of Salama. The converse is also true, as shown by U. Fiebig [6].

**Proposition 2.6.** Let $G$ be a connected oriented graph of positive entropy.

i) If there is a vertex $u$ such that $R = L_{uu}$ then there exists a connected subgraph $G' \subseteq G$ such that $h(G') = h(G)$.

ii) If there is a vertex $u$ such that $R < L_{uu}$ then for all proper subgraphs $G'$ one has $h(G') < h(G)$.
Proof. i) Suppose that \( R = L_{uu} \). If \( u_0 = u \) is followed by a unique vertex, let \( u_1 \) be this vertex. If \( u_1 \) is followed by a unique vertex, let \( u_2 \) be this vertex, and so on. If this leads to define \( u_n \) for all \( n \) then \( h(G) = 0 \), which is not allowed.

Let \( u_k \) be the last built vertex; there exist two distinct vertices \( v, v' \) such that \( u_k \to v \) and \( u_k \to v' \). Let \( G'_k \) be the graph \( G \) deprived of the arrow \( u_k \to v \) and \( G'_2 \) the graph \( G \) deprived of all the arrows \( u_k \to w, w \neq v \).

Call \( G_i \) the connected component of \( G'_i \) that contains \( u \) (\( i = 1, 2 \)); obviously \( G_i \not\subseteq G \). For all \( n \geq 1 \) one has

\[
   f_{uu}^G(n) = f_{uu}^{G_1}(n - k) = f_{u_k u}^{G_1}(n - k) + f_{u_k u}^{G_2}(n - k),
\]

thus there exists \( i \in \{1, 2\} \) such that \( L_{uu} = L_{u_k u}(G_i) \).

One has

\[
   R \leq R(G_i) \leq L_{u_k u}(G_i) = L_{uu} = R,
\]

thus \( R = R(G_i) \), that is, \( h(G) = h(G_i) \).

ii) Suppose that \( R < L_{uu} \) and consider \( G' \not\subseteq G \). Suppose first that \( u \) is a vertex of \( G' \). The graph \( G \) is positive recurrent by Proposition 2.2 so \( \sum_{n \geq 1} f_{uu}^G(n) R^n = 1 \). Since \( G' \not\subseteq G \) there exists \( n \) such that \( f_{uu}^{G'}(n) < f_{uu}^G(n) \), thus

\[
   \sum_{n \geq 1} f_{uu}^{G'} R^n < 1.
\]

Moreover \( L'_{uu} \geq L_{uu} \). If \( G' \) is transient then \( R' = L'_{uu} \) (Proposition 2.2) thus \( R' \geq L_{uu} > R \). If \( G' \) is recurrent then \( \sum_{n \geq 1} f_{uu}^{G'} R^n = 1 \) thus \( R' > R \) because of Equation (3). In both cases \( R' > R \), that is, \( h(G') < h(G) \).

Suppose now that \( u \) is not a vertex of \( G' \) and fix a vertex \( v \) in \( G' \). Let \((u_0, \ldots , u_p)\) a path (in \( G \)) of minimal length between \( u = u_0 \) and \( v = u_p \), and let \((v_0, \ldots , v_q)\) be a path of minimal length between \( v = v_0 \) and \( u = v_q \).

If \((u_0 = v, w_1, \ldots , w_n = v)\) is a loop in \( G' \) then

\[
   (u_0 = u, u_1, \ldots , u_p = w_0, w_1, \ldots , w_n = v_0, v_1, \ldots , v_q = u)
\]

is a first return loop based in \( u \) in the graph \( G \). For all \( n \geq 0 \) we get that

\[
   p_{uv}(n) \leq f_{uu}^G(n + p + q),
\]

thus \( R' \geq L_{uu} > R \), that is, \( h(G') < h(G) \). \( \square \)

The following result gives a characterization of strongly positive recurrent graphs. It is a straightforward corollary of Proposition 2.6 (see also [6]).

**Theorem 2.7.** Let \( G \) be a connected oriented graph of positive entropy. The following properties are equivalent:

i) For all \( u \) one has \( R < L_{uu} \) (that is, \( G \) is strongly positive recurrent),

ii) there exists \( u \) such that \( R < L_{uu} \),

iii) \( G \) has no proper subgraph of equal entropy.
2.2. Recurrent extensions of equal entropy of transient graphs.

We show that any transient graph $G$ can be extended to a recurrent graph without changing the entropy by adding a (possibly infinite) number of loops. If the series $\sum_{n>0} n f_{vu}^G(n) R^n$ is finite then the obtained recurrent graph is positive recurrent (but not strongly positive recurrent), otherwise it is null recurrent.

**Proposition 2.8.** Let $G$ be a transient graph of finite positive entropy. Then there exists a recurrent graph $G' \supset G$ such that $h(G) = h(G')$. Moreover $G'$ can be chosen to be positive recurrent if $\sum_{n>0} n f_{uu}^G(n) R^n < +\infty$ for some vertex $u$ of $G$, and $G'$ is necessarily null recurrent otherwise.

**Proof.** The entropy of $G$ is finite and positive thus $0 < R < 1$ and there exists an integer $p$ such that $\frac{1}{2} \leq pR < 1$. Define $\alpha = pR$. Let $u$ be a vertex of $G$ and define $D = 1 - \sum_{n\geq 1} f_{uu}^G(n) R^n$; one has $0 < D < 1$. Moreover

$$\sum_{n\geq 1} \alpha^n \geq \frac{1}{2^n} = 1,$$

thus

$$\sum_{n\geq k+1} \alpha^n = \alpha^k \sum_{n\geq 1} \alpha^n \geq \alpha^k.$$  \quad (4)

We build a sequence of integers $(n_i)_{i\in I}$ such that $2 \sum_{i\in I} \alpha^{n_i} = D$. For this, we define inductively a strictly increasing (finite or infinite) sequence of integers $(n_i)_{i\in I}$ such that for all $k \in I$

$$\sum_{i=0}^k \alpha^{n_i} \leq \frac{D}{2} < \sum_{i=0}^k \alpha^{n_i} + \sum_{n>n_k} \alpha^n.$$

— Let $n_0$ be the greatest integer $n \geq 2$ such that $\sum_{k\geq n} \alpha^k > \frac{D}{2}$. By choice of $n_0$ one has $\sum_{n\geq n_0+1} \alpha^n \leq \frac{D}{2}$, thus $\alpha^{n_0} \leq \frac{D}{2}$ by Equation (4). This is the required property at rank 0.

— Suppose that $(n_0, \ldots, n_k)$ is already defined. If $\sum_{i=0}^k \alpha^{n_i} = \frac{D}{2}$ then $I = \{0, \ldots, k\}$ and we stop the construction. Otherwise let $n_{k+1}$ be the greatest integer $n > n_k$ such that

$$\sum_{i=0}^k \alpha^{n_i} + \sum_{j \geq n} \alpha^{n_{k+1}} > \frac{D}{2}.$$

By choice of $n_{k+1}$ and Equation (4), one has

$$\alpha^{n_{k+1}} \leq \sum_{j \geq n_{k+1}+1} \alpha^j \leq \frac{D}{2} - \sum_{i=0}^k \alpha^{n_i}.$$

This is the required property at rank $k + 1$. 

Define a new graph \( G' \supset G \) by adding \( 2p^{n_i} \) loops of length \( n_i \) based at the vertex \( u \). Obviously one has \( R' \leq R \), and \( \sum_{i \in I} (pR)^{n_i} = \frac{D}{2} \) by construction. Therefore

\[
\sum_{n \geq 1} f^{G'}_{uu}(n)R^n = \sum_{n \geq 1} f^{G}_{uu}(n)R^n + \sum_{i \in I} 2(pR)^{n_i} = 1.
\]

This implies that \( R \leq L'_{uu} \). If \( G' \) is transient then \( \sum_{n \geq 1} f^{G'}_{uu}(n)R^n < 1 \) and \( R' = L'_{uu} \) by Proposition 2.2, thus \( R \leq R' \) and Equation (5) leads to a contradiction. Therefore \( G' \) is recurrent. By Lemma 2.4(ii) one has \( R' = R \), that is, \( h(G') = h(G) \). In addition,

\[
\sum_{n \geq 1} nf^{G'}_{uu}(n)R^n = \sum_{n \geq 1} nf^{G}_{uu}(n)R^n + \sum_{i \in I} n_i \alpha^{n_i}
\]

and this quantity is finite if and only if \( \sum nf^{G}_{uu}(n)R^n \) is finite. In this case the graph \( G' \) is positive recurrent.

If \( \sum nf^{G}_{uu}(n)R^n = +\infty \), let \( H \) be a recurrent graph containing \( G \) with \( h(H) = h(G) \). Then \( H \) is null recurrent because

\[
\sum_{n \geq 1} nf^{H}_{uu}(n)R^n \geq \sum_{n \geq 1} nf^{G}_{uu}(n)R^n = +\infty.
\]

\[\square\]

**Example 2.9.** We build a positive (resp. null) recurrent graph \( G \) such that \( \sum f^{G}_{uu}(n)L^n_{uu} = 1 \) and then we delete an arrow to obtain a graph \( G' \subset G \) which is transient and such that \( h(G') = h(G) \). First we give a description of \( G \) depending on a sequence of integers \( a(n) \) then we give two different values to the sequence \( a(n) \) so as to obtain a positive recurrent graph in one case and a null recurrent graph in the other case.

Let \( u \) be a vertex and \( a(n) \) a sequence of nonnegative integers for \( n \geq 1 \), with \( a(1) = 1 \). The graph \( G \) is composed of \( a(n) \) loops of length \( n \) based at the vertex \( u \) for all \( n \geq 1 \) (see Figure 1). More precisely, define the set of vertices of \( G \) as

\[ V = \{u\} \cup \bigcup_{n=1}^{+\infty} \{v^{n,i}_k | 1 \leq i \leq a(n), 1 \leq k \leq n - 1\}, \]

where the vertices \( v^{n,i}_k \) above are distinct. Let \( v^{n,i}_0 = v^{n,i}_n = u \) for \( 1 \leq i \leq a(n) \). There is an arrow \( v^{n,i}_k \rightarrow v^{n,i}_{k+1} \) for \( 0 \leq k \leq n - 1, 1 \leq i \leq a(n), n \geq 1 \) and there is no other arrow in \( G \). The graph \( G \) is connected and \( f^{G}_{uu}(n) = a(n) \) for \( n \geq 1 \).

The sequence \( (a(n))_{n \geq 2} \) is chosen such that it satisfies

\[
\sum_{n \geq 1} a(n)L^n = 1,
\]

(6)
Figure 1. The graphs $G$ and $G'$; the bold loop (on the left) is the only arrow that belongs to $G$ and not to $G'$, otherwise the two graphs coincide.

where $L = \nu > 0$ is the radius of convergence of the series $\sum a(n)z^n$. If $G$ is transient then $R = \nu$ by Proposition 2.2, but Equation (6) contradicts the definition of transient. Thus $G$ is recurrent. Moreover, $R = \nu$ by Lemma 2.4(ii).

The graph $G'$ is obtained from $G$ by deleting the arrow $u \to u$. Obviously one has $\nu' = \nu$ and

$$\sum_{n \geq 1} f_{uu}^G(n)L^n = 1 - \nu < 1.$$  

This implies that $G'$ is transient because $R' \leq \nu$. Moreover $R' = \nu$ by Proposition 2.2 thus $R' = R$, that is, $h(G') = h(G)$.

Now we consider two different sequences $a(n)$.

1) Let $a(n^2) = 2^{n^2-n}$ for $n \geq 1$ and $a(n) = 0$ otherwise. Then $L = \frac{1}{2}$ and

$$\sum_{n \geq 1} f_{uu}^G(n)L^n = \sum_{n \geq 1} 2^{n^2-n} \frac{1}{2^n} = \sum_{n \geq 1} \frac{1}{2^n} = 1.$$  

Moreover

$$\sum_{n \geq 1} nf_{uu}^G(n)L^n = \sum_{n \geq 1} \frac{n^2}{2^n} < +\infty,$$

hence the graph $G$ is positive recurrent.

2) Let $a(1) = 1$, $a(2^n) = 2^{2^n-n}$ for $n \geq 2$ and $a(n) = 0$ otherwise. One can compute that $L = \frac{1}{2}$, and

$$\sum_{n \geq 1} f_{uu}^G(n)L^n = \frac{1}{2} + \sum_{n \geq 2} 2^{2^n-n} \frac{1}{2^n} = \frac{1}{2} + \sum_{n \geq 2} \frac{1}{2^n} = 1.$$  

Moreover

$$\sum_{n \geq 1} nf_{uu}^G(n)L^n = \frac{1}{2} + \sum_{n \geq 2} 2^n \frac{1}{2^n} = +\infty.$$
hence the graph $G$ is null recurrent.

**Remark 2.10.** Let $G$ be a transient graph of finite entropy. Fix a vertex $u$ and choose an integer $k$ such that $\sum_{n \geq k} R^n < 1 - \sum_{n \geq 1} f_{uu}^G(n) R^n$. For every integer $n \geq k$ let $m_n = \lfloor R^{-n} \rfloor$, add $\lfloor R - (m_n - n) \rfloor$ loops of length $m_n$ based at the vertex $u$ and call $G'$ the graph obtained in this way. It can be shown that the graph $G'$ is transient, $h(G') = h(G)$ and $\sum_{n \geq 1} n f_{uu}^G R^n = +\infty$. Then Proposition 2.8 implies that every transient graph is included in a null recurrent graph of equal entropy.

**Remark 2.11.** In the more general setting of thermodynamic formalism for countable Markov chains, Sarig puts to the fore a subclass of positive recurrent potentials which he calls strongly positive recurrent [15]; his motivation is different, but the classifications agree. If $G$ is a countable oriented graph, a potential is a continuous map $\phi : \Gamma_G \rightarrow \mathbb{R}$ and the pressure $P(\phi)$ is the analogous of the Gurevich entropy, the paths being weighted by $e^{\phi}$; a potential is either transient or null recurrent or positive recurrent. Considering the null potential $\phi \equiv 0$, we retrieve the case of (non-weighted) topological Markov chains. In [15] Sarig introduces a quantity $\Delta_u[\phi]$; $\phi$ is transient (resp. recurrent) if $\Delta_u[\phi] < 0$ (resp. $\Delta_u[\phi] \geq 0$). The potential is called strongly positive recurrent if $\Delta_u[\phi] > 0$, which implies it is positive recurrent. A strongly positive recurrent potential $\phi$ is stable under perturbation, that is, any potential $\phi + t\psi$ close to $\phi$ is positive recurrent too. For the null potential, $\Delta_u[0] = \log \left( \sum_{n \geq 1} f_{uu}^G(n) L^n \right)$, thus $\Delta_u[0] > 0$ if and only if the graph is strongly positive recurrent (Lemma 2.4 and Theorem 2.7). In [9] strongly positive recurrent potentials are called stable positive.

Examples of (non-null) potentials which are positive recurrent but not strongly positive recurrent can be found in [15]; some of them resemble much the Markov chains of Example 2.9, their graphs being composed of loops as in Figure 1.

3. Existence of a maximal measure.

### 3.1. Positive recurrence and maximal measures.

A Markov chain on a finite graph always has a maximal measure [12], but it is not the case for infinite graphs [7]. In [8] Gurevich gives a necessary and sufficient condition for the existence of such a measure.

**Theorem 3.1** (Gurevich). Let $G$ be a connected oriented graph of finite positive entropy. Then the Markov chain $(\Gamma_G, \sigma)$ admits a maximal measure if and only if the graph is positive recurrent. Moreover, such a measure is unique if it exists, and it is an ergodic Markov measure.

In [10] Gurevich and Zargaryan show that if one can find a finite connected subgraph $H \subset G$ such that there are more paths inside than outside...
Let $G$ be a connected oriented graph, $W$ a subset of vertices and $u, v$ two vertices of $G$. Define $t_{uv}^W(n)$ as the number of paths $(v_0, \ldots, v_n)$ such that $v_0 = u, v_n = v$ and $v_i \in W$ for all $0 < i < n$, and put $\tau_{uv}^W = \limsup_{n \to +\infty} \frac{1}{n} \log t_{uv}^W(n)$.

**Theorem 3.2** (Gurevich-Zargaryan). Let $G$ be a connected oriented graph of finite positive entropy. If there exists a finite set of vertices $W$ such that $W$ is connected and for all vertices $u, v$ in $W, \tau_{uv}^W \leq h(W)$, then the graph $G$ is strongly positive recurrent.

For graphs that are not strongly positive recurrent the entropy is mainly concentrated near infinity in the sense that it is supported by the infinite paths that spend most of the time outside a finite subgraph (Proposition 3.3). This result is obtained by applying inductively the construction of Proposition 2.6(i). As a corollary, there exist “almost maximal measures escaping to infinity” (Corollary 3.4). These two results are proven and used as tools to study interval maps in [4], but they are interesting by themselves, that is why we state them here.

**Proposition 3.3.** Let $G$ be a connected oriented graph which is not strongly positive recurrent and $W$ a finite set of vertices. Then for all integers $n$ there exists a connected subgraph $G_n \subset G$ such that $h(G_n) = h(G)$ and for all $w \in W$, for all $0 \leq k < n$, $f_{uw}^{G_n}(k) = 0$.

**Corollary 3.4.** Let $G$ be a connected oriented graph which is not strongly positive recurrent. Then there exists a sequence of ergodic Markov measures $(\mu_n)_{n \geq 0}$ such that $\lim_{n \to +\infty} h_{\mu_n}(\Gamma_G, \sigma) = h(G)$ and for all finite subsets of vertices $W$, $\lim_{n \to +\infty} \mu_n((\{u_n\}_{n \in \mathbb{Z}} \in \Gamma_G \mid u_0 \in W}) = 0$.

### 3.2. Local entropy and maximal measures

For a compact system, the *local entropy* is defined according to a distance but does not depend on it. One may wish to extend this definition to noncompact metric spaces although the notion obtained in this way is not canonical.

**Definition 3.5.** Let $X$ be a metric space, $d$ its distance and let $T : X \to X$ be a continuous map.

The *Bowen ball* of centre $x$, of radius $r$ and of order $n$ is defined as

$$B_n(x, r) = \{y \in X \mid d(T^i x, T^i y) < r, 0 \leq i < n\}.$$

$E$ is a $(\delta, n)$-separated set if

$$\forall y, y' \in E, y \neq y', \exists 0 \leq k < n, d(T^k y, T^k y') \geq \delta.$$
The maximal cardinality of a \((\delta, n)\)-separated set contained in \(Y\) is denoted by \(s_n(\delta, Y)\).

The local entropy of \((X, T)\) is defined as \(h_{\text{loc}}(X) = \lim_{\varepsilon \to 0} h_{\text{loc}}(X, \varepsilon)\), where

\[
    h_{\text{loc}}(X, \varepsilon) = \lim_{\delta \to 0} \limsup_{n \to +\infty} \frac{1}{n} \sup_{x \in X} s_n(\delta, B_n(x, \varepsilon)).
\]

If the space \(X\) is not compact, these notions depend on the distance. When \(X = \Gamma_G\), we use the distance \(d\) introduced in Section 1.2. The local entropy of \(\Gamma_G\) does not depend on the identification of the vertices with \(\mathbb{N}\).

**Proposition 3.6.** Let \(\Gamma_G\) be the topological Markov chain on \(G\) and \(\Gamma_G\) its compactification as defined in Section 1.2. Then \(h_{\text{loc}}(\Gamma_G) = h_{\text{loc}}(\Gamma_G)\).

**Proof.** Let \(\bar{\sigma} = (u_n)_{n \in \mathbb{Z}} \in \Gamma_G\), \(\varepsilon > 0\) and \(k \geq 1\). By continuity there exists \(\eta > 0\) such that, if \(\bar{\sigma} \in \Gamma_G\) and \(d(\bar{\sigma}, \bar{\sigma}) < \eta\) then \(d(\sigma^i(\bar{\sigma}), \sigma^i(\bar{\sigma})) < \varepsilon\) for all \(0 \leq i < k\). By definition of \(\Gamma_G\) there is \(\bar{\sigma} \in \Gamma_G\) such that \(d(\bar{\sigma}, \bar{\sigma}) < \eta\), thus \(\bar{\sigma} \in B_k(\bar{\sigma}, \varepsilon)\), which implies that \(B_k(\bar{\sigma}, \varepsilon) \subset B_k(\bar{\sigma}, 2\varepsilon)\). Consequently \(h_{\text{loc}}(\Gamma_G, \varepsilon) \leq h_{\text{loc}}(\Gamma, 2\varepsilon)\), and \(h_{\text{loc}}(\Gamma_G) \leq h_{\text{loc}}(\Gamma_G)\). The reverse inequality is obvious. \(\square\)

We are going to prove that, if \(h_{\text{loc}}(\Gamma_G) < h(G)\), then \(G\) is strongly positive recurrent. First we introduce some notations.

Let \(G\) be an oriented graph. If \(V\) is a subset of vertices, \(H\) a subgraph of \(G\) and \(\bar{\sigma} = (u_n)_{n \in \mathbb{Z}} \in \Gamma_G\), define

\[
    \mathcal{C}^H(\bar{\sigma}, V) = \{(v_n)_{n \in \mathbb{Z}} \in \Gamma_H : \forall n \in \mathbb{Z}, u_n \in V \Rightarrow (v_n = u_n), u_n \notin V \Rightarrow v_n \notin V\}.
\]

If \(S \subset \Gamma_G\) and \(p, q \in \mathbb{Z} \cup \{-\infty, +\infty\}\), define

\[
    [S]^q_p = \{(v_n)_{n \in \mathbb{Z}} \in \Gamma_G : \exists (u_n)_{n \in \mathbb{Z}} \in S, \forall p \leq n \leq q, u_n = v_n\}.
\]

**Lemma 3.7.** Let \(G\) be an oriented graph on the set of vertices \(\mathbb{N}\).

i) If \(V \supseteq \{0, \ldots, p + 2\}\) then for all \(\bar{\sigma} \in \Gamma_G\) and all \(n \geq 1\), \(\mathcal{C}^G(\bar{\sigma}, V) \subset B_n(\bar{\sigma}, 2^{-p})\).

ii) If \(\bar{\sigma} = (u_n)_{n \in \mathbb{Z}}\) and \(\bar{\sigma} = (v_n)_{n \in \mathbb{Z}}\) are two paths in \(G\) such that \((u_0, \ldots, u_{n-1}) \neq (v_0, \ldots, v_{n-1})\) and \(u_i, v_i \in \{0, \ldots, q - 1\}\) for \(0 \leq i \leq n - 1\) then \((\bar{\sigma}, \bar{\sigma})\) is \((2^{-p}, n)\)-separated.

**Proof.** (i) Let \(\bar{\sigma} = (u_n)_{n \in \mathbb{Z}} \in \Gamma_G\). If \(\bar{\sigma} = (v_n)_{n \in \mathbb{Z}} \in \mathcal{C}^G(\bar{\sigma}, V)\), then \(D(u_j, v_j) \leq 2^{-i+p+2}\) for all \(j \in \mathbb{Z}\). Consequently for all \(0 \leq i < n\)

\[
    d(\sigma^i(\bar{\sigma}), \sigma^i(\bar{\sigma})) = \sum_{k \in \mathbb{Z}} D(u_{i+k}, v_{i+k}) \leq \sum_{k \in \mathbb{Z}} 2^{-i+p+2} = 3 \cdot 2^{-i+p+2} < 2^{-p}.
\]
(ii) Let $0 \leq i \leq n-1$ such that $u_i \neq v_i$. By hypothesis, $u_i, v_i \leq q-1$. Suppose that $u_i < v_i$. Then $d(\sigma^i(\overline{u}), \sigma^i(\overline{v})) \geq D(u_i, v_i) = 2^{-u_i}(1 - 2^{-(v_i-u_i)}) \geq 2^{-q}$.

\[\square\]

**Theorem 3.8.** Let $G$ be a connected oriented graph of finite entropy on the set of vertices $\mathbb{N}$. If $h_{\text{loc}}(\Gamma_G) < h(G)$, then the graph $G$ is strongly positive recurrent and the Markov chain $(\Gamma_G, \sigma)$ admits a maximal measure.

**Proof.** Fix $C$ and $\varepsilon > 0$ such that $h_{\text{loc}}(\Gamma_G, \varepsilon) < C < h(G)$. Let $p$ be an integer such that $2^{-(p-1)} < \varepsilon$. Let $G'$ be a finite subgraph such that $h(G') > C$ and let $V$ be a finite subset of vertices such that $V$ is connected and contains the vertices of $G'$ and the vertices $\{0, \ldots, p\}$. Define $W = V$, $V_q = \{n \leq q\}$ and $W_q = V_q \setminus V = W \cap V_q$ for all $q \geq 1$.

Our aim is to bound $t_{uu'}^q(n) = t_{uu'}^V(n)$. Choose $u, u' \in V$ and let $(w_0, \ldots, w_{n_0})$ be a path between $u'$ and $u$ with $w_i \in V$ for $0 \leq i \leq n_0$. Fix $n \geq 1$.

One has $t_{uu'}^W(n) = \lim_{q \to +\infty} t_{uu'}^q(n)$.

Fix $\delta_0 > 0$ such that

$$\forall \delta \leq \delta_0, \limsup_{n \to +\infty} \frac{1}{n} \sup_{\bar{v} \in \Gamma_G} \log s_n(\delta, B_n(\bar{v}, \varepsilon)) < C.$$ 

Take $q \geq 1$ arbitrarily large and $\delta \leq \min\{\delta_0, 2^{-(q+1)}\}$. Choose $N$ such that

$$\forall n \geq N, \forall \bar{v} \in \Gamma_G, \frac{1}{n} \log s_n(\delta, B_n(\bar{v}, \varepsilon)) < C.$$ (7)

If $t_{uu'}^q(n) \neq 0$, choose a path $(v_0, \ldots, v_n)$ such that $v_0 = u, v_n = u'$ and $v_i \in W_q$ for $0 < i < q$. Define $\bar{v}(n) = (v_i^{(n)})_{i \in \mathbb{Z}}$ as the periodic path of period $n + n_0$ satisfying $v_i^{(n)} = v_i$ for $0 \leq i \leq n$ and $v_{i+n_0}^{(n)} = w_i$ for $0 \leq i \leq n_0$.

Define the set $E_q(n, k)$ as follows (see Figure 2):

$$E_q(n, k) = \left[ C^{V_q}(\bar{v}(n), V) \right]_{0}^{k(n+n_0)} \cap \left[ \bar{v}(n) \right]_{0}^{-\infty} \cap \left[ \bar{v}(n) \right]_{k(n+n_0)}^{+\infty}.$$ 

The paths in $E_q(n, 1)$ are exactly the paths counted by $t_{uu'}^W(n)$ which are extended outside the indices $\{0, \ldots, n\}$ like the path $\bar{v}(n)$, thus $\#E_q(n, 1) = t_{uu'}^q(n)$. Similarly, $\#E_q(n, k) = \left(t_{uu'}^q(n)\right)^k$.

By definition, $E_q(n, k) \subset C^G(\bar{v}(n), V)$ and $\{0, \ldots, p\} \subset V$ thus $E_q(n, k) \subset B_{k(n+n_0)}(\bar{v}(n), \varepsilon)$ by Lemma 3.7(i). Moreover, if $(w_i)_{i \in \mathbb{Z}}$ and $(w_i')_{i \in \mathbb{Z}}$ are two distinct elements of $E_q(n, k)$, there exists $0 \leq i < k(n+n_0)$ such that $w_i \neq w_i'$ and $w_i, w_i' \leq q$, thus $E_q(n, k)$ is a $(\delta, k(n+n_0))$-separated set by Lemma 3.7(ii). Choose $k$ such that $k(n+n_0) \geq N$. Then by Equation (7)

$$\#E_q(n, k) \leq s_{k(n+n_0)}(\delta, B_{k(n+n_0)}(\bar{v}(n), \varepsilon)) < e^{k(n+n_0)C}.$$
As \( |E_q(n, k)| = \left( t_{uu'}^{(n)}(n) \right)^k \), one gets \( t_{uu'}^{(n)}(n) < e^{(n+n_0)C} \). This is true for all \( q \geq 1 \), thus
\[
t_{uu'}^{(n)}(n) = \lim_{q \to +\infty} t_{uu'}^{(n)}(n) \leq e^{(n+n_0)C}
\]
and
\[
\tau_{uu'}^{W} = \tau_{uu'}^{V} \leq C < h(V).
\]
Theorem 3.2 concludes the proof.

**Remark 3.9.** Define the entropy at infinity as \( h_\infty(G) = \lim_{n \to +\infty} h(G \setminus G_n) \) where \( (G_n)_{n \geq 0} \) is a sequence of finite graphs such that \( \bigcup_n G_n = G \). The local entropy satisfies \( h_{\text{loc}}(\Gamma_G) \geq h_\infty(G) \) but in general these two quantities are not equal and the condition \( h_\infty(G) < h(G) \) does not imply that \( G \) is strongly positive recurrent. This is illustrated by Example 2.9 (see Figure 1).

References


Received April 26, 2001 and revised June 20, 2002.

Institut de Mathématiques de Luminy
CNRS UPR 9016
163 avenue de Luminy, case 907
13288 Marseille Cedex 9
France
E-mail address: ruette@iml.univ-mrs.fr
REMOVABLE SINGULARITIES FOR YANG–MILLS CONNECTIONS IN HIGHER DIMENSIONS

Baozhong Yang

We prove several removable singularity theorems for singular Yang–Mills connections on bundles over Riemannian manifolds of dimensions greater than four. We obtain the local and global removability of singularities for Yang–Mills connections with $L^\infty$ or $L^{n/2}$ bounds on their curvature tensors, with weaker assumptions in the $L^\infty$ case and stronger assumptions in the $L^{n/2}$ case. With the global gauge construction methods we developed, we also obtain a ‘stability’ result which asserts that the existence of a connection with uniformly small curvature tensor implies that the underlying bundle must be isomorphic to a flat bundle.

1. Introduction.

Uhlenbeck’s original paper [11] on removing isolated singularities of Yang-Mills connections on four manifolds is important not only in its applications in the compactification of the moduli space of self-dual connections on Riemannian four manifolds, but also in the analytic techniques introduced in it. Later on, there has been much work on the removable singularities for Yang-Mills connections. Some of these work focused on the case of isolated singularities for connections in different dimensions and possibly coupled with a section (a Higgs field), such as [1], [4], [6] and [7]. Some other works treated more general singularities, such as [5], [3], [8] and [10]. The fundamental work of Tian in [10] on the analysis of Yang-Mills connections in higher dimensions gave some guidance on what we should expect about the singularities of Yang-Mills connections on manifolds of dimensions greater than 4. It turns out that if we are considering connections within the compactification of smooth Yang-Mills connections on an $n$-dimensional manifold, then the most general type of singularities to start with is probably an $H^{n-4}$-rectifiable closed set. While the most general removable singularity theorem in higher dimensions has not been proved yet, this paper is an effort to understand the removable singularities and the related gauge problems in higher dimensions.

We shall assume that all the vector bundles in this paper have a compact structure group $G$ and all connections and gauge transformations refer to
We fix a metric on $G$ by embedding $G$ into some orthogonal group and denote by $\text{Id}$ the unit element of $G$.

Before stating our main results in this paper, we shall first clarify the meaning of removable singularities we are going to use here. Because of the technical difficulty of defining suitable concepts of weak solutions to the Yang-Mills equations, instead of considering regularity theories for weak solutions (as in the case of harmonic maps and some other nonlinear PDE problems), people usually consider removable singularity theorems in the following (or similar and slightly different) context: We shall usually consider a Yang-Mills connection $A$ on a vector bundle $E$ on a manifold $M$ with a closed singular set $S$, that is, $A$ is defined and smooth on $M \setminus S$ and $A$ satisfies the Yang-Mills equation on $M \setminus S$. We say that the singularity of $A$ is removable if there exists a vector bundle $E'$ on $M$ such that $\phi : E|_{M \setminus S} \to E'$ gives an embedding of vector bundles preserving the $G$-structures, and there exists a gauge transformation $g$ of $E'$, defined and smooth on $M \setminus S$ such that the connection $g(A)$ is identified under $\phi$ with a connection on $E'|_{M \setminus S}$, which is the restriction of a smooth connection on $E'$ over $M$. We note that if the original connection $A$ is singular on $S$, then the smoothing gauge transformation $g$ must be discontinuous on $S$, hence in general $E'$ and $E$ may not be topologically isomorphic as vector bundles. Under a local trivialization of $E$, the connection $A$ may be identified with a $G$-valued 1-form, where $G$ is the Lie algebra of $G$. Then locally the removability of the singularity of $A$ is equivalent to the existence of a $G$-valued function $g$ (which is smooth away from $S$) such that $g(A) = gAg^{-1} - dg \cdot g^{-1}$ can be extended to a smooth $G$-valued 1-form.

In [10, 2.3], an admissible Yang-Mills connection is defined as a Yang-Mills connection $A$ with a closed singular set $S$ such that the $(n-4)$-dimensional Hausdorff measure of $S$ is locally finite and $\text{YM}(A) = \int_M |F_A|^2 < \infty$. The connections we considered in this work are within the class of admissible Yang-Mills connections. Another important notion is the stationarity of a connection, which in particular implies the monotonicity formula for the scaling-invariant Yang-Mills functional, see [10, 2.1]. Since we assume $L^\infty$ or $L^{n/2}$ boundedness of the curvature in this paper, the connections we consider satisfy the monotonicity for $L^{n/2}$ norm of the curvature trivially, hence we don’t need to assume stationarity here.

Our first result is the following local removable singularity theorem for singular connections with $L^\infty$ bounds on their curvatures.

**Theorem 1.** Let $E$ be the trivial bundle over the Euclidean unit cube $U = (0,1)^n \subset \mathbb{R}^n$ with the standard product metric. Assume that $A$ is an admissible Yang-Mills connection on $E$ with singular set $S$. Then there exists $\varepsilon_1 = \varepsilon_1(n, G) > 0$, such that if

$$
\|F_A\|_{L^\infty(U)} \leq \varepsilon_1,
$$

(1)
then the singularity of $A$ is removable over $U$. 

We note that there is no extra assumption about the singularity set $S$ except closedness and the dimensional requirement. With some global gauge patching arguments, we have the following global version of the above theorem:

**Theorem 2.** We assume that $M$ is a compact Riemannian manifold such that all representations $\pi_1(M) \to G$ are the trivial one and $E$ is the trivial smooth bundle over $M$ with a smooth $G$-structure. Assume that $A$ is a Yang-Mills connection on $E$ with singularity set $S$. Then there exists $\varepsilon_2 = \varepsilon_2(M,G) > 0$, such that if

$$\|FA\|_{L^\infty(M)} \leq \varepsilon_2,$$

then the singularity of $A$ is globally removable.

We should mention that without the triviality of the bundle $E$, there might not be global smoothing gauges for $A$ even if the singularity of $A$ is locally removable (see the remark at the end of Section 4). Our global patching arguments also yield some ‘stability’ results which roughly mean that a bundle close to a flat bundle (in some sense) must be flat. In particular, we have:

**Theorem 3** (Corollary 2). Assume that $M$ is a compact $n$-dimensional Riemannian manifold and $E$ is a smooth vector bundle over $M$ with a smooth connection $A$ on it. Then there exists a constant $\varepsilon_9 = \varepsilon_9(M) > 0$, such that if

$$\|FA\|_{L^\infty(M)} \leq \varepsilon_9,$$

then $E$ is smoothly isomorphic to a flat bundle.

It might be possible to improve the $L^\infty$ norm bounds in the above theorem to some $L^p$ ($p < \infty$) bounds. In the next two theorems, we use the $L^2$ norm instead of the $L^\infty$ norm of the curvature of the connection. We also assume the singularity set to be a manifold.

Assume that $4 \leq k \leq n$ is an integer. Let $B_1^k$ be the open unit $k$-ball in $\mathbb{R}^k$, and $D_1 = B_1^{n-k} \times B_1^k \subset \mathbb{R}^n$ be the Cartesian product of two balls. Assume that $E$ is a trivial vector bundle on $D_1 \setminus (B_1^{n-k} \times \{0\})$. We shall consider connections with singularity $B_1^{n-k} \times \{0\} \subset D_1$. This is the standard local model for connections with singularities being manifolds of codimension at least 4.

**Theorem 4.** Assume that $\tilde{A}$ is a Yang-Mills connection on $E$ with the singularity $B_1^{n-k} \times \{0\} \subset D_1$. Then there exists a constant $\varepsilon_3 = \varepsilon_3(n,k,G) > 0$, such that if

$$\|F\tilde{A}\|_{L^2(D_1 \setminus (B_1^{n-k} \times \{0\}))} \leq \varepsilon_3,$$

then the singularity of $A$ is removable over $D_{\frac{3}{2}} = B_{\frac{3}{2}}^{n-k} \times B_{\frac{3}{2}}^{n-k}$.
The following global version is a corollary of Theorem 2 and Theorem 4:

**Theorem 5.** We assume that $M$ is a compact Riemannian manifold such that all representations $\pi_1(M) \to G$ are the trivial one and $E$ is the trivial smooth bundle over $M$ with a smooth $G$-structure. Assume that $A$ is a Yang-Mills connection on $E$ such that the singularity set $S$ of $A$ is a closed smooth submanifold of codimension $\geq 4$. Then there exists $\varepsilon_4 = \varepsilon_4(M, G) > 0$, such that if

$$\|F_A\|_{L^n_2(M)} \leq \varepsilon_4, \quad \forall x \in M,$$

then the singularity of $A$ is removable.

We make the assumption that $S$ is a submanifold because we need the good product local model to prove Theorem 4. It is conceivable that this assumption may be relaxed. A conjecture is that we don’t need any additional assumption on $S$.

We remark here also that if we allow the constant $\varepsilon_4$ in Theorem 5 to depend also on the singularity set $S$, then the conclusion of Theorem 5 follows from the local theorem, Theorem 4 directly without the need of Theorem 2. We also would like to point out that a similar result to Theorem 4 for coupled Yang-Mills-Higgs fields has been proved in Thomas Otway’s paper [3]. Our proof of Theorem 4 here is based on the work of Rado [5] on singular connections on four manifolds with codimension two singularities. After we finished this work, we learned that in a recent work [9], Tao and Tian proved the local removability of singularities for stationary admissible Yang-Mills connections with singularities being manifolds. That will be a stronger result than Theorem 4 here.

In Section 2 we prove Theorem 4 and Theorem 5. In Section 3 we use some local gauge patching techniques to prove Theorem 1. In Section 4, we develop some global gauge patching results, including Theorem 3 and finally prove Theorem 2.

2. Removable singularities with $L^n_2$ norm bounds of curvatures.

Assume that $3 \leq k \leq n$ is an integer. Let $B^k_1$ be the open unit $k$-disk in $\mathbb{R}^k$, and $D_1 = B^{n-k}_1 \times B^k_1 \subset \mathbb{R}^n$. Assume that $E \to D_1 \setminus (B^{n-k}_1 \times \{0\})$ is a trivial vector bundle.

**Theorem 6.** There exist constants $\varepsilon_5 = \varepsilon_5(n, k, G) > 0$ and $C = C(n, k, G) > 0$, such that if $\nabla + A$ is a connection on $E$ with the singularity $B^{n-k}_1 \times \{0\}$, $A \in L^n_2_{1,\text{loc}}(D_1 \setminus (B^{n-k}_1 \times \{0\}))$, and $F_A \in L^n_2(D_1)$, and

$$\|F_A\|_{L^n_2(D_1)} \leq \varepsilon_5,$$
then $\nabla + A$ is gauge equivalent, by a gauge transformation in $L_{2, \text{loc}}^n(D_1 \setminus (B_1^{n-k} \times \{0\}))$ to a connection of the form $\nabla + \widetilde{A}$, $\widetilde{A} \in L_{2}^n(D_1)$, and

$$\|\widetilde{A}\|_{L_{2}^n(D_1)} \leq C \|F_A\|_{L_{2}^n(D_1)}.$$  

Proof. This theorem is a higher dimensional generalization of Theorem 2.1 in Johan Råde [5]. Actually the setting of the theorem in [5] is even more complicated because Råde considered the possibility of nontrivial holonomy around a codimension 2 singularity. Since the codimension of the singularity here is at least 3, the complement of the singular set, $D_1 \setminus (B_n - k \times \{0\})$, is simply connected and we can follow the proof in [5] omitting the part involving holonomy to prove the theorem. Of course, since we are in the $n$-dimensional setting, we need to modify the statements of the lemmas and results in [5] (stated in 4-dimensional setting there) accordingly. Mainly we just need to adjust the indices of the various objects and norms. It is not hard to check that the proofs there still work out with these changes and we shall not reproduce the details here. \hfill \Box

**Proposition 1.** An admissible Yang-Mills connection $A$ on a vector bundle $E$ over $M$ is a weak solution of the Yang-Mills equation, i.e.,

$$\int_M \langle d_A \omega, F_A \rangle = 0, \quad \text{for any } \omega \in C_c^\infty(M, T^*M \otimes \text{Ad } E).$$

Proof. This type of results are well-known to analysts. For completeness, we give a proof here. Since the question is local, we may assume that $M$ and the singular set $S$ of $A$ are compact. Assume that $H^{n-4}(M) = m < \infty$. For any $\delta > 0$, we may find finitely many (geodesic) open balls $B_{r_i}(x_i)$ of radii $r_i < \delta$, such that $x_i \in S$, $S \subset \cup B_i$ and $\sum r_i^{n-4} \leq C m$. Choose cutoff functions $\phi_i$ such that $\phi_i = 0$ on $B_{r_i}(x_i)$, $\phi_i = 1$ on $M \setminus B_{2r_i}(x_i)$, $0 \leq \phi_i \leq 1$ on $M$ and $|\nabla \phi_i| \leq Cr_i^{-1}$ on $M$. Let $\phi = \phi_\delta = \inf \phi_i$ on $M$. Then $\phi(x)$ is supported away from $S$ and if we let $N_{2\delta}(S) = \{y \in M : \text{dist}(y, S) \leq 2\delta\}$, then $\phi(x) = 1$ if $x \in M \setminus N_{2\delta}S$. We have

$$|\nabla \phi(x)| \leq \sup_i |\nabla \phi_i(x)|, \quad \forall x \in M.$$

Now we have, for any $\omega \in \Omega(\text{Ad } E)$,
\[ \int_M \phi \langle d_A \omega, F_A \rangle \]
\[ = - \int_M \phi \text{tr}(d_A \omega \wedge *F_A) \]
\[ = - \int_M d(\phi \text{tr}(\omega \wedge *F_A)) + \int_M \phi \wedge \text{tr}(\omega \wedge *F_A) + (-1) \int_M \phi \text{tr}(\omega \wedge d_A *F_A) \]
\[ = \int_M \phi \wedge \text{tr}(\omega \wedge *F_A), \text{ because } A \text{ is smooth Yang-Mills away from } S. \]

Now
\[ \left| \int_M \phi \wedge \text{tr}(\omega \wedge *F_A) \right| \leq C \int_M |\nabla \phi| |F_A| \]
\[ \leq C \left( \int_{N_{2\delta}(S)} |\nabla \phi|^2 \right)^{\frac{1}{2}} \left( \int_{N_{2\delta}(S)} |F_A|^2 \right)^{\frac{1}{2}} \]
\[ \leq C \left( \sum_{i} \int_{B_{2\varepsilon_i}(x_i)} |\nabla \phi_i|^2 \right)^{\frac{1}{2}} \left( \int_{N_{2\delta}(S)} |F_A|^2 \right)^{\frac{1}{2}} \]
\[ = C \left( \sum_{i} r_i^{n-2} \right) \left( \int_{N_{2\delta}(S)} |F_A|^2 \right)^{\frac{1}{2}} \]
\[ \leq C \left( \int_{N_{2\delta}(S)} |F_A|^2 \right)^{\frac{1}{2}}. \]

If we let \( \delta \rightarrow 0 \), then the last right-hand side goes to 0 by the \( L^2 \) integrability of \( F_A \). On the other hand
\[ \int_M \phi \theta(d_A \omega, F_A) \rightarrow \int_M (d_A \omega, F_A). \]
This gives the weak Equation (6). \( \square \)

**Remark.** We note here that if \( A \in L^\frac{n}{2} \) and \( A \) satisfies the Yang-Mills equation weakly, then for any gauge transformation \( g \in L^\frac{n}{2}, g(A) = gAg^{-1} - dgg^{-1} \) is still a weak solution of the Yang-Mills equation. The reason is as follows: We have \( d^*_{g(A)}F_{g(A)} = g(d_A^*F_A)g^{-1} \). By Sobolev embedding theorems, \( g, g^{-1} \in L^\frac{n}{2} \) and \( d_A^*F_A = 0 \) in \( L^\frac{n}{2} \) imply that \( g(d_A^*F_A)g^{-1} \) is
well-defined in $L^{p-1}$, for any $p < \frac{n}{2}$. Hence $d^*_g(A^g(A) = 0$ in $L^{p-1}_-$.
That implies $g(A)$ satisfies the Yang-Mills equation weakly.

Now we are ready to prove Theorem 4 — the $\varepsilon$-regularity theorem for admissible Yang-Mills connections with $L^n_2$ bounds on the curvature.

**Proof of Theorem 4.** Assume that $\varepsilon_3 < \varepsilon_5$. Then the connection satisfies the assumption of Theorem 6. We first apply Theorem 6 to the given connection to obtain a $L^n_2$ gauge in which the connection, still denoted $A$, is in $L^n_1$. Now we may apply the existence theorem of Hodge gauges, Theorem 1.3 in Uhlenbeck [12] to see that if $\varepsilon_3$ is sufficiently small, then after a further $L^n_2$ gauge transformation, we can make the resulting connection, again denoted $A$, to be in the Hodge gauge, i.e.,

$$d^*A = 0,$$

and with the following elliptic boundary condition:

$$*A = 0, \quad \text{on } \partial D_1.$$  

Since the original admissible Yang-Mills connection is a weak solution of the Yang-Mills equation by Prop. 1, and the gauges we have used are all in $L^n_2$, it follows from the previous remark that we have the weak equation,

$$d^*A F_A = 0.$$  

Now (7), (8) and (9) form a uniform elliptic system with $A \in L^n_1(D_1)$. Therefore, by standard elliptic theory, we may obtain higher regularities and the smoothness of $A$ in $D_1$. This removes the singularity. □

**Proof of Theorem 5.** If $\varepsilon_4$ is sufficiently small, then the assumptions of Theorem 4 are satisfied locally and we have that locally the singularity of $A$ is removable. Since $|F_A|$ is gauge-invariant, we have in particular that $|F_A|^2$ is a smooth function on $M$. However, $|F_A|^2$ satisfies a Bochner-Weitzenböck formula and hence the following a priori estimates (by Uhlenbeck, see also [2]):

$$|F_A|^2(x) \leq C \left( \rho^{-n} \int_{B_\rho(x)} |F_A|^{\frac{n}{2}} dv \right)^{\frac{4}{n}}, \quad \forall x \in M.$$  

We remark here that although the a priori estimates are usually stated for smooth (or stationary) Yang-Mills connections and for scaling invariant $L^2$ energies $\rho^{4-n} \int_{B_\rho(x)} |F_A|^2 dv$, we have here the bounds on the $L^n_2$ energy which is itself scaling invariant. Corresponding to the monotonicity formula used in the proof of the usual a priori estimates, we have trivially that

$$\int_{B_\rho(x)} |F_A|^{\frac{n}{2}} dv \leq \int_{B_\sigma(x)} |F_A|^{\frac{n}{2}} dv, \quad \text{if } \rho \leq \sigma.$$
Hence the proof of the usual a priori estimates (see \[2\] or \[10, 2.2.1\]) can be adapted (almost word by word) to give our version (10). We shall not give the details and refer the reader to the references.

By (10), if $\epsilon_4$ is small enough, we have

$$\|F_A\|_{L^\infty(M)} \leq \epsilon_2,$$

and hence Theorem 2 applies to give the global removability of the singularity of $A$. $\square$

3. Removable singularities with $L^\infty$ norm bounds of curvatures.

The essence of the Proof of Theorem 1 is a construction of a smoothing gauge transformation. The method we shall use here is a refinement of Uhlenbeck’s method to construct global gauges on compact manifolds used in §3 of \[12\]. The idea of this argument is to modify and glue suitable gauges on different patches inductively to obtain a global gauge. Here because we have infinitely many patches, we have to keep a careful track of the gluing procedure to make sure the gauges we obtained are always suitably bounded, thus amenable for further gluing in the induction.

Before giving the proofs, we introduce the following definitions to make the statements simpler. Let $M$ be a Riemannian manifold and let the index set $I$ be either the set of natural numbers or the set \{1, 2, \ldots, n\} for some integer $n$.

**Definition.** Let $c > 0$ be a constant and $K > 0$ be an integer. We call a countable collection of open subsets of $M$, \{$U^i_\alpha\}_{\alpha \in I, 1 \leq i \leq K}$, a \((c,K)\)-uniform nested covering of $M$ if the following conditions are satisfied:

1) $\overline{U^i_{\alpha + 1}} \subset U^i_\alpha$, for $1 \leq i \leq K - 1$.
2) $M \subset \bigcup_{\alpha \in I} U^K_\alpha$.
3) $\# \{\alpha : x \in U^1_\alpha\} \leq K$, $\forall x \in M$.
4) The diameters $r_\alpha = \text{diam}(U^1_\alpha)$ satisfy,

$$r_\alpha \leq cr_\beta, \quad \text{if } U^1_\alpha \cap U^1_\beta \neq \emptyset.$$

This definition of nested open sets is natural because we shall see that in the gluing procedure, we need to shrink the open sets each time we try to glue gauges on overlapping open patches. Let \{$U^i_\alpha\}_{\alpha \in I, 1 \leq i \leq K}$ be a \((c,K)\)-uniform covering of $M$. We shall define integers $i^\alpha_\beta$ for all pairs $(\alpha, \beta)$ satisfying $\alpha \geq \beta$. First we define for any $\alpha \in I$, $i^\alpha_\alpha = 1$. Then we define inductively for $\alpha > \beta$ that

$$i^\alpha_\beta = \begin{cases} i^{\alpha - 1}_\beta + 1, & \text{if } U^{\alpha - 1}_\beta \cap U^1_\alpha \neq \emptyset, \\ i^\alpha_\beta, & \text{otherwise.} \end{cases}$$

(11)
Note that because of 3) in the definition, for any \( \alpha \in I \), \( 1 = i_\alpha^0 \leq i_\alpha^{a+1} \leq \cdots \leq K \), and the increasing sequence stabilizes to an integer, which we will denote by \( i_\alpha \).

**Definition.** We call a \((c, K)\)-uniform nested covering \( \{U_\alpha^1\}_{\alpha \in I}, 1 \leq i \leq K \) of \( M \) a good \((c, K)\)-uniform nested covering if there exist functions \( \psi_\alpha \in C^\infty(U_\alpha^1) \) such that for each \( \alpha \in I \),

\[
\psi_\alpha \equiv 1, \quad \text{on } U_\alpha^1 \cap \left( \bigcup_{\beta < \alpha} U_\beta^{i_\beta^\alpha} \right),
\]

\[
\psi_\alpha \equiv 0, \quad \text{on } U_\alpha^1 \setminus \left( \bigcup_{\beta < \alpha} U_\beta^{i_\beta^\alpha-1} \right),
\]

\[
0 \leq \psi_\alpha \leq 1, \quad r_\alpha |\nabla \psi_\alpha| \leq c, \quad \text{on } U_\alpha^1, \quad \text{where } r_\alpha = \text{diam}(U_\alpha^1).
\]

We recall that a collection of transition functions \( \{g_\alpha\beta\} \) with respect to an open covering \( \{U_\alpha\} \) of \( M \) consist of functions \( g_\alpha\beta : U_\alpha \cap U_\beta \to G \) on \( U_\alpha \cap U_\beta \neq \emptyset \) such that:

1) \( g_\alpha\beta g_\beta\alpha = \text{Id}, \) on \( U_\alpha \cap U_\beta, \)

2) \( g_\alpha\beta g_\beta\gamma g_\gamma\alpha = \text{Id}, \) on \( U_\alpha \cap U_\beta \cap U_\gamma. \)

**Definition.** We call a collection of transition functions \( g_\alpha\beta \in C^1(U_\alpha \cap U_\beta, G) \) a collection of \( \delta \)-small transition functions with respect to a covering \( \{U_\alpha\}_{\alpha \in I} \) with length scales \( \{r_\alpha\}_{\alpha \in I} \) if, setting \( r_\alpha = \text{diam}(U_\alpha) \), we have

\[
|g_\alpha\beta - \text{Id}| + r_\alpha |\nabla g_\alpha\beta| \leq \delta r_\alpha, \quad \text{on } U_\alpha \cap U_\beta.
\]

The above technical definitions will enable us to track the bounds effectively in the gluing procedure.

**Lemma 1.** For any \( c_0 > 0 \) and \( K \) a positive integer, there exists \( \delta_0 = \delta_0(c_0, K, \text{diam}(M), G) > 0 \) such that if \( \delta < \delta_0 \), \( \{U_\alpha^1\}_{\alpha \in I}, 1 \leq i \leq K \) is a good \((c_0, K)\)-uniform nested covering of \( M \) and \( g_\alpha\beta \) is a collection of \( \delta \)-small transition functions with respect to the covering \( \{U_\alpha^1\}_{\alpha \in I} \), then there exist a collection of functions \( h_\alpha \in C^1(U_\alpha^1, G) \) and a constant \( C = C(c_0, n, K, G) > 0 \) such that

\[
g_\alpha\beta = h_\alpha^{-1}h_\beta, \quad \text{on } U_\alpha^1 \cap U_\beta^{i_\beta^\alpha},
\]

\[
|h_\alpha - \text{Id}| + r_\alpha |\nabla h_\alpha| \leq C\delta r_\alpha, \quad \text{on } U_\alpha^1.
\]

**Proof.** For simplicity, we use \( U_\alpha \) to denote \( U_\alpha^1 \) in the proof. We shall prove by induction on \( \alpha \in I \) that there exist \( h_\alpha \in C^1(U_\alpha, G) \) and constants \( C(k) = C(k, n, c_0, G) > 0 \) for \( 1 \leq k \leq K \) such that for any \( \alpha \in I \),

\[
g_\beta\gamma = h_\beta^{-1}h_\gamma, \quad \text{on } U_\beta^{i_\beta^\alpha} \cap U_\gamma^{i_\gamma^\alpha}, \quad \beta, \gamma \leq \alpha,
\]

\[
|h_\alpha(x) - \text{Id}| + r_\alpha |\nabla h_\alpha(x)| \leq C(L_\alpha^2)\delta, \quad \forall x \in U_\alpha,
\]
where \( |\beta| = \# \{ \beta \leq \alpha : x \in U_\beta \} \leq K \). We note that if (18) and (19) hold for all \( \alpha \in I \), then by taking \( C = \max_{1 \leq k \leq K} C(k) \), (16) and (17) will follow immediately.

Let \( \alpha_0 \) be the smallest element of \( I \), we define \( h_{\alpha_0}(x) = \Id \in G \) for \( x \in U_{\alpha_0} \). Then (18) and (19) are trivially satisfied for \( \alpha = \alpha_0 \) (for any value of \( C(1) > 0 \)). Now suppose that we have defined \( h_\beta \in C^1(U_\beta, G) \) for all \( \beta < \alpha \) such that (18) and (19) hold for indices \( < \alpha \).

We define

\[
(20) \quad h_\alpha(x) = \exp(\psi_\alpha(x) \exp^{-1}(h_\beta(x)g_\beta(x))), \quad \forall x \in U_\alpha \cap U_\beta^{\alpha-1}, \forall \beta < \alpha, \\
(21) \quad h_\alpha(x) = \Id, \quad \forall x \in U_\alpha \setminus \left( \bigcup_{\beta < \alpha} U_\beta^{\alpha-1} \right).
\]

We note that by the induction hypothesis, if \( \beta < \alpha \), then \( |h_\beta(x)g_\beta(x) - \Id| \leq C|\delta r_\beta| \leq C\delta \) and \( |\phi_\beta(x)| \leq 1 \) by (14). Hence if \( \delta \) is sufficiently small, then the expression in (20) involving \( \exp \) and \( \exp^{-1} \) in (20) is meaningful. The definition in (20) is unambiguous for different choices of \( \beta \) because of the induction hypothesis (18) for indices \( \beta, \gamma < \alpha \) and the assumption that \( \{g_{\alpha\beta}\} \) are transition functions. We note that the definition (20) and (21) determine a well-defined \( h_\alpha \in C^1(U_\alpha, G) \) because of (12) and (13). It follows easily by (12) and (20), and the fact that \( g_{\beta\gamma} \) are transition functions that (18) holds for all \( \beta, \gamma \leq \alpha \).

If \( x \in U_\alpha \cap U_\beta^{\alpha-1} \) with \( \beta < \alpha \), then by (14), (15), (20), (21) and the induction hypothesis, there exists a constant \( C_1 = C_1(n, c_0) > 0 \), such that

\[
|h_\alpha(x) - \Id| \leq C_1|h_\beta(x)g_\beta(x) - \Id| \leq C_1(1 + C(l^{\alpha-1}))|\delta r_\alpha|,
\]

\[
r_{\alpha}|\nabla h_\alpha(x)| \leq C_1r_\alpha(|\nabla \psi_\alpha(x)||h_\beta(x)g_\beta(x) - \Id| + |\nabla h_\beta(x)| + |\nabla g_\beta(x)|) \leq C_1(1 + C(l^{\alpha-1}))|\delta r_\alpha|.
\]

It follows that if at the beginning we define \( C(1) = C_1 \) and inductively define \( C(k + 1) = C_1(1 + C(k)) \) (these definitions only depend on \( n, K \) and \( c_0 \)), then (19) will be true for \( \alpha \). This finishes the induction step and the proof of the lemma.

**Proof of Theorem 1.** We shall choose a collection \( \tilde{F} \) of disjoint open dyadic cubes in \( U \setminus S \) step by step in the following way (the Whitney decomposition): At the first step, we divide \( U \) into \( 2^n \) congruent disjoint open cubes with edges of length \( 1/2 \). For \( k \geq 2 \), in the \( k \)-th step, we consider dyadic cubes with edges of length \( 1/2^{k-1} \) which (or a cube containing it) have not been put in \( \tilde{F} \) in the previous steps; if such a cube \( C \) satisfies the following condition:

\[
\text{dist}(C, S) \geq \text{diam}(C),
\]
then we put $C$ in the collection $\tilde{F}$; otherwise, we subdivide $C$ into $2^n$ congruent disjoint smaller open cubes with edges of length $1/2^k$. It is easy to see that we obtain a collection of disjoint open cubes $\tilde{F} = \{C_\alpha : \alpha \in I\}$ this way and we have

$$U \cap \bigcup_{\alpha \in I} \overline{C}_\alpha = U \setminus S.$$  

We then let

$$\mathcal{F} = \left\{ U_\alpha = \left( \frac{9}{8} C_\alpha \right) \cap U : \alpha \in I \right\},$$

where $\frac{9}{8} C_\alpha$ is the dilation of $C_\alpha$ at the center of $C_\alpha$ with a factor $\frac{9}{8}$. The collection $\mathcal{F} = \{U_\alpha\}_{\alpha \in I}$ of open sets now satisfies:

1) $\bigcup_{\alpha \in I} U_\alpha = U \setminus S$.
2) If $U_\alpha \in \mathcal{F}$, then $\text{dist}(U_\alpha, S) \geq \frac{1}{2} \text{diam}(U_\alpha)$.
3) If $U_\alpha, U_\beta \in \mathcal{F}$, and $U_\alpha \cap U_\beta \neq \emptyset$, then $\text{diam} U_\beta \leq 3 \text{diam}(U_\alpha)$.
4) There exists a number $K = K(n)$, such that $\# \{\alpha : x \in U_\alpha \in \mathcal{F}\} \leq K$.

Let $\lambda(i) = \exp\left( \frac{K+1-i}{K} \log \frac{9}{8} \right)$ for $1 \leq i \leq K$ and define $U^i_\alpha = \lambda(i)C_\alpha \cap U$ for $\alpha \in I$ and $1 \leq i \leq K$. We note that $U^1_\alpha = U_\alpha$ and $\overline{C}_\alpha \subset U^K_\alpha$, $\mathcal{U}^{i+1}_\alpha \subset U^i_\alpha$ for $1 \leq i \leq K - 1$. It is then easy to check that there exists a constant $c_0 = c_0(n)$ such that $\{U^i_\alpha\}_{\alpha \in I, 1 \leq i \leq K}$ is a good $(c_0, K)$-uniform nested covering.

Under the trivialization of the bundle on $U$, the connection may be identified as a matrix valued 1-form $A$ and a gauge transformation can be viewed simply as a function from $U$ to $G$. We fix a point $x_0 \in U \setminus S$. Let $x_\alpha$ be the center of the cube $C_\alpha$. For any point $x \in U_\alpha$, we let $\gamma^x_\alpha$ be the shortest geodesic from $x_\alpha$ to $x$ inside $U_\alpha$ and define $\mu_\alpha(x) \in G$ to be the parallel transport of the bundle from $x_\alpha$ to $x$ along $\gamma^x_\alpha$, using the trivialization of the bundle.

Note that $\mu_\alpha(x_\alpha) = \text{Id}$. We regard $\mu_\alpha^{-1}$ as gauge transformations on $U_\alpha$, and denote $\mu_\alpha^{-1}(A)$ by $\tilde{A}_\alpha$. We use the normal spherical coordinates $\{r, \theta^i\}_{i=1,\ldots,n-1}$ centered at $x_\alpha$, where $r$ is the distance to $x_\alpha$. Assume that

$$\tilde{A}_\alpha = \tilde{A}_{\alpha,r} dr + \tilde{A}_{\alpha,i} d\theta^i, \quad \text{on } U_\alpha$$

and

$$F_{\tilde{A}_\alpha} = F_{\alpha,r} dr \wedge d\theta^i + F_{\alpha,ij} d\theta^i \wedge d\theta^j, \quad \text{on } U_\alpha.$$  

Then by the definition of $\mu_\alpha$, we have $\tilde{A}_{\alpha,r} \equiv 0$ on $U_\alpha$. Hence

$$(22) \quad \partial_r(\tilde{A}_{\alpha,i}) = F_{\alpha,ri}, \quad i = 1, \ldots, n - 1.$$
By integrating (22) and using that $\tilde{A}_\alpha(0) = 0$, we have

$$|\tilde{A}_\alpha(x)| \leq C|x| \int_0^1 |F_{\tilde{A}_\alpha}(tx)| dt$$

$$= C|x| \int_0^1 |F_A(tx)| dt \leq C|x| \varepsilon_1 \leq C\varepsilon_1 r_\alpha. \quad (23)$$

**Remark.** The above gauge coming from the parallel transport along a geodesic leading from a central point was introduced by Uhlenbeck in [11]. It may be called the *radially flat gauge* at $x_\alpha$. Properties (22) and (23) were also given in [11].

Since $\dim S < n - 2$, we may perturb the geodesic from $x_\alpha$ to $x_0$ slightly to be disjoint from $S$ and denote the perturbed curve by $l_\alpha$. We define $\sigma_\alpha(0) \in G$ to be the parallel transport of the bundle from $x_0$ to $x_\alpha$ along the curve $l_\alpha$ and let

$$\sigma_\alpha(x) = \mu_\alpha(x) \sigma_\alpha(0), \quad \forall x \in U_\alpha.$$ 

If $U_\alpha \cap U_\beta \neq \emptyset$, we denote the difference between the gauge transformations $\sigma_\alpha$ and $\sigma_\beta$ by

$$g_{\alpha\beta} = \sigma_\alpha^{-1} \sigma_\beta = \sigma_\alpha^{-1}(0) \mu_\alpha^{-1}(x) \mu_\beta(x) \sigma_\beta(0). \quad (24)$$

Now assume that $x \in U_\alpha \cap U_\beta$. Consider the closed curve

$$\gamma = l_\alpha^{-1}(\gamma_\alpha^{-1})^{-1} \gamma_\beta l_\beta.$$ 

We notice that $g_{\alpha\beta}(x) \in G$ represents the parallel transport of the bundle along the closed curve $\gamma$. Because $\dim(S) < n - 2$, by perturbation, there exists a triangle $\Delta$ spanning $\gamma$ with $\text{Area}(\Delta) \leq cr_\alpha$ for some constant $c = c(n)$ and $\Delta \cap S = \emptyset$. By a well-known relation between holonomy and curvature, we have

$$|g_{\alpha\beta}(x) - \text{Id}| \leq \int_\Delta |F_{\tilde{A}_\alpha}(y)| dy \leq c r_\alpha \varepsilon_1. \quad (25)$$

$$|g_{\alpha\beta}(x)| \leq C\varepsilon_1 r_\alpha, \quad \forall x \in U_\alpha.$$ 

Then (23) and the compactness of $G$ imply that

$$|\tilde{A}_\alpha(x)| \leq C\varepsilon_1 r_\alpha, \quad \forall x \in U_\alpha.$$ 

We have by the definition of $g_{\alpha\beta}$ that

$$dg_{\alpha\beta} = g_{\alpha\beta} \tilde{A}_\beta - \tilde{A}_\alpha g_{\alpha\beta}. \quad (27)$$

It follows from (26) and (27) that

$$|\nabla g_{\alpha\beta}(x)| \leq C\varepsilon_1 r_\alpha. \quad (28)$$
Given any $\delta > 0$, if we take $\varepsilon_1 = \delta/C$, then (25) and (28) imply that $\{g_{\alpha\beta}\}$ is a $\delta$-small collection of transition functions with respect to the covering $\{U_{\alpha}\}$ of $U \setminus S$ with length scales $\{r_{\alpha}\}$. Therefore Lemma 1 applies and gives the correction term $h_\alpha$ on each $U_{\alpha}$. Let $V_{\alpha} = U_{\alpha} \cap V$. The new gauges $\rho_{\alpha} = h_\alpha \sigma_{\alpha}^{-1}$ on $V_{\alpha}$ given by the correction of $h_\alpha$ now satisfy $\rho_{\alpha} = \rho_{\beta}$ on $V_{\alpha} \cap V_{\beta}$ (because of (16) and (24)). Hence the $\rho_{\alpha}$’s define a global gauge $\rho$ on $U \setminus S$. We have

$$|\rho(A)(x)| = |h_\alpha(\mathcal{A}_\alpha(x))| = |h_\alpha \mathcal{A}_\alpha h_\alpha^{-1} - dh_\alpha h_\alpha^{-1}|$$

$$\leq C(|\mathcal{A}_\alpha| + |\nabla h_\alpha|) \leq C\varepsilon_1, \quad \forall x \in V_{\alpha}.$$

By inspecting our gluing procedure, we know that we can actually require $\rho(A)$ to be smooth away from $S$, which implies that $\rho(A)$ is admissible and a weak solution of the Yang-Mills equation. (29) and the assumption that $\|F_A\| \leq \varepsilon_1$ implies that $\|A\|_{L^p_\rho} \leq C\varepsilon_1, \quad \forall p \leq \infty$

because $F_A = dA + A \wedge A$. Fix $\frac{n}{2} < p < \infty$, if $\varepsilon_1$ is sufficiently small, we may apply the implicit function theorem (see Theorem 2.7 in [11]) to obtain an $L^p_\rho$ gauge transformation $g$ on $U$ so that the connection $A' = g(\rho(A))$ satisfies

$$d^* A' = 0, \quad \text{on } U, \quad *A' = 0, \quad \text{on } \partial U$$

(30) as (7) and (8) in Theorem 4’s proof. In the new gauge, we have $A' \in L^p_\rho$. $A'$ is also the weak solution of

$$d^* A' F_{A'} = 0, \quad \text{on } U,$$

by the remark following the proof of Proposition 1. Therefore we obtain the smoothness of the connection by elliptic regularity and finishes the proof of the theorem.

4. Global removable singularity theorems.

Before we carry out the Proof of Theorem 2, we shall first prove the following theorem, which is also of independent interest:

**Theorem 7.** Assume that $M$ is a compact $n$-dimensional manifold. Let $U = \{U_{\alpha}\}_{\alpha \in I}$ be a finite open covering of $M$ and $\{g_{\alpha\beta}\}, g_{\alpha\beta} : U_{\alpha\beta} = U_{\alpha} \cap U_{\beta} \to G$, be a set of smooth transition functions with respect to $U$. Then there exist constants $\varepsilon_6 = \varepsilon_6(M, U) > 0$ and $C = C(M, U) > 0$, such that if

$$\sup_{x, y \in U_{\alpha\beta}} |g_{\alpha\beta}(x) - g_{\alpha\beta}(y)| = \delta \leq \varepsilon_6,$$

(31)
then there exist a collection of constant transition functions \( \{g^0_{\alpha\beta}\} \), a smaller covering \( \mathcal{V} = \{V_\alpha\} \) of \( M \), with \( V_\alpha \subset U_\alpha \) and \( M \subset \bigcup \mathcal{V} \), and a set of smooth functions \( \rho_\alpha : V_\alpha \to G \), such that

\[
\rho_\alpha g_{\alpha\beta} \rho_\beta^{-1} = g^0_{\alpha\beta}, \quad \text{on } V_\alpha \cap V_\beta
\]

(32)

and

\[
\sup_{x \in V_\alpha} |\rho_\alpha(x) - \text{Id}| \leq C\delta.
\]

(33)

In particular, the bundle defined by \( \{g_{\alpha\beta}\} \) is isomorphic to a flat bundle (defined by \( \{g^0_{\alpha\beta}\} \)).

**Proof.** We first claim that there exists an increasing continuous function \( \mu : [0, \infty) \to [0, \infty) \) with \( \mu(0) = 0 \), depending only on \( M \) and \( U \), such that if \( \delta \) is defined by the left-hand side of (31), then there exists a collection of constant transition functions \( \{g^0_{\alpha\beta}\} \) with respect to the covering \( U \) such that

\[
\sup_{x \in U_\alpha \cap U_\beta} |g_{\alpha\beta}(x) - g^0_{\alpha\beta}(x)| \leq \mu(\delta).
\]

(34)

We let \( J = \{(\alpha, \beta) | \alpha < \beta \in I, U_\alpha \cap U_\beta \neq \emptyset\} \) and \( K = \{(\alpha, \beta, \gamma) | \alpha < \beta < \gamma \in I, U_\alpha \cap U_\beta \cap U_\gamma \neq \emptyset\} \). Denote by \( G^J \) be the Cartesian product of \( |J| \) copies of \( G \) indexed by \( J \), with a general element \( a = (a_{\alpha\beta}) \), \( (\alpha, \beta) \in J \). Define \( G^K \) similarly. We then define a map \( \Phi : G^J \to G^K \) by

\[
\Phi((a_{\alpha\beta})) = (a_{\alpha\beta}a_{\beta\gamma}^{-1}a_{\alpha\gamma}^{-1}) \in G^K, \quad \text{for any } a = (a_{\alpha\beta}) \in G^J.
\]

(35)

\( \Phi \) is clearly a continuous map. We note that an element \( a = (a_{\alpha\beta}) \) of \( G^K \) gives a collection of constant transition functions with respect to \( U \) if and only if \( \Phi(a) = (\text{Id}, \ldots, \text{Id}) := 1 \in G^K \), i.e., if and only if \( a \in \Phi^{-1}(1) \). Assume that there doesn’t exist such a function \( \mu \) as claimed above, then there exists \( \varepsilon > 0 \), a sequence \( \delta_i \) decreasing to 0 and a sequence of sets of transition functions \( \{g^i_{\alpha\beta}\} \) with respect to \( U \), such that

\[
\sup_{x, y \in U_\alpha \cap U_\beta, a, \beta \in I} |g^i_{\alpha\beta}(x) - g^i_{\alpha\beta}(y)| = \delta_i,
\]

(35)

and

\[
\sup_{x \in U_\alpha \cap U_\beta, a, \beta \in I} |g^i_{\alpha\beta}(x) - a_{\alpha\beta}| \geq \varepsilon, \quad \forall (a_{\alpha\beta}) \in \Phi^{-1}(1), \ \forall i.
\]

(36)

We fix points \( x_{\alpha\beta} \in U_\alpha \cap U_\beta \). Because \( G^J \) is compact, we know that the sequence \( (g^i_{\alpha\beta}(x_{\alpha\beta})) \) in \( G^J \) contains a subsequence converging to an element \( (g^0_{\alpha\beta}) \in G^J \). Now the fact that \( \{g^i_{\alpha\beta}\} \) are sets of transition functions imply
that \( \{g^0_{\alpha\beta}\} \) is also a set of transition functions with respect to \( \mathcal{U} \), i.e., \( (g^0_{\alpha\beta}) \in \Phi^{-1}(1) \). The fact that \( g^i_{\alpha\beta}(x_{\alpha\beta}) \to g^0_{\alpha\beta} \) together with (35) imply that

\[
\sup_{x \in U_\alpha \cap U_\beta} |g^i_{\alpha\beta}(x) - g^0_{\alpha\beta}| \to 0, \text{ as } i \to \infty.
\]

This clearly contradicts (36) when \( i \). Thus the claim is verified.

Now we recall the following proposition by Uhlenbeck (Prop. 3.2 in [12]):

**Proposition 2.** Let \( \{g_{\alpha\beta}\} \) and \( \{h_{\alpha\beta}\} \) be two sets of smooth transition functions with respect to a covering \( \mathcal{U} = \{U_\alpha\}_{\alpha \in \mathcal{I}} \) of a compact manifold \( M \). There exist constants \( \varepsilon_7 = \varepsilon_7(M, \mathcal{U}) > 0 \) and \( C = C(M, \mathcal{U}) > 0 \), such that if

\[
\sup_{x \in U_\alpha \cap U_\beta} |g_{\alpha\beta}(x) - h_{\alpha\beta}(x)| \leq \varepsilon_7,
\]

then there exists a smaller covering \( \mathcal{V} = \{V_\alpha\}_{\alpha \in \mathcal{I}} \) of \( M \), with \( V_\alpha \subset U_\alpha \) and \( M \subset V \), and a set of smooth functions \( \rho_\alpha : V_\alpha \to G \), such that

\[
\rho_\alpha g_{\alpha\beta} \rho_\beta^{-1} = h_{\alpha\beta}, \quad \text{on } V_\alpha \cap V_\beta
\]

and

\[
\sup_{x \in V_\alpha} |\rho_\alpha(x) - \text{Id}| \leq C\delta.
\]

In particular, the bundle defined by \( \{g_{\alpha\beta}\} \) is smoothly isomorphic to the bundle defined by \( \{h_{\alpha\beta}\} \).

We apply Proposition 2 to \( \{g_{\alpha\beta}\} \) and \( \{g^0_{\alpha\beta}\} \) (found by the claim) and immediately see that the theorem holds.

There are some corollaries of Theorem 7 in the following:

**Corollary 1.** Assume that \( M \) is a compact Riemannian \( n \)-dimensional manifold. Let \( \mathcal{U} = \{U_\alpha\}_{\alpha \in \mathcal{I}} \) be a finite open covering of \( M \) such that any two points \( x, y \) in a nonempty intersection \( U_\alpha \cap U_\beta \) can be connected by a \( C^1 \) curve within \( U_\alpha \cap U_\beta \) with length \( \leq l \), a uniform constant, and let \( \{g_{\alpha\beta}\} \) be a set of smooth transition functions with respect to \( \mathcal{U} \). Then there exist constants \( \varepsilon_8 = \varepsilon_8(M, l, \mathcal{U}) > 0 \) and \( C = C(M, \mathcal{U}) > 0 \), such that if

\[
\sup_{x \in U_{\alpha\beta}} |\nabla g_{\alpha\beta}(x)| = \delta \leq \varepsilon_8,
\]

then we have the same conclusions as in Theorem 7. In particular, the bundle defined by \( \{g_{\alpha\beta}\} \) is smoothly isomorphic to a flat bundle.

**Proof.** We may easily deduce from (40) and the assumptions that the inequality (31) holds if we take \( \varepsilon_8 = \varepsilon_6/l \). Then we can apply Theorem 7 here. \( \square \)
Corollary 2. Assume that $M$ is a compact $n$-dimensional Riemannian manifold and $E$ is a smooth vector bundle over $M$ with a smooth connection $A$ on it. Then there exists $\varepsilon_9 = \varepsilon_9(M) > 0$, such that if
\[
\|F_A\|_{L^\infty(M)} \leq \varepsilon_9,
\]
then $E$ is smoothly isomorphic to a flat bundle.

Proof. We cover $M$ with coordinate balls $\{U_\alpha\}$, such that any two points $x, y$ in a nonempty intersection $U_\alpha \cap U_\beta$ can be connected by a $C^1$ curve within $U_\alpha \cap U_\beta$ with length \( \leq \text{diam}(M) \). Let $\phi_\alpha : E|_{U_\alpha} \to B_1(0) \times \mathbb{R}^l$ be trivializations on $U_\alpha$ and $A_\alpha$ be the $G$-valued 1-form on $U_\alpha$ corresponding to $A$ under $\phi_\alpha$. We use the radially flat gauge of $A$ on $U_\alpha$, i.e., we find $h_\alpha : U_\alpha \to G$ such that $h_\alpha(A_\alpha)(0) = 0$ and $h_\alpha(\partial_r) = 0$ under the local coordinates $U_\alpha \cong B_1(0) \subset \mathbb{R}^n$. We have from (23) that
\[
|h_\alpha(A_\alpha)(x)| \leq C\|F_A\|_{L^\infty(M)} \leq C\varepsilon_9, \ \forall x \in U_\alpha.
\]
We define $h_{\alpha\beta} = h_\alpha\phi_\alpha\phi_\beta^{-1}h_\beta^{-1}$ on $U_\alpha \cap U_\beta$ and we can check that $\{h_{\alpha\beta}\}$ is a set of transition functions. Now we have
\[
\text{d}h_{\alpha\beta} = h_\alpha(A_\alpha) \circ h_{\alpha\beta} - h_{\alpha\beta} \circ (h_\beta(A_\alpha))
\]
and hence from (42), we have
\[
|\text{d}h_{\alpha\beta}| \leq C\varepsilon_9, \text{ on } U_\alpha \cap U_\beta.
\]
Now by taking $\varepsilon_9$ sufficiently small we may apply Corollary 1 to establish the theorem. \qed

Proof of Theorem 2. We choose a covering $U = \{U_\alpha\}$ of $M$ such that each $U_\alpha$ is a coordinate cube and the metrics of $E$ and $M$ on $U_\alpha$ can be uniformly compared with the product metric and the Euclidean metric. We also require that any two points $x, y$ in a nonempty intersection $U_\alpha \cap U_\beta$ can be connected by a $C^1$ curve within $U_\alpha \cap U_\beta$ with length \( \leq \text{diam}(M) \). If $\varepsilon_2$ is taken small, we can now apply Theorem 1 for the connection $A$ on each $U_\alpha$ to obtain local gauge transformations $h_\alpha : U_\alpha \to G$ such that $h_\alpha(A)$ are smooth on $U_\alpha$ and furthermore, from the Proof of Theorem 1, we may require that
\[
|h_\alpha(A)| \leq C\|F_A\|_{L^\infty(M)} \leq C\varepsilon_2,
\]
for some uniform constant $C$. Now we define $h_{\alpha\beta} = h_\alpha h_\beta^{-1}$. $h_{\alpha\beta}$ must be smooth because it intertwines smooth connections $h_\alpha(A)$ and $h_\beta(A)$ on $U_\alpha \cap U_\beta$. We can then follow the lines of the last part in the proof of Cor. 2 to establish that there exist a refinement $V = \{V_\alpha\}$ of the covering $U$, a collection of smooth functions $\rho_\alpha : V_\alpha \to G$ and a collection of constant transition functions $g^0_{\alpha\beta}$ such that
\[
\rho_\alpha h_{\alpha\beta} \rho_\beta^{-1} = g^0_{\alpha\beta}, \text{ on } U_\alpha \cap U_\beta.
\]
Because every representation of $\pi_1(M) \to G$ is the trivial one, every flat bundle over $M$ is trivial. Hence there exists constants $\lambda_\alpha \in G$ such that $g_{\alpha\beta}^0 = \lambda_\alpha^{-1}\lambda_\beta$ for any $\alpha, \beta$. It follows that

$$\lambda_\alpha \rho_\alpha h_\alpha = \lambda_\beta \rho_\beta h_\beta, \text{ on } U_\alpha \cap U_\beta.$$ 

Hence we may define a global gauge transformation $g$ by letting $g = \lambda_\alpha \rho_\alpha h_\alpha$, on $U_\alpha$. $g(A)$ is smooth on $M$ and hence $g$ gives the desired smoothing gauge.

Remark. We remark that in Theorem 2 if the original bundle $E$ is not trivial, then although locally the singularity of the connection $A$ is removable, there may not exist a global gauge transformation which makes $A$ smooth on $M$. The following is a simple example of this lack of global smoothing gauge. In fact, this example has something to do with the lack of global smooth gauge transformations on certain bundles. Let $M = S^2 = \mathbb{C}P^1$ with the covering given by $U_1 = \{z : z \in \mathbb{C}\}$ and $U_2 = \{w : w \in \mathbb{C}\}$ with the coordinate change given by $z = 1/w$ on $U_{12} = U_1 \cap U_2$. We give a transition function for a line bundle $g_{12} : U_{12} \to \mathbb{C}^\ast$ via

$$g_{12}(z) = \frac{z}{|z|} \in \mathbb{C}^\ast.$$ 

Let $E$ be the smooth line bundle determined by $g_{12}$. It is easy to see that $E$ has the same smooth structure as the hyperplane bundle on $\mathbb{P}^1$. We then define a connection $A$ on $E$ by letting its local forms be $A_1 = -i\theta$ on $U_1$ and $A_2 = 0$ on $U_2$, where $\theta$ is the usual angle coordinate on $\mathbb{C}$. We can check that $g_{12}(A_1) = A_2$ and hence $A$ is a well-defined connection on $E$ with singularity $p = \{z = 0\} \subset U_1$. This singularity of $A$ can be removed on $U_1$ as follows. We define $\rho : U_1 \to \mathbb{C}$ by $\rho(z) = \bar{z}/|z|$. Then $\rho(A_1) = 0$ gives the smoothing of $A_1$ on $U_1$. However, it is clear from homotopical considerations that $\rho$ cannot be extended to a global gauge transformation smooth on $M - \{p\}$.

We may similarly construct an example of a nontrivial $SU(2)$ bundle on $S^4$ with a flat connection $A$ which is singular at one point but does not have a global smoothing gauge. We note that in these examples the bundles don’t allow global smooth gauge transformations.

Acknowledgment. The author would like to thank Rick Schoen and Gang Tian for many helpful discussions on this work and the referee for many useful suggestions on the revision of this paper.

References


Received November 27, 2001 and revised May 19, 2002. Research partially supported by NSF Grant DMS-0104163.

DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CA 94305
E-mail address: byang@math.stanford.edu
<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The discriminant of a symplectic involution</td>
<td>Grégory Berhuy, Marina Monsurrò and Jean-Pierre Tignol</td>
<td>201</td>
</tr>
<tr>
<td>Fusion and fission in graph complexes</td>
<td>James Conant</td>
<td>219</td>
</tr>
<tr>
<td>Some planar algebras related to graphs</td>
<td>Brian Curtin</td>
<td>231</td>
</tr>
<tr>
<td>Boolean algebras of projections and algebras of spectral operators</td>
<td>H.R. Dowson, M.B. Ghaemi and P.G. Spain</td>
<td>1</td>
</tr>
<tr>
<td>On rings which are sums of two PI-subrings: a combinatorial approach</td>
<td>B. Felzenszwalb, A. Giambruno and G. Leal</td>
<td>17</td>
</tr>
<tr>
<td>Complete contractivity of maps associated with the Aluthge and Duggal transforms</td>
<td>Ciprian Foiaș, Il Bong Jung, Eungil Ko and Carl Pearcy</td>
<td>249</td>
</tr>
<tr>
<td>Sommes de modules de sommes d’exponentielles</td>
<td>Etienne Fouvry and Philippe Michel</td>
<td>261</td>
</tr>
<tr>
<td>A method of Washington applied to the derivation of a two-variable p-adic L-function</td>
<td>Glenn J. Fox</td>
<td>31</td>
</tr>
<tr>
<td>Boolean algebras of projections and algebras of spectral operators</td>
<td>M.B. Ghaemi with H.R. Dowson and P.G. Spain</td>
<td>1</td>
</tr>
<tr>
<td>On rings which are sums of two PI-subrings: a combinatorial approach</td>
<td>A. Giambruno with B. Felzenszwalb and G. Leal</td>
<td>17</td>
</tr>
<tr>
<td>Bass numbers of semigroup-graded local cohomology</td>
<td>David Helm and Ezra Miller</td>
<td>41</td>
</tr>
<tr>
<td>Examples of bireducible Dehn fillings</td>
<td>James A. Hoffman and Daniel Matignon</td>
<td>67</td>
</tr>
<tr>
<td>Homotopy minimal periods for maps of three dimensional nilmanifolds</td>
<td>Jerzy Jezierski and Wacław Marzantowicz</td>
<td>85</td>
</tr>
<tr>
<td>A zeta function for flip systems</td>
<td>Young-One Kim, Jungseob Lee and Kyewon K. Park</td>
<td>289</td>
</tr>
<tr>
<td>On some pointwise inequalities concerning tent spaces and sharp maximal functions</td>
<td>Andrei K. Lerner</td>
<td>303</td>
</tr>
</tbody>
</table>
Zhi-Guo Liu: Some Eisenstein series identities related to modular equations of the seventh order 103

Charles Livingston: Observations on Lickorish knotting of contractible 4-manifolds 319

Harald Löwe: Sixteen-dimensional locally compact translation planes admitting $\text{SL}_2\mathbb{H}$ as a group of collineations 325

Waclaw Marzantowicz with Jerzy Jezierski 85

Daniel Matignon with James A. Hoffman 67

Philippe Michel with Etienne Fouvry 261

Ezra Miller with David Helm 41

Marina Monsurrò with Grégory Berhuy and Jean-Pierre Tignol 201

Matthew Neal and Bernard Russo: Operator space characterizations of $C^*$-algebras and ternary rings 339

Kyewon K. Park with Young-One Kim and Jungseob Lee 289

Carl Pearcy with Ciprian Foiaş, Il Bong Jung and Eungil Ko 249

J.C. Rosales and M.B. Branco: Irreducible numerical semigroups 131

Sylvie Ruette: On the Vere–Jones classification and existence of maximal measures for countable topological Markov chains 365

Bernard Russo with Matthew Neal 339

P.G. Spain with H.R. Dowson and M.B. Ghaemi 1

Michael Taylor: The Schrödinger equation on spheres 145

Jean-Pierre Tignol with Grégory Berhuy and Marina Monsurrò 201

Cynthia E. Will: The Meromorphic continuation of the resolvent of the Laplacian on line bundles over $\mathbb{C}H(n)$ 157

Baozhong Yang: Removable singularities for Yang–Mills connections in higher dimensions 381

Fuliu Zhu: The heat kernel and the Riesz transforms on the quaternionic Heisenberg groups 175
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 Bib\TeX 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.