Pacific Journal of Mathematics

Volume 247 No. 1 September 2010

PACIFIC JOURNAL OF MATHEMATICS

http://www.pjmath.org

Founded in 1951 by
E. F. Beckenbach (1906–1982) and F. Wolf (1904–1989)

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

The subscription price for 2010 is US $420/year for the electronic version, and $485/year for print and electronic.
Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163, U.S.A. Prior back issues are obtainable from Periodicals Service Company, 11 Main Street, Germantown, NY 12526-5635. The Pacific Journal of Mathematics is indexed by Mathematical Reviews, Zentralblatt MATH, PASCAL CNRS Index, Referativnyi Zhurnal, Current Mathematical Publications and the Science Citation Index.

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

PJM peer review and production are managed by EditFlow™ from Mathematical Sciences Publishers.

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

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

CLASSIFICATION RESULTS FOR EASY QUANTUM GROUPS

TEODOR BANICA, STEPHEN CURRAN AND ROLAND SPEICHER

We study the orthogonal quantum groups satisfying the "easiness" assumption axiomatized in our previous paper, with the construction of some new examples and with some partial classification results. The conjectural conclusion is that the easy quantum groups consist of the previously known 14 examples, plus a hypothetical multiparameter "hyperoctahedral series", related to the complex reflection groups $H_n^s = \mathbb{Z}_s \wr S_n$. We also discuss the general structure and the computation of asymptotic laws of characters for the new quantum groups that we construct.

Introduction

One of the strengths of the theory of compact Lie groups is that these objects can be classified. It is indeed extremely useful to know that the symmetry group of a classical or a quantum mechanical system falls into an advanced classification machinery, and applications of this method abound in mathematics and physics.

The quantum groups were introduced by Drinfel'd [1987] and Jimbo [1985], in order to deal with quite complicated systems, basically coming from number theory or quantum mechanics, whose symmetry groups are not "classical". There are now available several extensions and generalizations of the Drinfel'd–Jimbo construction, all of them more or less motivated by the same philosophy. A brief account of the whole story, focusing on constructions that are of interest here, is as follows:

(1) Let $G \subset U_n$ be a compact group, and consider the algebra $A = C(G)$. The matrix coordinates $u_{ij} \in A$ satisfy the commutation relations $ab = ba$. The original idea of Drinfel'd and Jimbo, further processed by Woronowicz [1987], was that these commutation relations are in fact the $q = 1$ case of the q-commutation relations $ab = qba$, where $q > 0$ is a parameter. The algebra A itself appears then as the $q = 1$ case of an algebra A_q. While A_q is no longer commutative, we can formally write $A = C(G_q)$, where G_q is a quantum group.

MSC2000: primary 46L65; secondary 20F55, 46L54.
Keywords: quantum group, noncrossing partition.
Banica was supported by the ANR grants "Galoisint" and "Granma". Speicher was supported by a Discovery grant from NSERC.

(2) Wang [1995; 1998] proposed an interesting modification of this construction. His idea was to construct a new algebra A^+, by somehow "removing" the commutation relations $ab = ba$. Once again we can formally write $A^+ = C(G^+)$, where G^+ is a so-called free quantum group. This construction, while originally coming only with a vague motivation from mathematical physics, has been studied intensively in the last 15 years. Among the partial conclusions that we have so far is the fact that the combinatorics of G^+ is definitely interesting, and should have something to do with physics. In other words, G^+, while being by definition a quite abstract object, is probably the symmetry group of something very concrete.

(3) Several variations of Wang's construction appeared in recent years, notably in connection with the construction and classification of intermediate quantum groups $G \subset G^* \subset G^+$. For instance in the case $G = O_n$, it was shown in our previous paper [BS 2009] that the commutation relations $ab = ba$ can be successfully replaced with the so-called half-commutation relations $abc = cba$, in order to obtain a new quantum group O_n^*. Some other commutation-type relations, for instance of type $(ab)^s = (ba)^s$, will be described in the present paper.

(4) As a conclusion, the general idea that tends to emerge from these considerations is that a very large class of compact quantum groups should appear in the following way: start with a compact Lie group $G \subset U_n$; build a noncommutative version of $C(G)$ by replacing the commutation relations $ab = ba$ by some weaker relations; and deform this latter algebra, by using a positive parameter $q > 0$, or more generally a whole family of such positive parameters.

This was the motivating story. In practice, now, while the construction (1) is now basically understood, thanks to about 25 years of effort, (2) is just at the very beginning of an axiomatization, (3) is still at the level of pioneering examples, and (4) is just a dream. As for the possible applications to physics, basically nothing is known so far, but the hope for such an application increases as more and more interesting formulas emerge from the study of compact quantum groups.

This paper, a continuation of [BS 2009], will advance on the classification work there, for the easy quantum groups in the orthogonal case, and will present a detailed study of the new quantum groups we find.

The objects of interest will be the compact quantum groups with $S_n \subset G \subset O_n^+$. Here O_n^+ is the free analogue of the orthogonal group, constructed by Wang [1995], and for the compact quantum groups we use Woronowicz's formalism [1987].

As in [BS 2009] we restrict attention to the "easy" case. The easiness assumption, essential to our considerations, roughly states that the tensor category of G should be spanned by certain partitions, coming from the tensor category of S_n.

This might look like a quite technical condition. The point, though, is that imposing this technical condition is the price to pay for restricting attention to the "truly easy" case.

As explained in [BS 2009], our motivating belief is that "any result that holds for S_n, O_n should have a suitable extension to all easy quantum groups". This is of course a quite vague statement, whose target is actually informed by some results at the borderline between representation theory and probability. Here, however, we would rather focus on the classification problem. The further development of our "S_n, O_n philosophy", leading perhaps to some interesting applications, will be left to future papers. See Section 8 for more comments in this direction.

So, for the purposes of the present work, the easy quantum groups can be just thought of as being a carefully chosen collection of basic objects of the theory.

There are 14 natural examples of easy quantum groups, all but one described in [BS 2009], and the remaining one to be studied in detail in this paper. In addition, there are at least two infinite series, once again to be introduced here. The list is as follows:

(1) Groups: O_n, S_n, H_n, B_n, S'_n, B'_n.

(2) Free versions: O_n^+, S_n^+, H_n^+, B_n^+, $S_n'^+$, $B_n'^+$.

(3) Half liberations: O_n^*, H_n^*.

(4) Hyperoctahedral series: $H_n^{(s)}$, $H_n^{[s]}$.

This list doesn't cover all the easy quantum groups, but we will present here some partial classification results, with the conjectural conclusion that the full list should consist of (1)–(3), and of a multiparameter series unifying (4). We will also investigate the new quantum groups that we find, by using various techniques from [Banica et al. 2007a; 2007b; BS 2009; Banica and Vergnioux 2009a; 2009b].

As already mentioned, we expect the list above to be a useful, fundamental starting point for a number of representation theory and probability considerations. We also expect that the new quantum groups that we find this way will lead to some interesting applications. We have several projects here, to be discussed at the end of the paper.

The paper is organized as follows. In Sections 1 and 2 we recall our previous results from [BS 2009], and we study the quantum group H_n^* by using techniques from [BS 2009; Banica and Vergnioux 2009b]. In Sections 3 and 4, we introduce the one-parameter series, and we study their basic properties by using techniques from [Banica et al. 2007a; Banica and Vergnioux 2009a]. In Sections 5 and 6, we state and prove the classification results, by making heavy use of the capping method in [BS 2009; Banica and Vergnioux 2009b]. Sections 7 and 8 contain the computation of asymptotic laws of characters, and some concluding remarks.

Notation

As in [BS 2009], the basic object we consider will be a compact quantum group G. Concrete examples include the usual compact groups G and, to some extent, the duals of discrete groups $\widehat{\Gamma}$. In the general case, G is just a fictional object, which exists only via its associated Hopf C^*-algebra of "complex continuous functions", denoted $A = C(G)$.

For simplicity of notation, we would rather use the quantum group G instead of the Hopf algebra A. For instance $\int_G u_{i_1 j_1} \cdots u_{i_k j_k}\, du$ will denote the complex number obtained by applying the Haar functional $\varphi : A \to \mathbb{C}$ to the well-defined quantity $u_{i_1 j_1} \cdots u_{i_k j_k} \in A$.

We will use the quantum group notation depending on the setting; in cases where this can lead to confusion, we will switch back to the Hopf algebra notation.

1. Easy quantum groups

We briefly recall some notions and results from [BS 2009]. This material is here mostly for fixing the formalism and the notation.

Consider first a compact group satisfying $S_n \subset G \subset O_n$. That is, $G \subset O_n$ is a closed subgroup containing the subgroup $S_n \subset O_n$ formed by the permutation matrices.

Let u, v be the fundamental representations of G, S_n. By functoriality we have an inclusion $\mathrm{Hom}(u^{\otimes k}, u^{\otimes l}) \subset \mathrm{Hom}(v^{\otimes k}, v^{\otimes l})$ for any k and l. On the other hand, the Hom-spaces for v are well known: they are spanned by operators T_p, with p belonging to $P(k,l)$, the set of partitions between k points and l points. More precisely, if $e_1, \ldots, e_n$ denotes the standard basis of $\mathbb{C}^n$, the formula for T_p is

$$T_p(e_{i_1} \otimes \cdots \otimes e_{i_k}) = \sum_{j_1 \cdots j_l} \delta_p \begin{pmatrix} i_1 \cdots i_k \\ j_1 \cdots j_l \end{pmatrix} e_{j_1} \otimes \cdots \otimes e_{j_l} \tag{1-1}$$

Here the δ symbol on the right is 0 or 1, depending on whether the indices "fit" or not, that is, $\delta = 1$ if all blocks of p contain equal indices, and $\delta = 0$ if not.

Thus the space $\mathrm{Hom}(u^{\otimes k}, u^{\otimes l})$ consists of linear combinations of operators of type T_p with $p \in P(k,l)$.

We call G easy if its tensor category is spanned by partitions.

Definition 1.1. We say a compact group $S_n \subset G \subset O_n$ is *easy* if there exist sets $D(k,l) \subset P(k,l)$ such that $\mathrm{Hom}(u^{\otimes k}, u^{\otimes l}) = \mathrm{span}(T_p \mid p \in D(k,l))$ for any k, l.

It follows from the axioms of tensor categories that the collection of sets $D(k,l)$ must be closed under certain categorical operations, notably vertical and horizontal concatenation, and upside-down turning. The corresponding algebraic structure formed by the sets $D(k,l)$, axiomatized in [BS 2009], will be called *category of crossing partitions*.

We recall that a matrix is called monomial if it has exactly one nonzero entry on each row and each column. The basic examples are the permutation matrices.

Definition 1.2. We consider the following groups:

(1) O_n, the orthogonal group;

(2) S_n, the symmetric group, formed by the permutation matrices

(3) H_n, the hyperoctahedral group, formed by monomial matrices with ± 1 entries;

(4) B_n, the bistochastic group, formed by orthogonal matrices with sum 1 on each row;

(5) $S'_n = \mathbb{Z}_2 \times S_n$, the group formed by the permutation matrices times ± 1;

(6) $B'_n = \mathbb{Z}_2 \times B_n$, the group formed by the bistochastic matrices times ± 1.

It follows from definitions that all these groups satisfy $S_n \subset G \subset O_n$. Among all these groups, only O_n and S_n are "irreducible", because we have canonical isomorphisms $H_n = \mathbb{Z}_2 \wr S_n$ and $B_n \simeq O_{n-1}$. See [BS 2009].

The partitions in $P(k,l)$ with $k+l$ even are themselves called even.

Theorem 1.3 [BS 2009]. *The only easy groups are the ones in Definition 1.2, and the corresponding categories of crossing partitions are as follows*:

(1) P_o, *all pairings*;

(2) P_s, *all partitions*;

(3) P_h, *partitions with blocks of even size*;

(4) P_b, *singletons and pairings*;

(5) $P_{s'}$, *all partitions* (*even part*);

(6) $P_{b'}$, *singletons and pairings* (*even part*).

The second assertion follows from some well-known results about the groups O_n, S_n and their versions, and the first can be proved by carefully manipulating the categorical axioms.

We now discuss the free analogue of the above results. Let O_n^+ and S_n^+ be respectively the free orthogonal and symmetric quantum groups corresponding to the Hopf algebras $A_o(n)$ and $A_s(n)$ constructed by Wang [1995; 1998]. Here and in what follows, we use Woronowicz's Hopf algebra formalism [1987] and its subsequent quantum group interpretation.

We have $S_n \subset S_n^+$, so by functoriality the Hom-spaces for S_n^+ appear as subspaces of the corresponding Hom-spaces for S_n. The Hom-spaces for S_n^+ have in fact a very simple description. They are spanned by the operators T_p with $P \in \mathrm{NC}(k,l)$, the set of noncrossing partitions between k upper points and l lower points.

Definition 1.1. has a free analogue.

Definition 1.4. A compact quantum group $S_n^+ \subset G \subset O_n^+$ is called *free* if there exist sets $D(k,l) \subset \mathrm{NC}(k,l)$ such that $\mathrm{Hom}(u^{\otimes k}, u^{\otimes l}) = \mathrm{span}(T_p \mid p \in D(k,l))$ for any k, l.

In this definition, the word “free” has a quite subtle meaning, to be fully justified later on. Forn now, let us note that the passage from Definition 1.1 to Definition 1.4 is basically done by restricting attention to the noncrossing partitions, which, according to [Speicher 1994], should indeed lead to freeness.

As in the classical case, the sets of partitions $D(k,l)$ must be stable under certain categorical operations, coming this time from the axioms in [Woronowicz 1988]. The corresponding algebraic structure, axiomatized in [BS 2009], is called the category of noncrossing partitions.

We denote by H_n^+ the hyperoctahedral quantum group constructed in [Banica et al. 2007b], and by B_n^+, $S_n'^+$ and $B_n'^+$ the free analogues of the groups B_n, S_n' and B_n' constructed in [BS 2009].

Definition 1.5. We consider the following quantum groups, all given with the defining relations between the basic coordinates $u_{ij} \in C(G)$:

(1) O_n^+, orthogonality ($u_{ij} = u_{ij}^*$ and $u^t = u^{-1}$);

(2) S_n^+, magic condition (all rows and columns of u are partitions of unity);

(3) H_n^+, cubic condition (orthogonality and $u_{ij}u_{ik} = u_{ji}u_{ki} = 0$ for $j \neq k$);

(4) B_n^+, bistochastic condition (orthogonality and on each row the sum is 1);

(5) $S_n'^+$, cubic condition, with the same sum on rows and columns;

(6) $B_n'^+$, orthogonality, with the same sum on rows and columns;

Perhaps the very first observation is that for any of the groups G appearing in Definition 1.2 we have $C(G) = C(G^+)/I$, where $I \subset C(G^+)$ is the commutator ideal. In other words, G^+ is indeed a noncommutative version of G. We refer to [BS 2009] and to its predecessors [Banica et al. 2007b; Wang 1995; 1998] for the whole story, and for a careful treatment of all this material.

The free analogue of Theorem 1.3 is this:

Theorem 1.6 [BS 2009]. *Definition 1.5 lists the only free quantum groups. The corresponding categories of noncrossing partitions are as follows*:

(1) NC_o, *all noncrossing pairings*;

(2) NC_s, *all noncrossing partitions*;

(3) NC_h, *noncrossing partitions with blocks of even size*;

(4) NC_b, *singletons and noncrossing pairings*;

(5) $\mathrm{NC}_{s'}$, *all noncrossing partitions* (*even part*);

(6) $\mathrm{NC}_{b'}$, *singletons and noncrossing pairings* (*even part*).

The proof of this theorem follows that of Theorem 1.3. The symmetry between Theorems 1.3 and 1.6 corresponds to the liberation operation for orthogonal Lie groups, further investigated in [BS 2009].

2. Half-liberation

We consider now the general situation where we have a compact quantum group satisfying $S_n \subset G \subset O_n^+$. Once again, we can ask for the tensor category of G to be spanned by certain partitions, coming from the tensor category of S_n.

Definition 2.1. A compact quantum group $S_n \subset G \subset O_n^+$ is called easy if there exist sets $D(k,l) \subset P(k,l)$ such that $\mathrm{Hom}(u^{\otimes k}, u^{\otimes l}) = \mathrm{span}(T_p \mid p \in D(k,l))$ for any k, l.

This definition generalizes at the same time Definitions 1.1 and 1.4. Indeed, the easy quantum groups $S_n \subset G \subset O_n^+$ satisfying the extra assumption $G \subset O_n$ are the easy groups, and those satisfying the extra assumption $S_n^+ \subset G$ are the free quantum groups. This follows from definitions; see [BS 2009].

Once again, the sets of partitions $D(k,l)$ must be stable under certain categorical operations coming from the axioms in [Woronowicz 1988]. The corresponding algebraic structure, axiomatized in [BS 2009], will be called simply "category of partitions".

We already know that the easy quantum groups include the 6 easy groups and the 6 free quantum groups. In general, the world of easy quantum groups is quite rigid, but we can produce some more examples in the following way.

Definition 2.2. The half-liberated version of an easy group G is the quantum group G^* given by $C(G^*) = C(G^+)/I$, where I is the ideal generated by the half-commutation relations $abc = cba$, imposed on the basic matrix coordinates $u_{ij} \in C(G^+)$.

In other words, instead of removing the commutativity relations of type $ab = ba$ from the standard presentation of $C(G)$, which would produce the algebra $C(G^+)$, we replace these commutativity relations by the weaker relations $abc = cba$.

To study the half-liberated versions, we need a categorical interpretation of the relations $abc = cba$. Let us agree that the upper points of a partition $p \in P(k,l)$ are labeled $1, 2, \dots, k$, and the lower points are labeled $1', 2', \dots, l'$.

Lemma 2.3 [BS 2009]. *For a compact quantum group $G \subset O_n^+$, the following are equivalent*:

(1) *The basic coordinates u_{ij} satisfy $abc = cba$.*

(2) T_p *belongs to* $\mathrm{End}(u^{\otimes 3})$, *where* $p = (13')(22')(3'1)$.

Proof. By the definition (1-1) of T_p, we have $T_p(e_a \otimes e_b \otimes e_c) = e_c \otimes e_b \otimes e_a$. This gives the formulas

$$T_p u^{\otimes 3}(e_a \otimes e_b \otimes e_c) = \sum_{ijk} e_k \otimes e_j \otimes e_i \otimes u_{ia} u_{jb} u_{kc}$$

$$u^{\otimes 3} T_p(e_a \otimes e_b \otimes e_c) = \sum_{ijk} e_i \otimes e_j \otimes e_k \otimes u_{ic} u_{jb} u_{ka}$$

The identification of the right terms gives the equivalence in the statement. □

We now go back to the quantum groups G^*. Observe first that we have inclusions $G \subset G^* \subset G^+$. As pointed out in [BS 2009], the cases $G = S_n, B_n, S'_n, B'_n$ are not interesting, because here we have $G = G^*$. This can be checked by a direct computation with generators and relations, or with the partition p appearing in Lemma 2.3, and will follow as well from the general classification results in Sections 5 and 6.

In the cases $G = O_n, H_n$, however, we obtain new quantum groups. Label the legs of each partition by $1, 2, 3, \dots,$ clockwise starting from top left.

Theorem 2.4. *The half-liberated versions of O_n and H_n are easy quantum groups, and the corresponding categories of partitions are*

(1) E_o, *pairings with each string connecting an odd number to an even number*;

(2) E_h, *partitions with each block having the same number of odd and even legs.*

Proof. Our claim is that E_o and E_h are categories of partitions, corresponding respectively to the quantum groups O_n^* and H_n^*.

(1) Here E_o is nothing but the set of pairings with each string having an even number of crossings, and the result was proved in [BS 2009]. The idea is that E_o is generated in the categorical sense by the partition p appearing in Lemma 2.3.

(2) The fact that E_h is indeed a category of partitions follows from definitions. Thinking of each block as being "balanced" with respect to the odd and even labels, we see that the categorical operations preserve the balancing. For instance when checking the stability under composition, which is the crucial axiom, we see that given a connected union of blocks of the two partitions that are composed, the "balancing in the middle" is subject to canceling.

The fact that E_h corresponds to the above quantum group H_n^* can be checked in several ways. Consider for instance the diagram

$$\begin{array}{ccc} O_n^* & \subset & O_n^+ \\ \cup & & \cup \\ H_n^* & \subset & H_n^+. \end{array}$$

We know from definitions that H_n^* is obtained by putting together the relations for O_n^* and for H_n, so we have the quantum group equality $H_n^* = O_n^* \cap H_n^+$. Now by the general properties of Tannakian duality, it follows that the category of partitions of H_n^* is generated by the category of partitions for H_n^+, namely the noncrossing partitions having even blocks, and by the half-liberation partition p in Lemma 2.3.

This category is by definition included into E_h, and the reverse inclusion can be checked as well by a straightforward computation. □

The quantum group O_n^*, appearing first in [BS 2009], was further investigated in [Banica and Vergnioux 2009b]. To get some insight into the structure of H_n^*, we will use similar methods.

Definition 2.5. The projective version of a quantum group $G \subset U_n^+$ is the quantum group $PG \subset U_{n^2}^+$, having as basic coordinates the elements $v_{ij,kl} = u_{ik}u_{jl}^*$.

In other words, $C(PG) \subset C(G)$ is the algebra generated by the elements $v_{ij,kl} = u_{ik}u_{jl}^*$. In the case where G is a classical group we recover the well-known formula $PG = G/(G \cap T)$, where $T \subset U_n$ are the unitary diagonal matrices. We refer to [Banica and Vergnioux 2009b] for a full discussion and a list of concrete examples.

Consider now the compact group $K_n = \mathbb{T} \wr S_n$ consisting of monomial (that is, permutation-like) matrices, with elements on the unit circle $\mathbb{T}$ as nonzero entries.

The next result, whose first claim is from [Banica and Vergnioux 2009b], will play a key role in the study of H_n^* and the other quantum groups introduced here.

Theorem 2.6. *The projective versions of half-liberations are as*

(1) $PO_n^* = PU_n$, *and*

(2) $PH_n^* = PK_n$.

Proof. The first claim is proved using that the partitions for PO_n^* and PU_n are the same. For the second, we use a similar method. Observe first that from $H_n^* \subset O_n^*$, we get $PH_n^* \subset PO_n^* = PU_n$, so PH_n^* is indeed a classical group.

To compute this group, consider the diagram

$$\begin{tikzcd} K_n \arrow[r, phantom, "\subset"] & U_n^+ \\ H_n \arrow[u, phantom, "\cup"] \arrow[r, phantom, "\subset"] & H_n^*. \arrow[u, phantom, "\cup"] \end{tikzcd}$$

We fix $k, l \geq 0$ and consider the formal words $\alpha = (u \otimes \bar{u})^{\otimes k}$ and $\beta = (u \otimes \bar{u})^{\otimes l}$. Our claim is that the corresponding spaces $\mathrm{Hom}(\alpha, \beta)$ for our 4 quantum groups appear as span of the operators T_p, with p belonging to the following 4 sets of partitions:

$$\begin{tikzcd} E_h(2k,2l) \arrow[r, phantom, "\supset"] & E_o(2k,2l) \\ P_h(2k,2l) \arrow[u, phantom, "\cup"] \arrow[r, phantom, "\supset"] & E_h(2k,2l). \arrow[u, phantom, "\cup"] \end{tikzcd}$$

Indeed, the bottom left set is a good one, by Theorem 1.3. The bottom right set is also a good one, by Theorem 2.4. For the top right set, this follows from the equality $PO_n^* = PU_n$ and from Theorem 2.4, and for full details see [Banica and Vergnioux 2009b]. As for the top left set, this follows for instance from the various results in [Banica et al. 2007a; Banica 2008; Banica and Vergnioux 2009a] regarding K_n^+, after "adding a crossing". A direct proof can be obtained as well, by working out the categorical interpretation of the various relations defining K_n.

In summary, we have computed the relevant diagrams for the projective versions of our four algebras. So, let us look now at these projective versions:

$$\begin{array}{ccc} PK_n & \subset & PU_n^+ \\ \cup & & \cup \\ PH_n & \subset & PH_n^* . \end{array}$$

The quantum groups PH_n^* and PK_n appear as subgroups of the same quantum group, namely PU_n^+, and the discussion above tells us that these subgroups have the same diagrams. The same argument of [Banica and Vergnioux 2009b] tells us that $PH_n^* = PK_n$. □

3. The hyperoctahedral series

We now introduce a new series of quantum groups, $H_n^{(s)}$ with $s \in \{2, 3, \dots, \infty\}$. These will intermediate between $H_n^{(2)} = H_n$ and $H_n^{(\infty)} = H_n^*$.

The quantum group $H_n^{(s)}$ is obtained from H_n^* by imposing the s-commutation condition $abab\cdots = baba\cdots$ (words of length s) on the basic coordinates u_{ij}.

Definition 3.1. $C(H_n^{(s)})$ is the universal C^*-algebra generated by n^2 self-adjoint variables u_{ij}, subject to the relations

(1) (orthogonality) $uu^t = u^t u = 1$, where $u = (u_{ij})$ and $u^t = (u_{ji})$;

(2) (cubic relations) $u_{ij}u_{ik} = u_{ji}u_{ki} = 0$ for any i and any $j \neq k$;

(3) (half-commutation) $abc = cba$ for any $a, b, c \in \{u_{ij}\}$;

(4) (s-mixing relation) $abab\cdots = baba\cdots$ (length s words) for any $a, b \in \{u_{ij}\}$.

That $H_n^{(s)}$ is a quantum group follows from the elementary fact that the cubic relations are of Hopf type, that is, they allow the construction of the Hopf algebra maps Δ, ε, S. This can be checked by a routine computation.

At $s = 2$, the s-mixing is the usual commutation $ab = ba$. Since this relation is stronger than the half-commutation $abc = cba$, we are led to the algebra generated by n^2 commuting self-adjoint variables satisfying (1) and (2), which is $C(H_n)$.

In the case $s = \infty$, the s-mixing relation disappears by definition. Thus we are led to the algebra defined by the relations (1)–(3), which is $C(H_n^*)$.

Summarizing, we have $H_n^{(2)} = H_n$ and $H_n^{(\infty)} = H_n^*$, as previously claimed. In what follows we present a detailed study of $H_n^{(s)}$.

Lemma 3.2. *For a compact quantum group $G \subset H_n^*$, the following are equivalent*:

(1) *The basic coordinates u_{ij} satisfy $abab\cdots = baba\cdots$ (length s words).*

(2) *T_p belongs to* $\mathrm{End}(u^{\otimes s})$, *where $p = (135\cdots 2'4'6'\cdots)(246\cdots 1'3'5'\cdots)$.*

Proof. According to the definition of T_p given in (1-1), the operator associated to the partition in the statement is given by the formula

$$T_p(e_{a_1}\otimes e_{b_1}\otimes e_{a_2}\otimes e_{b_2}\otimes \cdots) = \delta(a)\delta(b)e_b\otimes e_a\otimes e_b\otimes e_a\otimes \cdots.$$

Here we use the convention $\delta(a) = 1$ if all the indices a_i are equal and $\delta(a) = 0$ if not, along with a similar convention for $\delta(b)$. The indices a and b appearing on the right are the common values of the a indices and b indices, respectively, in the case $\delta(a) = \delta(b) = 1$, and are irrelevant quantities in the remaining cases.

This gives the formulas

$$T_p u^{\otimes s}(e_{a_1}\otimes e_{b_1}\otimes e_{a_2}\otimes\cdots) = \sum_{ij} e_i\otimes e_j\otimes e_i\otimes\cdots\otimes u_{ia_1}u_{jb_1}u_{ia_2}\cdots,$$

$$u^{\otimes s}T_p(e_{a_1}\otimes e_{b_1}\otimes e_{a_2}\otimes\cdots) = \delta(a)\delta(b)\sum_{ij} e_{i_1}\otimes e_{j_1}\otimes e_{i_2}\otimes\cdots\otimes u_{i_1b}u_{j_1a}u_{i_2b}\cdots.$$

Here the upper sum is over all indices i and j, and the lower sum is over all multiindices $i = (i_1,\dots,i_s)$ and $j = (j_1,\dots,j_s)$. The identification of the terms on the right, after a suitable relabeling of indices, gives the equivalence in the statement. □

We now work out the s-analogue of Theorem 2.4.

Theorem 3.3. *$H_n^{(s)}$ is an easy quantum group, and its associated category E_h^s is that of the s-*balanced *partitions, that is, partitions satisfying the conditions that*

(1) *the total number of legs is even, and*

(2) *in each block, the number of odd and even legs are equal, modulo s.*

Proof. At $s = 2$ the first condition implies the second one, so here we simply get the partitions having an even number of legs, corresponding to H_n. At $s = \infty$ we get the partitions that are balanced, in the sense of the proof of Theorem 2.4, which are known from there to correspond to the quantum group H_n^*.

We first claim that E_h^s is a category. This follows simply by adapting the $s = \infty$ argument in the proof of Theorem 2.4, just by adding “modulo s” everywhere.

It remains to prove that this category corresponds to $H_n^{(s)}$. This follows from the fact that the partition p of Lemma 3.2 generates the category of s-balanced partitions, as one can check by a routine computation. □

Consider now the complex reflection group $H_n^s = \mathbb{Z}_s \wr S_n$, consisting of monomial matrices having the s-roots of unity as nonzero entries. Observe that we have $PH_n^{(s)} = H_n^s/\mathbb{T}$.

We have the following s-analogue of Theorem 2.6.

Theorem 3.4. $PH_n^{(s)} = PH_n^s$.

Proof. This statement holds at $s = 2$, because here we have $H_n^{(2)} = H_n^2 = H_n$; it holds at $s = \infty$ due to Theorem 2.6.

In the general case, it follows by adapting the proof of Theorem 2.6. Observe first that from $H_n^{(s)} \subset H_n^*$ we get $PH_n^{(s)} \subset PH_n^* = PK_n$, so $PH_n^{(s)}$ is a classical group.

To compute this group, consider the diagram

$$\begin{array}{ccc} H_n^s & \subset & U_n^+ \\ \cup & & \cup \\ S_n & \subset & H_n^{(s)}. \end{array}$$

The corresponding sets of partitions, as in the proof of Theorem 2.6, are

$$\begin{array}{ccc} E_h^s(2k,2l) & \supset & E_o(2k,2l) \\ \cup & & \cup \\ P(2k,2l) & \supset & E_h^s(2k,2l). \end{array}$$

The bottom left set is a good one, by Theorem 1.3, as is bottom right one, by Theorem 3.3. For the top right set, this was already discussed in the proof of Theorem 2.6. For the top left set, this follows either from the results in [Banica et al. 2007a; Banica and Vergnioux 2009a] regarding the free version H_n^{s+}, after adding a crossing, or from the $s = \infty$ computation in the proof of Theorem 2.6. A direct proof can be obtained as well.

We now look at the projective versions of the above quantum groups:

$$\begin{array}{ccc} PH_n^s & \subset & PU_n^+ \\ \cup & & \cup \\ PH_n & \subset & PH_n^{(s)}. \end{array}$$

As in the proof of Theorem 2.6, we have two quantum subgroups having the same diagrams, and we conclude that $PH_n^{(s)} = PH_n^s$. □

4. The higher hyperoctahedral series

We introduce a second one-parameter series of quantum groups, $H_n^{[s]}$ with s in $\{2, 3, \dots, \infty\}$, having as main particular case the group $H_n^{[2]} = H_n$.

Definition 4.1. $C(H_n^{[s]})$ is the universal C^*-algebra generated by n^2 self-adjoint variables u_{ij}, subject to the relations

(1) (orthogonality) $uu^t = u^t u = 1$, where $u = (u_{ij})$ and $u^t = (u_{ji})$.

(2) (ultracubic relations) $acb = 0$ for any $a \neq b$ on the same row or column of u.

(3) (s-mixing relation) $abab \cdots = baba \cdots$ (length s words) for any $a, b \in \{u_{ij}\}$.

That $H_n^{[s]}$ is a quantum group follows from the elementary fact that the ultracubic relations are of "Hopf type", that is, that they allow the construction of the Hopf algebra maps Δ, ε and S. This can be checked by a routine computation.

We first compare the defining relations for $H_n^{[s]}$ with those for $H_n^{(s)}$. To deal at the same time with the cubic and ultracubic relations, it is convenient to use a statement about a unifying notion, *k-cubic* relations.

Lemma 4.2. *For a compact quantum group $G \subset O_n^+$, the following are equivalent*:

(1) *The basic coordinates u_{ij} satisfy the k-cubic relations $ac_1 \cdots c_k b = 0$ for any $a \neq b$ on the same row or column of u, and for any $c_1, \dots, c_k$.*

(2) $T_p \in \mathrm{End}(u^{\otimes k+2})$, *where* $p{=}(1, 1', k{+}2, k{+}2')(2, 2')\cdots(k{+}1, k{+}1')$.

Proof. According to (1-1), the operator associated to the partition in the statement is given by

$$T_p(e_a \otimes e_{c_1} \otimes \cdots \otimes e_{c_k} \otimes e_b) = \delta_{ab} e_a \otimes e_{c_1} \otimes \cdots \otimes e_{c_k} \otimes e_a.$$

This gives the formulas

$$\begin{aligned}
&T_p u^{\otimes k+2}(e_a \otimes e_{c_1} \otimes \cdots \otimes e_{c_k} \otimes e_b)\\
&\qquad = \sum_{ij} e_i \otimes e_{j_1} \otimes \cdots \otimes e_{j_k} \otimes e_i \otimes u_{ia} u_{j_1 c_1} \cdots u_{j_k c_k} u_{ib},\\
&u^{\otimes k+2} T_p(e_a \otimes e_{c_1} \otimes \cdots \otimes e_{c_k} \otimes e_b\\
&\qquad = \delta_{ab} \sum_{ijl} e_i \otimes e_{j_1} \otimes \cdots \otimes e_{j_k} \otimes e_l \otimes u_{ia} u_{j_1 c_1} \cdots u_{j_k c_k} u_{la},
\end{aligned}$$

Here the sums are over all indices i and l, and over all multiindices $j = (j_1, \dots, j_k)$. The identification of the terms on the right gives the desired equivalence. □

We can now establish the precise relationship between $H_n^{[s]}$ and $H_n^{(s)}$ and also show that no further series can appear in this way.

Proposition 4.3. *For $k \geq 1$, the k-cubic relations are all equivalent to the ultracubic relations, and they imply the cubic relations.*

Proof. This follows from two observations.

First, the k-cubic relations imply the $2k$-cubic relations. Indeed, one can connect two copies of the partition p in Lemma 4.2, by gluing them with two semicircles in the middle, and the resulting partition is the one implementing the $2k$-cubic relations.

Second, the k-cubic relations imply the $(k-1)$-cubic relations. By capping the partition p in Lemma 4.2 with a semicircle at bottom right, we get a partition $p' \in P(k+2,k)$, and by rotating the upper right leg of this partition we get the partition $p'' \in P(k+1,k+1)$ implementing the $(k-1)$-cubic relations. □

Proposition 4.3 shows that replacing in Definition 4.1 the ultracubic condition by any of the k-cubic conditions with $k \geq 2$ won't change the resulting quantum group. The other consequences of Proposition 4.3 are summarized as follows.

Proposition 4.4. *The quantum groups $H_n^{[s]}$ have the properties that*

(1) $H_n^{(s)} \subset H_n^{[s]} \subset H_n^+$;

(2) $H_n^{[2]} = H_n^{(s)} = H_n$ *at* $s = 2$;

(3) $H_n^{(s)} \neq H_n^{[s]}$ *at* $s \geq 3$.

Proof. All the assertions basically follow from Lemma 4.2:

(1) For the first inclusion, we need to show half-commutation plus cubic implies ultracubic; this can be done by placing the half-commutation partition next to the cubic partition, then using 2 semicircle cappings in the middle. The second inclusion follows from Proposition 4.3, because the ultracubic relations (1-cubic relations) imply the cubic relations (0-cubic relations).

(2) At $s = 2$ the s-commutation is the usual commutation $ab = ba$. Thus we are led here to the algebra generated by n^2 commuting self-adjoint variables satisfying the cubic condition, which is $C(H_n)$.

(3) Finally, $H_n^{(s)} \neq H_n^{[s]}$ will be a consequence of Theorem 4.5 below, because at $s \geq 3$ the half-commutation partition $p = (14)(25)(36)$ is s-balanced but not locally s-balanced. □

Theorem 4.5. *$H_n^{[s]}$ is an easy quantum group, and its associated category is that of the* locally s-balanced *partitions, that is, partitions having the property that each of their subpartitions* (*that is, partitions obtained by removing certain blocks*) *are s-balanced.*

Proof. At $s = 2$ the locally s-balancing condition is automatic for a partition having blocks of even size, so we get indeed the category corresponding to H_n.

In the general case, we first claim that the locally s-balanced partitions form a category. This follows simply by adapting the proof of Theorem 3.3, just by adding "locally" everywhere.

It remains to prove that this category corresponds to $H_n^{[s]}$. This follows from Lemma 3.2 and from the fact that the partition generating the category of locally balanced partitions, namely, $p = (1346)(25)$, is nothing but the one implementing the ultracubic relations, as one can check by a routine computation. □

5. Classification: General strategy

In this section and the next we advance the classification work started in [BS 2009]. We will prove that the easy quantum groups constructed so far are the only ones, modulo a hypothetical multiparameter "hyperoctahedral series", unifying the series constructed in the previous sections, and still waiting to be constructed.

Let G be an easy quantum group with category of partitions denoted P_g. It follows from definitions that $P_g \cap \mathrm{NC}$ is a category of noncrossing partitions; by the results in Section 1, this latter category must come from a free quantum group K^+. Since $\mathrm{NC}_k = P_g \cap \mathrm{NC}$ is included into P_g, we have $G \subset K^+$.

Definition 5.1. Associated to an easy quantum group G is the easy group K given by the equality of categories $P_g \cap \mathrm{NC} = \mathrm{NC}_k$.

According to the easy group classification in Theorem 1.3, there are six cases to be studied; five of these will be studied in Section 6, and the remaining case, $K = H_n$, will be left open.

The reason these cases are separated comes from the question, Do we have $K \subset G$? In the reminder of this section we will try to answer this question.

We begin with the technical lemma, valid in the general case. Let $\Lambda_g, \Lambda_k \subset \mathbb{N}$ be the set of the possible sizes of blocks of elements of P_g, NC_k.

Lemma 5.2. *Let G and K be as above.*

(1) $\Lambda_k \subset \Lambda_g \subset \Lambda_k \cup (\Lambda_k - 1)$.

(2) $1 \in \Lambda_g$ *implies* $1 \in \Lambda_k$.

(3) *If* NC_k *is even, so is* P_g.

Proof. We will heavily use the various abstract notions and results in [BS 2009].

(1) The first inclusion follows from $\mathrm{NC}_k \subset P_g$. The second is equivalent to the statement, "If b is a block of a partition $p \in P_g$, then there exists a certain block b' of a certain partition $p' \in P_g \cap \mathrm{NC}$, having size $\#b$ or $\#b - 1$." This then follows by using the capping method in [BS 2009]. We can cap p with semicircles, so that b remains unchanged, and we end up with a partition p' consisting of b and some extra points, at most one point between any two legs of b, which may or may not be connected. Since the semicircle capping is a categorical operation, this partition p' remains in P_g.

Now by further capping p' with semicircles, so as to get rid of the extra points, the size of b can only increase, and we end up with a one-block partition having size at least that of b. This one-block partition is obviously noncrossing, and by capping it again with semicircles we can reduce the number of legs up to $\#b$ or $\#b - 1$, and we are done.

(2) The condition $1 \in \Lambda_g$ means that there exists $p \in P_g$ having a singleton. By capping p with semicircles outside this singleton, we can obtain a singleton or a double singleton. Since both these partitions are noncrossing and have a singleton, we are done.

(3) Assume that P_g is not even, and consider a partition $p \in P_g$ having an odd number of legs. By capping p with enough semicircles we ensure ending up with a singleton, and since this singleton is by definition in $P_g \cap \mathrm{NC}$, we are done. □

We are now in position of splitting the classification.

Proposition 5.3. *Let G, K be as above.*

(1) *If $K \neq H_n$, then $K \subset G \subset K^+$.*

(2) *If $K = H_n$, then $S_n' \subset G \subset H_n^+$.*

Proof. Recall that the inclusion $G \subset K^+$ follows from definitions. For the other inclusion, we have 6 cases, depending on the exact value of the easy group K:

(1a) $K = O_n$. Here $\Lambda_k = \{2\}$, so by Lemma 5.2(1) we get $\{2\} \subset \Lambda_g \subset \{1,2\}$. Moreover, from Lemma 5.2(2), we get $\Lambda_g = \{2\}$. Thus $P_g \subset P_o$, which gives $O_n \subset G$.

(1b) $K = S_n$. Here there is nothing to prove, since $S_n \subset G$ by definition.

(1c) $K = B_n$. Here $\Lambda_k = \{1,2\}$, so by Lemma 5.2(1) we get $\Lambda_g = \{1,2\}$. Thus we have $P_g \subset P_b$, which gives $B_n \subset G$.

(1d) $K = S_n'$. Here $P_g \subset P_s$ by definition, and by using Lemma 5.2(3) we deduce that $P_g \subset P_{s'}$, which gives $S_n' \subset G$.

(1e) $K = B_n'$. Here $\Lambda = \{1,2\}$, so by Lemma 5.2(1) we get $\Lambda_g = \{1,2\}$. This gives $P_g \subset P_b$, and by Lemma 5.2(3), we get $P_g \subset P_{b'}$, which gives $B_n' \subset G$.

(2) $K = H_n$. Here $P_g \subset P_s$ by definition, and by using Lemma 5.2(3) we deduce that $P_g \subset P_{s'}$, which gives $S_n' \subset G$. □

With a little more care, one can prove that the easy group K in statement (1) is nothing but the classical version of G, obtained as dual object to the commutative Hopf algebra $C(G)/I$, where $I \subset C(G)$ is the commutator ideal.

Statement (2) cannot be improved. The reason is that for the quantum group $H_n^{(s)}$ with s odd, we have $K = H_n$, and $K \not\subset G$.

6. The nonhyperoctahedral case

We classify the easy quantum groups, under the nonhyperoctahedral assumption $K \neq H_n$. Here K is as usual the easy group from Definition 5.1.

We know from Proposition 5.3 that our easy quantum group G appears as an intermediate quantum group, $K \subset G \subset K^+$. To classify these intermediate quantum groups, we use the method in [Banica and Vergnioux 2009b], where the problem was solved in the case $G = O_n$. For uniformity, we will also include this case.

Definition 6.1. Let $p \in P(k, l)$ be a partition, with the points counted modulo $k+l$ counterclockwise starting from bottom left.

(1) We call *semicircle capping* of p any partition obtained from p by connecting with a semicircle a pair of consecutive neighbors.

(2) We call *singleton capping* of p any partition obtained from p by capping one of its legs with a singleton.

(3) We call *doubleton capping* of p any partition obtained from p by capping two of its legs with singletons.

The semicircle, singleton and doubleton cappings are elementary operations on partitions that lower the total number of legs by 2, 1 and 2 respectively. There are $k+l$ possibilities for placing the semicircle or the singleton, and $(k+l)(k+l-1)/2$ possibilities for placing the double singleton. In the case of semicircle cappings at left or at right, the semicircle in question is in fact a vertical bar, but we will still call it semicircle.

The various cappings of p will be generically denoted p'.

Consider now the $5+5+1 = 11$ categories of partitions P_x, NC_x, E_x, with $x = o, s, b, s', b'$ described in Sections 1 and 2.

Lemma 6.2. *Let p be a partition having j legs.*

(1) *If $p \in P_o - E_o$ and $j > 4$, there exists a semicircle capping $p' \in P_o - E_o$.*

(2) *If $p \in E_o - \mathrm{NC}_o$ and $j > 6$, there exists a semicircle capping $p' \in E_o - \mathrm{NC}_o$.*

(3) *If $p \in P_s - \mathrm{NC}_s$ and $j > 4$, there exists a singleton capping $p' \in P_s - \mathrm{NC}_s$.*

(4) *If $p \in P_b - \mathrm{NC}_b$ and $j > 4$, there exists a singleton capping $p' \in P_b - \mathrm{NC}_b$.*

(5) *If $p \in P_{s'} - \mathrm{NC}_{s'}$ and $j > 4$, there exists a doubleton capping $p' \in P_{s'} - \mathrm{NC}_{s'}$.*

(6) *If $p \in P_{b'} - \mathrm{NC}_{b'}$ and $j > 4$, there exists a doubleton capping $p' \in P_{b'} - \mathrm{NC}_{b'}$.*

Proof. We write $p \in P(k, l)$, so that the number of legs is $j = k + l$. In the cases where our partition is a pairing, we use as well the number of strings, $s = j/2$.

Let us agree that all partitions are drawn to have a minimal number of crossings.

We use the same idea for all the proofs, namely to isolate a block of p having a crossing, or an odd number of crossings, then to cap p as in the statement, so this block remains crossing, or with an odd number of crossings. Here we use the observation that the balancing condition that defines the categories E_o and E_h can be interpreted as saying that each block has an even number of crossings when the picture of the partition is drawn so that this number of crossings is minimal.

(1) The assumption $p \notin E_o$ means that p has strings having an odd number of crossings. We fix such a string, and we try to cap p so that this string remains odd in the resulting partition p'. An examination of all possible pictures shows that this is possible, provided that our partition has $s > 2$ strings.

(2) The assumption $p \notin \mathrm{NC}_o$ means that p has crossing strings. We fix such a pair of strings, and we try to cap p so these strings remain crossing in p'. Once again, looking at all possible pictures shows that this is possible, provided that our partition has $s > 3$ strings.

(3) Since p is crossing, we can choose two of its blocks that are intersecting. If there are some other blocks left, we can cap one of their legs with a singleton, and we are done. If not, this means that our two blocks have a total of $j' \geq j > 4$ legs, so at least one of them has $j'' > 2$ legs. One of these j'' legs can always be capped with a singleton, so the capped partition remains crossing, and we are done.

(4) Here we can simply cap with a singleton, as in (3).

(5)–(6) Here we can cap with a doubleton, by proceeding twice as in (3). □

For a collection of subsets $X(k,l) \subset P(k,l)$ we denote by $\langle X \rangle \subset P$ the category of partitions generated by X. In other words, the elements of $\langle X \rangle$ come from those of X via the categorical operations for the categories of partitions, which are the vertical and horizontal concatenation and the upside-down turning. See [BS 2009].

Lemma 6.3. *Let p be a partition.*

(1) *If* $p \in P_o - E_o$, *then* $\langle p, \mathrm{NC}_o \rangle = P_o$.

(2) *If* $p \in E_o - \mathrm{NC}_o$, *then* $\langle p, \mathrm{NC}_o \rangle = E_o$.

(3) *If* $p \in P_s - \mathrm{NC}_s$, *then* $\langle p, \mathrm{NC}_s \rangle = P_s$.

(4) *If* $p \in P_b - \mathrm{NC}_b$, *then* $\langle p, \mathrm{NC}_b \rangle = P_b$.

(5) *If* $p \in P_{s'} - \mathrm{NC}_{s'}$, *then* $\langle p, \mathrm{NC}_{s'} \rangle = P_{s'}$.

(6) *If* $p \in P_{b'} - \mathrm{NC}_{b'}$, *then* $\langle p, \mathrm{NC}_{b'} \rangle = P_{b'}$.

Proof. We use Lemma 6.2 and the observation that the "capping partition" appearing there is always in the good category.

That is, we use that the semicircle is in NC_o, $\mathrm{NC}_{s'}$, the singleton is in NC_s, NC_b, and the doubleton is in $\mathrm{NC}_{b'}$. This observation tells us that, in each of the cases under consideration, the category to be computed can only decrease when replacing p by one of its cappings p'. For the singleton and doubleton cappings this is clear from definitions; for the semicircle capping this is also clear from definitions, except in the case where the capping semicircle is actually a bar added at left or at right, where we can use a categorical rotation operation as in [BS 2009].

(1)–(2) These claims can be proved by recurrence on the number $s = (k+l)/2$ of strings. Indeed, by using Lemma 6.2(1)–(2), for $s > 3$ we have a descent procedure $s \to s-1$, and this leads to the situation $s \in \{1, 2, 3\}$, where the statement is clear.

(3) We can proceed by recurrence on the number of legs of p. If the number of legs is $j = 4$, then p is a basic crossing, and we have $\langle p \rangle = P_s$. If the number of legs is $j > 4$ we can apply Lemma 6.2(3), and the result follows from $\langle p \rangle \supset \langle p' \rangle = P_s$.

(4)–(6) This is similar to the proof of (1)–(2), by using Lemma 6.2(4)–(6). □

Lemma 6.4. *Let p be a partition.*

1. *If* $p \in P_o$, *then* $\langle p, \mathrm{NC}_o \rangle \in \{P_o, E_o, \mathrm{NC}_o\}$.
2. *If* $p \in P_s$, *then* $\langle p, \mathrm{NC}_s \rangle \in \{P_s, \mathrm{NC}_s\}$.
3. *If* $p \in P_b$, *then* $\langle p, \mathrm{NC}_b \rangle \in \{P_b, \mathrm{NC}_b\}$.
4. *If* $p \in P_{s'}$, *then* $\langle p, \mathrm{NC}_{s'} \rangle \in \{P_{s'}, \mathrm{NC}_{s'}\}$.
5. *If* $p \in P_{b'}$, *then* $\langle p, \mathrm{NC}_{b'} \rangle \in \{P_{b'}, \mathrm{NC}_{b'}\}$.

Proof. This follows by rearranging the results in Lemma 6.3. □

We may now state our main result. We call *nonhyperoctahedral* any easy quantum group G such that $K \neq H_n$.

Theorem 6.5. *There are exactly* 11 *nonhyperoctahedral easy quantum groups*:

1. O_n, O_n^* *and* O_n^+, *the orthogonal quantum groups*;
2. S_n *and* S_n^+, *the symmetric quantum groups*;
3. B_n *and* B_n^+, *the bistochastic quantum groups*;
4. S_n' *and* $S_n'^+$, *the modified symmetric quantum groups*;
5. B_n' *and* $B_n'^+$, *the modified bistochastic quantum groups.*

Proof. By Proposition 5.3, we have to classify the easy quantum groups satisfying $K \subset G \subset K^+$. More precisely, we have to prove that for $K = S_n, B_n, S_n', B_n'$ there is no such partial liberation, and that for $K = O_n$ there is only one partial liberation, namely the quantum group K^* mentioned above. This follows from Lemma 6.4, via the Tannakian results in [BS 2009]. □

The classification in the hyperoctahedral case seems to be a difficult problem, which we have to leave open.

7. Laws of characters

We discuss the computation of the asymptotic law of the fundamental character $\chi = \mathrm{Tr}(u)$, and of its truncated versions $\chi_t = \sum_{i=1}^{[tn]} u_{ii}$ with $t \in (0, 1]$.

These computations, which might seem quite technical, are in fact of great relevance in the general context of representation theory. Given a compact group

$G \subset U_n$, or more generally a compact quantum group $G \subset U_n^+$, the main problem in representation theory is to classify the irreducible representations of G. By the Peter–Weyl theory, these irreducible representations appear in the tensor powers $u^{\otimes k}$ of the fundamental representation, and they can be in fact identified with the minimal projections of the algebra $\mathrm{End}(u^{\otimes k})$.

The exact computation of $\mathrm{End}(u^{\otimes k})$ is generally quite difficult. However, an easier problem whose answer is generally extremely useful is the computation of the dimension of this algebra. Since this dimension can be simply obtained by integrating χ^{2k}, we are led to the fundamental problem of computing the law of χ.

In the quantum group context, the difference between the law of χ and the corresponding classical result can be quite puzzling. The problem appears for instance with S_n and S_n^+, where the law of χ is respectively Poisson with $n \to \infty$, and free Poisson with $n \geq 4$. The lack of symmetry was conceptually understood in [Banica and Collins 2007], where it was shown that the correct invariant to look at is the law of the truncated character χ_t, with $t \in (0, 1]$.

Definition 7.1. Associated to an easy quantum group $G \subset U_n^+$ is the truncated character

$$\chi_t = \sum_{i=1}^{[tn]} u_{ii},$$

where $u = (u_{ij})$ is the matrix of standard coordinates, and $t \in (0, 1]$.

Recall some basic results from [BS 2009]. Let G be an easy quantum group, and denote by $D_k \subset P(0, k)$ the corresponding sets of diagrams, having no upper points. We define the Gram matrix to be $G_{kn}(p, q) = n^{b(p\vee q)}$, where $b(\cdot)$ is the number of blocks. The Weingarten matrix is by definition its inverse, $W_{kn} = G_{kn}^{-1}$. In order for this inverse to exist, n must be big enough, and the assumption $n \geq k$ is sufficient. In the general case the notion of quasiinverse must be used; see [Collins and Matsumoto 2009] for a detailed discussion.

Theorem 7.2. *The Haar integration over G is given by*

$$\int_G u_{i_1 j_1} \cdots u_{i_k j_k}\, du = \sum_{p,q\in D_k} \delta_p(i)\delta_q(j) W_{kn}(p, q),$$

where the δ symbols are 0 *or* 1, *depending on whether the indices fit or not.*

Proof. This is proved in [BS 2009], using the idea that the integrals on the left, taken altogether, form the orthogonal projection on $\mathrm{Fix}(u^{\otimes k}) = \mathrm{span}(D_k)$. □

The Weingarten formula is particularly effective in the classical and free cases, where complete computations were performed in [BS 2009].

Theorem 7.3. *The asymptotic law of $\chi_t = \sum_{i=1}^{[tn]} u_{ii}$ with $t \in (0, 1]$ is as follows*:

(1) *For O_n, S_n, H_n and B_n, we get the Gaussian, Poisson, Bessel and shifted Gaussian laws, which form convolution semigroups.*

(2) *For O_n^+, S_n^+, H_n^+ and B_n^+ we get the semicircular, free Poisson, free Bessel and shifted semicircular laws, which form free convolution semigroups.*

(3) *For S_n', B_n', $S_n'^+$ and $B_n'^+$ we get symmetrized versions of the laws for S_n, B_n, S_n^+ and B_n^+, which do not form classical or free convolution semigroups.*

Proof. This is proved in [BS 2009] by using the Weingarten formula and cumulants. Note that the semigroups in (1) and (2) are in Bercovici–Pata [1999] bijection. □

We should mention that the measures in (3), while not forming semigroups due to the canonical copy of $\mathbb{Z}_2$, which produces a "correlation", are very close to forming some kind of semigroup. We come back to this question in our next papers [Banica et al. 2009a; 2009b].

In the remaining cases, the Weingarten formula is less effective, because counting partitions and their blocks is a delicate task. In the case of half-liberations and of the hyperoctahedral series we will use instead the projective versions computed in the previous sections, which reduce the problem to a classical computation.

Definition 7.4. We use the following complex probability measures:

(1) The complex Gaussian law of parameter $t > 0$ is the law of $x + iy$, where x and y are independent Gaussian variables of parameter t.

(2) The s-Bessel law of parameter $t > 0$ is the law of $\sum_{r=1}^{s} e^{2\pi ir/s} x_i$, where $x_1, \dots, x_s$ are independent Poisson variables of parameter t/s.

The complex Gaussian laws are well known to form a convolution semigroup. The same holds for the s-Bessel laws, and we refer to [Banica et al. 2007a] for a complete discussion. The "Bessel" terminology comes from the fact that at $s = 2$, the density of the corresponding discrete measure on $\mathbb{R}$ is given by a Bessel function of the first kind.

Definition 7.5. Given a complex probability measure μ, we call *squeezed version* of it the law of $\sqrt{zz^*}$, where z follows the law μ.

This law doesn't depend of course on the choice of z.

For example, the squeezed version of the complex Gaussian law of parameter 1 is the Rayleigh law. This is because with $z = x + iy$, we have $zz^* = x^2 + y^2$.

Another interesting example, of key relevance in free probability, is the fact that the squeezed version of Voiculescu's circular law is Wigner's semicircle law. See for example [Nica and Speicher 2006].

Theorem 7.6. *The asymptotic law of $\chi_t = \sum_{i=1}^{[tn]} u_{ii}$ with $t \in (0, 1]$ is as follows:*

(1) *For O_n^*, we get the squeezed complex Gaussian semigroup.*

(2) *For $H_n^{(s)}$, we get the squeezed s-Bessel semigroup.*

Proof. The Weingarten formula shows that the odd moments of the variables in the statement are all 0, so all computations actually take place over the projective versions. With this remark in hand, the results simply follow from the well-known fact that χ_t is asymptotically complex Gaussian for U_n and s-Bessel for H_n^s. See [Banica et al. 2007a]. □

The squeezed s-Bessel laws seem to have a quite interesting combinatorics, but this is beyond the purposes of this paper. We would like however to present one such combinatorial statement, in the simplest case, $s = \infty$ and $t = 1$.

Proposition 7.7. *The asymptotic even moments of the character $\chi \in C(H_n^*)$ satisfy*

$$c_k = \sum_{s=0}^{k-1} \binom{k}{s} \binom{k-1}{s} c_s$$

and are equal to the number of games of simple patience with n cards.

Proof. This follows from Theorem 7.6, but we will present below a direct proof, which we found at an early stage of this work. According to the general theory, the numbers in the statement are given by $c_k = \#E_h(2k)$, that is, they count the partitions of $\{1, \dots, 2k\}$ having the property that each block has the same number of odd and even legs.

It is convenient to do the following manipulation: We keep the sequence of odd legs fixed, and we pull downwards the sequence of even legs. In this way, $E_h(2k)$ becomes the set of partitions between an upper and a lower sequence of k points, such that each block is balanced in the sense that it has the same number of upper and lower legs.

Now observe that these partitions can be obtained as follows: pick a number $r \in \{1, \dots, n\}$; connect the first point on the upper line to some $r - 1$ other points on the upper line; choose r points on the lower line, and connect them to the already connected upper r points; and finally connect the remaining $k - r$ upper points to the remaining $k - r$ lower points by means of a balanced partition.

With $s = k - r$ this gives the formula in the statement. For the patience game interpretation, see Aldous and Diaconis [1999] and Sloane's comments [2008] about the sequence A023998, which is the sequence of moments of χ. □

For the higher hyperoctahedral quantum group $H_n^{[s]}$, our standard methods simply do not work. We don't know if this quantum group produces the squeezed version of some known semigroup.

8. Concluding remarks

We have seen in this paper that the easy quantum groups consist in principle of 6 groups, their free versions, 2 half-liberations, and one infinite series still waiting to be constructed. The construction of this hypothetical multiparameter "hyper-octahedral series", and the continuation and completion of our classification work, are of course the main two questions that we would like to address here.

The situation here, which is unexpectedly complex, brings to mind the algebraic difficulty and subtlety of the usual complex reflection groups [Broué et al. 1998].

At the level of applications, as explained in the introduction, we intend to use the easy quantum group list we know of as input for a number of representation theory and probability considerations; again, we believe that "any result that holds for S_n and O_n should have a suitable extension to all easy quantum groups".

In the noneasy case, there are of course of large number of results, classical or even free, having something to do with diagrams and with the easy quantum group technology in general, and that might fall one day into an extension of our formalism.

Here is a list of topics waiting to be developed:

(1) De Finetti theorems. These are available for S_n, O_n from the book [Kallenberg 2005], for S_n^+ from [Köstler and Speicher 2009] and then [Curran 2009a], and for O_n^+ from [Curran 2009b]. We develop a global approach to the problem by using easy quantum groups in our forthcoming paper [Banica et al. 2009a].

(2) Eigenvalue computations. The key results of Diaconis and Shahshahani [1994] about S_n, O_n can also be obtained by using Weingarten functions and cumulants; this is extended to all easy quantum groups in the preprint [Banica et al. 2009b].

(3) Invariant theory. The groups S_n, O_n and their versions S_n^+, O_n^+, O_n^* have served as a guiding example for the study of many invariants; see [Collins and Śniady 2006; Banica and Collins 2007; Novak 2007; Banica and Vergnioux 2009a; 2009b; Collins and Matsumoto 2009]. Some of these results are expected to extend to all easy quantum groups.

(4) Geometric aspects. The groups S_n, O_n and their free versions S_n^+, O_n^+ were also involved in many other "classical versus free" considerations. Let us mention here the Poisson boundary results in [Vaes and Vergnioux 2007], and the quantum isometry groups in [Bhowmick and Goswami 2009]. Once again, the easy quantum groups can lead to some new results here.

(5) Generalizations. One interesting question would be to understand the twisting and deformation of the easy quantum groups, say with the goal of extending our formalism to the $S^2 \neq \mathrm{id}$ case, via monoidal equivalence [Bichon et al. 2006]. Another question is whether the half-liberation operation can be applied to locally

compact real algebraic groups $G \subset M_n(\mathbb{R})$, so as to fit into the general axioms in [Kustermans and Vaes 2000].

In addition to these questions, one basic problem is to classify the intermediate quantum groups $K \subset G \subset K^+$, where K is a fixed easy group. This looks like a quite difficult question; but a possible way forward comes from a conjecture in [Banica et al. 2007c], stating that there is no intermediate quantum group $S_n \subset G \subset S_n^+$. This is actually a quite subtle question, whose study leads straight into the core of the "noneasy" problems.

Acknowledgements

Banica would like to thank Queen's University, where part of this work was done.

References

[Aldous and Diaconis 1999] D. Aldous and P. Diaconis, "Longest increasing subsequences: From patience sorting to the Baik–Deift–Johansson theorem", *Bull. Amer. Math. Soc.* (*N.S.*) **36**:4 (1999), 413–432. MR 2000g:60013 Zbl 0937.60001

[Banica 2008] T. Banica, "A note on free quantum groups", *Ann. Math. Blaise Pascal* **15**:2 (2008), 135–146. MR 2010f:46107 Zbl 05382764

[Banica and Collins 2007] T. Banica and B. Collins, "Integration over quantum permutation groups", *J. Funct. Anal.* **242**:2 (2007), 641–657. MR 2008g:46115 Zbl 1170.46059

[Banica and Vergnioux 2009a] T. Banica and R. Vergnioux, "Fusion rules for quantum reflection groups", *J. Noncommut. Geom.* **3**:3 (2009), 327–359. MR 2511633 Zbl 05578949

[Banica and Vergnioux 2009b] T. Banica and R. Vergnioux, "Invariants of the half-liberated orthogonal group", preprint, 2009. arXiv 0902.2719

[Banica et al. 2007a] T. Banica, S. Belinschi, M. Capitaine, and B. Collins, "Free Bessel laws", preprint, 2007. To appear in *Canad. J. Math.* arXiv 0710.5931v2

[Banica et al. 2007b] T. Banica, J. Bichon, and B. Collins, "The hyperoctahedral quantum group", *J. Ramanujan Math. Soc.* **22**:4 (2007), 345–384. MR 2009d:46125 Zbl 05252140

[Banica et al. 2007c] T. Banica, J. Bichon, and B. Collins, "Quantum permutation groups: A survey", pp. 13–34 in *Noncommutative harmonic analysis with applications to probability*, edited by M. Bożejko et al., Banach Center Publ. **78**, Polish Acad. Sci. Inst. Math., Warsaw, 2007. MR 2009f:46094 Zbl 1140.46329

[Banica et al. 2009a] T. Banica, S. Curran, and R. Speicher, "De Finetti theorems for easy quantum groups", preprint, 2009. arXiv 0907.3314v2

[Banica et al. 2009b] T. Banica, S. Curran, and R. Speicher, "Stochastic aspects of easy quantum groups", preprint, 2009. arXiv 0909.0188v1

[Bercovici and Pata 1999] H. Bercovici and V. Pata, "Stable laws and domains of attraction in free probability theory", *Ann. of Math.* (2) **149**:3 (1999), 1023–1060. MR 2000i:46061 Zbl 0945.46046

[Bhowmick and Goswami 2009] J. Bhowmick and D. Goswami, "Quantum group of orientation-preserving Riemannian isometries", *J. Funct. Anal.* **257**:8 (2009), 2530–2572. MR 2555012 Zbl 1180.58005

[Bichon et al. 2006] J. Bichon, A. De Rijdt, and S. Vaes, "Ergodic coactions with large multiplicity and monoidal equivalence of quantum groups", *Comm. Math. Phys.* **262**:3 (2006), 703–728. MR 2007a:46072 Zbl 1122.46046

[Broué et al. 1998] M. Broué, G. Malle, and R. Rouquier, "Complex reflection groups, braid groups, Hecke algebras", *J. Reine Angew. Math.* **500** (1998), 127–190. MR 99m:20088 Zbl 0921.20046

[BS 2009] T. Banica and R. Speicher, "Liberation of orthogonal Lie groups", *Adv. Math.* **222**:4 (2009), 1461–1501. MR 2554941 Zbl 05614878

[Collins and Matsumoto 2009] B. Collins and S. Matsumoto, "On some properties of orthogonal Weingarten functions", *J. Math. Phys.* **50**:11 (2009), 113516, 14. MR 2567222

[Collins and Śniady 2006] B. Collins and P. Śniady, "Integration with respect to the Haar measure on unitary, orthogonal and symplectic group", *Comm. Math. Phys.* **264**:3 (2006), 773–795. MR 2007c:60009 Zbl 1108.60004

[Curran 2009a] S. Curran, "Quantum exchangeable sequences of algebras", *Indiana Univ. Math. J.* **58**:3 (2009), 1097–1125. MR 2010f:46096 Zbl 1178.46064

[Curran 2009b] S. Curran, "Quantum rotatability", preprint, version 2, 2009. To appear in *Trans. Amer. Math. Soc.* arXiv 0901.1855v2

[Diaconis and Shahshahani 1994] P. Diaconis and M. Shahshahani, "On the eigenvalues of random matrices", *J. Appl. Probab.* **31A** (1994), 49–62. MR 95m:60011 Zbl 0807.15015

[Drinfel'd 1987] V. G. Drinfel'd, "Quantum groups", pp. 798–820 in *Proceedings of the International Congress of Mathematicians* (Berkeley, 1986), vol. 1, edited by A. M. Gleason, Amer. Math. Soc., Providence, RI, 1987. MR 89f:17017 Zbl 0667.16003

[Jimbo 1985] M. Jimbo, "A q-difference analogue of $U(\mathfrak{g})$ and the Yang–Baxter equation", *Lett. Math. Phys.* **10**:1 (1985), 63–69. MR 86k:17008 Zbl 0587.17004

[Kallenberg 2005] O. Kallenberg, *Probabilistic symmetries and invariance principles*, Springer, New York, 2005. MR 2006i:60002 Zbl 1084.60003

[Köstler and Speicher 2009] C. Köstler and R. Speicher, "A noncommutative de Finetti theorem: Invariance under quantum permutations is equivalent to freeness with amalgamation", *Comm. Math. Phys.* **291**:2 (2009), 473–490. MR 2530168 Zbl 1183.81099

[Kustermans and Vaes 2000] J. Kustermans and S. Vaes, "Locally compact quantum groups", *Ann. Sci. École Norm. Sup.* (4) **33**:6 (2000), 837–934. MR 2002f:46108 Zbl 1034.46508

[Nica and Speicher 2006] A. Nica and R. Speicher, *Lectures on the combinatorics of free probability*, London Math. Soc. Lecture Note Ser. **335**, Cambridge University Press, 2006. MR 2008k:46198 Zbl 1133.60003

[Novak 2007] J. Novak, "Truncations of random unitary matrices and Young tableaux", *Electron. J. Combin.* **14**:1 (2007), 21–33. MR 2008f:05202 Zbl 1112.05101

[Sloane 2008] N. J. A. Sloane, "The online encyclopedia of integer sequences", 2008, Available at www.research.att.com/~njas/sequences.

[Speicher 1994] R. Speicher, "Multiplicative functions on the lattice of noncrossing partitions and free convolution", *Math. Ann.* **298**:4 (1994), 611–628. MR 95h:05012 Zbl 0791.06010

[Vaes and Vergnioux 2007] S. Vaes and R. Vergnioux, "The boundary of universal discrete quantum groups, exactness, and factoriality", *Duke Math. J.* **140**:1 (2007), 35–84. MR 2010a:46166 Zbl 1129.46062

[Wang 1995] S. Wang, "Free products of compact quantum groups", *Comm. Math. Phys.* **167**:3 (1995), 671–692. MR 95k:46104 Zbl 0838.46057

[Wang 1998] S. Wang, "Quantum symmetry groups of finite spaces", *Comm. Math. Phys.* **195**:1 (1998), 195–211. MR 99h:58014 Zbl 1013.17008

[Woronowicz 1987] S. L. Woronowicz, "Compact matrix pseudogroups", *Comm. Math. Phys.* **111**:4 (1987), 613–665. MR 88m:46079 Zbl 0627.58034

[Woronowicz 1988] S. L. Woronowicz, "Tannaka–Kreĭn duality for compact matrix pseudogroups: Twisted SU(N) groups", *Invent. Math.* **93**:1 (1988), 35–76. MR 90e:22033 Zbl 0664.58044

Received June 21, 2009.

TEODOR BANICA
DEPARTMENT OF MATHEMATICS
CERGY-PONTOISE UNIVERSITY
95000 CERGY-PONTOISE
FRANCE

Teodor.Banica@u-cergy.fr

STEPHEN CURRAN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CA 94720
UNITED STATES

curransr@math.berkeley.edu

ROLAND SPEICHER
DEPARTMENT OF MATHEMATICS AND STATISTICS
QUEEN'S UNIVERSITY
JEFFERY HALL
KINGSTON, ONTARIO K7L 3N6
CANADA

speicher@mast.queensu.ca

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

BATALIN–VILKOVISKY COALGEBRA OF STRING TOPOLOGY

XIAOJUN CHEN AND WEE LIANG GAN

We prove that the reduced Hochschild homology of a commutative DG Frobenius algebra has the natural structure of a Batalin–Vilkovisky coalgebra, and the reduced cyclic homology has the natural structure of a gravity coalgebra. As an application, this gives an algebraic model for a Batalin–Vilkovisky coalgebra structure on the reduced homology of the free loop space of a simply connected closed oriented manifold, and a gravity coalgebra structure on the reduced equivariant homology.

1. Introduction

Let M be a simply connected closed oriented m-manifold, and let LM be its free loop space. Félix and Thomas [2008] gave a construction of the Batalin–Vilkovisky algebra structure on the homology of LM in terms of Hochschild homology of a Poincaré duality model of M. The aim of this paper is to show that the reduced Hochschild homology, which gives the homology of LM relative to constant loops, has the structure of a Batalin–Vilkovisky coalgebra. As a consequence it is also shown that the reduced cyclic homology of the Poincaré duality model, which models the equivariant homology of LM relative to the constant loops, has the structure of a gravity coalgebra.

Throughout this paper, we shall work over the field of rational numbers. By $C_*(\cdot)$ and $C^*(\cdot)$, we mean the complex of singular chains and the complex of singular cochains. We shall grade $C^*(\cdot)$ negatively. By applying [Lambrechts and Stanley 2008, Theorem 1.1] to the Sullivan minimal model of M, it follows that there is a commutative differential graded (DG) algebra A such that

- A is connected, finite-dimensional, and quasiisomorphic to the DG algebra $C^*(M)$; and
- there is an A-module isomorphism $A \to A^\vee$ of degree m commuting with the differential and inducing the Poincaré duality isomorphism $H^*(M) \to H_{m+*}(M)$ on homology.

MSC2000: 18G60.
Keywords: Batalin–Vilkovisky, coalgebra, string topology.
Gan was partially supported by NSF grant DMS-0726154.

Following Félix and Thomas [2008], we call A a *Poincaré duality model* for M.

Let $C = A^\vee$, the dual space of A. Since A is a commutative DG algebra, C is a cocommutative DG coalgebra. The linear isomorphism $A \xrightarrow{\cong} C[m]$ induces the structure of a commutative DG algebra on C whose product is of degree $-m$. Moreover, the coproduct

$$\Delta : C \to C \otimes C, \quad x \mapsto x' \otimes x''$$

is a map of C-bimodules. Thus, C forms a commutative DG Frobenius algebra in the following sense, which models the chain complex of M:

Definition 1. Let C be a chain complex over a field k. A *commutative DG Frobenius algebra* of degree m on C is a triple $(C, \cdot, \Delta)$ such that $(C, \cdot)$ is a DG commutative algebra whose product is of degree $-m$, (C, Δ) is a DG cocommutative coalgebra, and

$$(1) \quad (x \cdot y)' \otimes (x \cdot y)'' = (x \cdot y') \otimes y'' = (-1)^{m|x'|} x' \otimes (x'' \cdot y) \quad \text{for any } x, y \in C.$$

In Definition 1, C is not necessarily finite-dimensional.

From now on, we shall denote by C a commutative DG Frobenius algebra of degree m with differential d, counit ε, and a coaugmentation $\mathbb{Q} \hookrightarrow C$. By the Hochschild homology $HH_*(C)$ and cyclic homology $HC_*(C)$ of C, we mean the Hochschild homology and cyclic homology of the underlying DG *coalgebra* structure of C. We recall their definitions:

Definition 2. The *Hochschild homology* $HH_*(C)$ of C is the homology of the normalized cocyclic cobar complex $(CC_*(C), b)$, where

$$CC_*(C) = \prod_{n=0}^{\infty} C \otimes (\Sigma \overline{C})^{\otimes n},$$

and

$$\begin{aligned} &b(a_0[a_1|\cdots|a_n]) \\ &\quad := da_0[a_1|\cdots|a_n] + \sum_{i=1}^{n} (-1)^{|a_0|+|[a_1|\cdots|a_{i-1}]|} a_0[a_1|\cdots|da_i|\cdots|a_n] \\ &\qquad\qquad + \sum_{i=1}^{n} (-1)^{|a_0|+|[a_1|\cdots|a_{i-1}|a_i']|} a_0[a_1|\cdots|a_i'|a_i''|\cdots|a_n] \\ &\qquad\qquad + (-1)^{|a_0'|} a_0' \big([a_0''|a_1|\cdots|a_n] - (-1)^{(|a_0''|-1)|[a_1|\cdots|a_n]|}[a_1|\cdots|a_n|a_0'']\big). \end{aligned}$$

Here, $\overline{C} := C/\mathbb{Q} \simeq \ker\{\varepsilon : C \to \mathbb{Q}\}$ and Σ is the desuspension functor (shifting the degrees of $\overline{C}$ down by one), and we write the elements of $C \otimes (\Sigma \overline{C})^{\otimes n}$ in the form $a_0[a_1|\cdots|a_n]$. In particular, the degree $|[a_1|\cdots|a_n]|$ is $(|a_1|-1)+\cdots+(|a_n|-1)$.

One easily checks that $b^2 = 0$. Connes' cyclic operator on the normalized cocyclic cobar complex is given by

$$\begin{aligned} B: \quad CC_*(C) &\to CC_{*+1}(C), \\ a_0[a_1|\cdots|a_n] &\mapsto \sum_{i=1}^{n}(-1)^{|[a_i|\cdots|a_n]||[a_1|\cdots|a_{i-1}]|} \\ &\qquad\qquad \cdot\, \varepsilon(a_0)a_i[a_{i+1}|\cdots|a_n|a_1|\cdots|a_{i-1}]. \end{aligned}$$

One has $B^2 = 0$ and $bB + Bb = 0$.

Definition 3. The *cyclic homology* $HC_*(C)$ of the coalgebra C is the homology of the chain complex $CC_*(C)[u]$, where u is a parameter of degree 2, with differential $b + u^{-1}B$ defined by

$$(b+u^{-1}B)(\alpha\otimes u^n) = \begin{cases} b\alpha\otimes u^n + B\alpha\otimes u^{n-1} & \text{if } n>0, \\ b\alpha & \text{if } n=0 \end{cases}$$

for $\alpha \in CC_*(C)$.

As in the algebra case, one has Connes' exact sequence:

$$\begin{tikzcd} \cdots \arrow[r] & HH_*(C) \arrow[r, "\mathrm{E}"] & HC_*(C) \\ \arrow[r] & HC_{*-2}(C) \arrow[r, "\mathrm{M}"] & HH_{*-1}(C) \arrow[r] & \cdots; \end{tikzcd} \tag{2}$$

compare with [Chen 2007, Theorem 8.3]. We now recall the Batalin–Vilkovisky algebra structure on the Hochschild homology of a commutative DG Frobenius algebra.

Definition 4. A *Batalin–Vilkovisky algebra* is a graded commutative algebra $(V, \bullet)$ together with a linear map $\Delta : V_* \to V_{*+1}$ such that $\Delta\circ\Delta = 0$, and for all $a, b, c \in V$,

$$\begin{aligned} \Delta(a\bullet b\bullet c) = \Delta(a\bullet b)\bullet c + (-1)^{|a|}a\bullet\Delta(b\bullet c) + (-1)^{(|a|-1)|b|}b\bullet\Delta(a\bullet c) \\ - (\Delta a)\bullet b\bullet c - (-1)^{|a|}a\bullet(\Delta b)\bullet c - (-1)^{|a|+|b|}a\bullet b\bullet(\Delta c). \end{aligned} \tag{3}$$

Now for a commutative DG Frobenius algebra C, define a product

$$\bullet : CC_*(C)\otimes CC_*(C) \to CC_*(C)$$

by

$$a_0[a_1|\cdots|a_n]\bullet b_0[b_1|\cdots|b_r] := (-1)^{|b_0||[a_1|\cdots|a_n]|}a_0b_0[a_1|\cdots|a_n|b_1|\cdots|b_r].$$

Theorem 5 [Tradler 2008]. *The Hochschild homology $HH_*(C)[m]$ is a Batalin–Vilkovisky algebra with differential B and product $\bullet$.*

Using the maps E and M in Connes' exact sequence (2), define, for each integer $n \geq 2$, a map c_n of degree $2-n$ by

$$\begin{aligned} c_n : HC_*(C)[m-2]^{\otimes n} &\to HC_*(C)[m-2] \\ \alpha_1 \otimes \cdots \otimes \alpha_n &\to (-1)^{\epsilon}\, \mathrm{E}(\mathrm{M}(\alpha_1) \bullet \cdots \bullet \mathrm{M}(\alpha_n)), \end{aligned}$$

where $\epsilon = (n-1)|\alpha_1| + (n-2)|\alpha_2| + \cdots + |\alpha_{n-1}|$. The corollary, which follows from a general result (see Proposition 24), is this:

Corollary 6. *The cyclic homology* $(HC_*(C)[m-2], \{c_n\})$ *is a gravity algebra.*

Definition 7. A *gravity algebra* is a graded vector space V with a sequence of graded skew-symmetric operators

$$\{x_1, \ldots, x_k\} : V^{\otimes k} \to V \quad \text{for } k = 2, 3, \ldots$$

of degree $2-k$ that satisfy the generalized Jacobi identities

$$\sum_{1 \leq i < j \leq k} (-1)^{\epsilon(i,j)} \{\{x_i, x_j\}, x_1, \ldots, \hat{x}_i, \ldots, \hat{x}_j, \ldots, x_k, y_1, \ldots, y_l\} = \begin{cases} \{\{x_1, \ldots, x_k\}, y_1, \ldots, y_l\} & \text{if } l > 0, \\ 0 & \text{if } l = 0. \end{cases} \tag{4}$$

where $\epsilon(i, j) = (|x_1| + \cdots + |x_{i-1}|)|x_i| + (|x_1| + \cdots + |x_{j-1}|)|x_j| + |x_i||x_j|$.

In this paper, by the *reduced Hochschild homology* $\widetilde{HH}_*(C)$ of C, we mean the homology of

$$\widetilde{CC}_*(C) := CC_*(C)/C = \prod_{n=1}^{\infty} C \otimes (\Sigma \overline{C})^{\otimes n}.$$

By the *reduced cyclic homology* $\widetilde{HC}_*(C)$ of C, we mean the homology of

$$\widetilde{CC}_*(C)[u] = CC_*(C)[u]/C[u].$$

As above, we have $\mathrm{E} : \widetilde{HH}_*(C) \to \widetilde{HC}_*(C)$ and $\mathrm{M} : \widetilde{HC}_*(C) \to \widetilde{HH}_{*+1}(C)$.

Define a coproduct

$$\vee : \widetilde{CC}_*(C) \to \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C)$$

by

$$\vee(a_0[a_1 | \cdots | a_n]) := \sum_{i=2}^{n-1} (-1)^{\epsilon(i)}\, (a_0 a_i)'[a_1 | \cdots | a_{i-1}] \otimes (a_0 a_i)''[a_{i+1} | \cdots | a_n], \tag{5}$$

where $\epsilon(i) = |a_0| + (1 + |a_i| + |(a_0 a_i)''|)|[a_1 | \cdots | a_{i-1}]|$.

Our main result is the following.

Theorem 8. *The reduced Hochschild homology $\widetilde{HH}_*(C)[1-m]$ is a Batalin–Vilkovisky coalgebra with differential B and coproduct $\vee$.*

The proof of Theorem 8 uses several identities at the chain level involving certain homotopies which we will give in Section 5.

Similarly to above, define a map $s_n : \widetilde{HC}_*(C)[2-m] \to \widetilde{HC}_*(C)[2-m]^{\otimes n}$ of degree $2-n$ by

$$s_n(\alpha) := (\mathrm{E} \otimes \cdots \otimes \mathrm{E}) \circ (\vee \otimes \mathrm{id}^{\otimes n-2}) \circ \cdots \circ (\vee \otimes \mathrm{id}) \circ \vee \circ \mathrm{M}(\alpha)$$

for any $\alpha \in \widetilde{HC}_*(C)[2-m]$.

Corollary 9. *The reduced cyclic homology $(\widetilde{HC}_*(C)[2-m], \{s_n\})$ is a gravity coalgebra.*

In the two statements above, the Batalin–Vilkovisky coalgebra and gravity coalgebra are defined as dual versions of the corresponding algebras (see Definitions 14 and 15). The Batalin–Vilkovisky algebra and the gravity algebra structures of $HH_*(C)$ and $HC_*(C)$ descend to $\widetilde{HH}_*(C)$ and $\widetilde{HC}_*(C)$, respectively. Thus, we obtain both Batalin–Vilkovisky algebra and coalgebra structures on $\widetilde{HH}_*(C)$, and gravity algebra and coalgebra structures on $\widetilde{HC}_*(C)$.

Let A be a Poincaré duality model for M and $C = A^\vee$. Let LM be the free loop space of M. From [Jones 1987], one has isomorphisms

$$H_*(LM, M) \cong \widetilde{HH}_*(C) \quad \text{and} \quad H_*^{S^1}(LM, M) \cong \widetilde{HC}_*(C).$$

Following [Chas and Sullivan 2004], we call $H_*(LM, M)$ the reduced homology of the free loop space, and $H_*^{S^1}(LM, M)$ the reduced equivariant homology of the free loop space. As a consequence, the choice of a Poincaré duality model for M gives the reduced homology of the free loop space the structure of a Batalin–Vilkovisky coalgebra, and the reduced equivariant homology of the free loop space the structure of a gravity coalgebra. In string topology, the loop product $\bullet$ was first introduced in [Chas and Sullivan 1999]; see also [Cohen and Jones 2002]. The coproduct $\vee$ was introduced in [Sullivan 2004]. The operators c_n and s_n were first introduced in [Chas and Sullivan 2004] and discussed further in [Sullivan 2004]; see also [Westerland 2008].

Getzler [1994a; 1994b; 1995] studied Batalin–Vilkovisky algebras and gravity algebras in his works on topological conformal field theories (TCFT). He showed that a (genus zero) TCFT (respectively, an equivariant TCFT) with one output is the same as a Batalin–Vilkovisky algebra (respectively, a gravity algebra). If we consider multiple inputs and outputs, we then obtain both Batalin–Vilkovisky algebra and coalgebra (respectively, gravity algebra and coalgebra). Our construction gives an algebraic proof that string topology is a part of a (genus zero) TCFT. We expect that the constructions above can be generalized to homotopy versions of commutative DG Frobenius algebras.

Remark 10. Sullivan's coproduct $\vee$ is not the same as the loop coproduct introduced in [Cohen and Godin 2004]; see also [Godin 2007].

Remark 11. Theorem 5 is not new; it is well known that the Hochschild cohomology of a Frobenius algebra has the structure of a Batalin–Vilkovisky algebra; see, for example, [Menichi 2004] and [Tradler 2008]. However, notice that the formulas we give above in terms of the Hochschild homology of a Frobenius coalgebra are really explicit and simple. The proof of Theorem 5 is included in this paper so that the reader can compare it with the proof of Theorem 8. As far as we are aware, Theorem 8 is new. Its statement is not true in general at the chain level, and the homotopy operators that appear in its proof are also new.

The BV coalgebra structure in Theorem 8 also appears to be related to [Eu and Schedler 2009, Question 2.3.72].

The rest of this paper is organized as follows. We recall the definitions of Batalin–Vilkovisky algebras and gravity algebras in Section 2 and the proof of Theorem 5 in Section 3. We give the proof of Corollary 6 in Section 4, the proof of Theorem 8 in Section 5, and the proof of Corollary 9 in Section 6.

Koszul sign rule. All ± signs in this paper are determined by the Koszul rule for signs. Thus, whenever we switch two elements $a \otimes b \mapsto b \otimes a$, we put $(-1)^{|a||b|}$ in front of $b \otimes a$ and write $\pm b \otimes a$. Also, if f and g are operators of homogeneous degree, then $(f \otimes g)(a \otimes b) = \pm f(a) \otimes g(b) = (-1)^{|g||a|} f(a) \otimes g(b)$. For example, in (5), to see that $\epsilon(i) = |a_0| + (1 + |a_i| + |(a_0a_i)''|)|[a_1 | \cdots | a_{i-1}]|$ is given by the Koszul sign rule, note that the term $(1 + |a_i|)|[a_1 | \cdots | a_{i-1}]|$ comes from first moving $[a_i]$ to the left of $[a_1 | \cdots | a_{i-1}]$, the term $|a_0|$ comes from moving a suspension operator to the right of a_0 to apply it to $[a_i]$, and $|(a_0a_i)''||[a_1 | \cdots | a_{i-1}]|$ comes from moving $(a_0a_i)''$ to the right of $[a_1 | \cdots | a_{i-1}]$. Similarly, the signs in the formulas above for b, B, the product $\bullet$, c_n, and so on, are also all given by the Koszul sign rule.

2. Batalin–Vilkovisky algebras and gravity algebras

Lemma 12. *Let* $(V, \bullet, \Delta)$ *be a Batalin–Vilkovisky algebra. Define*

$$\{\cdot, \cdot\} : V \otimes V \to V$$

by

$$\{a, b\} := (-1)^{|a|} \Delta(a \bullet b) - (-1)^{|a|} (\Delta a) \bullet b - a \bullet (\Delta b).$$

Then $(V[-1], \{\cdot, \cdot\}, \Delta)$ *is a DG Lie algebra.*

Proof. See [Getzler 1994a, Proposition 1.2]. □

More generally, one has the following result proved by Getzler; see [1994b, Theorem 4.5] and [1995, §3.4].

Theorem 13. *Let* $(V, \bullet, \Delta)$ *be a Batalin–Vilkovisky algebra. For* $k = 2, 3, \ldots,$ *define*

$$\{\cdot, \ldots, \cdot\} : V^{\otimes k} \to V$$

by

$$\{a_1, \ldots, a_k\} := (-1)^{\epsilon} \Big(\Delta(a_1 a_2 \cdots a_k) - \sum_{i=1}^{k} (-1)^{|a_1|+\cdots+|a_{i-1}|} a_1 \cdots (\Delta a_i) \cdots a_k \Big),$$

where $\epsilon = (k-1)|a_1| + (k-2)|a_2| + \cdots + |a_{k-1}|$. *Then* $V[-1]$ *is a DG gravity algebra with differential* Δ *and brackets* $\{a_1, \ldots, a_k\}$.

A DG gravity algebra is a gravity algebra with a differential commuting with all the brackets. Thus, for a Batalin–Vilkovisky algebra $(V, \bullet, \Delta)$, its homology $H(V, \Delta)[-1]$ has a gravity algebra structure. Taking $k = 3$ and $l = 0$ in (4) gives the graded Jacobi identity. Hence, a gravity algebra has a graded Lie algebra structure.

Analogously, we may introduce the notions of a Batalin–Vilkovisky coalgebra and a gravity coalgebra.

Definition 14. A *Batalin–Vilkovisky coalgebra* is a graded cocommutative coalgebra $(V, \vee)$ together with a linear map $\Delta : V_* \to V_{*+1}$ such that $\Delta \circ \Delta = 0$, and

$$\begin{aligned}(\Delta \otimes \mathrm{id}^{\otimes 2} + \mathrm{id} \otimes \Delta \otimes \mathrm{id} + \mathrm{id}^{\otimes 2} \otimes \Delta) \circ (\vee \otimes \mathrm{id}) \circ \vee(a) \\ = (\tau^2 + \tau + \mathrm{id}) \circ (\vee \circ \Delta \otimes \mathrm{id}) \circ \vee(a) + (\vee \otimes \mathrm{id}) \circ \vee \circ \Delta(a)\end{aligned}$$

for all $a \in V$, where τ is the cyclic permutation $\tau : a \otimes b \otimes c \mapsto c \otimes a \otimes b$.

Similarly to the Batalin–Vilkovisky algebra case, the chain complex (V, Δ) is a DG gravity coalgebra:

Definition 15. A *gravity coalgebra* is a graded vector space V with a sequence of graded skew-symmetric operators

$$m_k : V \to V^{\otimes k} \quad \text{for } k = 2, 3, 4, \ldots$$

of degree $2 - k$, such that

$$S_{2,k-2} \circ (m_2 \otimes \mathrm{id}^{\otimes k-2}) \circ m_{k-1+l} = (m_k \otimes \mathrm{id}^{\otimes l}) \circ m_{l+1} : V \to V^{k+l}, \tag{6}$$

where the range of the mapping $(m_2 \otimes \mathrm{id}^{\otimes k-2}) \circ m_{k-1+l} : V \to V^{k+l}$ is identified with $V^{\otimes 2} \otimes V^{\otimes k-2} \otimes V^{\otimes l}$ and $S_{2,k-2}$ is the shuffle product $V^{\otimes 2} \otimes V^{\otimes k-2} \to V^{\otimes k}$, and if $l = 0$, we set $m_1 = 0$.

Theorem 16. *Let* $(V, \vee, \Delta)$ *be a Batalin–Vilkovisky coalgebra. For any* $x \in V$, *let*

$$\vee_k(x) := (\vee \otimes \mathrm{id}^{\otimes k-2}) \circ \cdots \circ (\vee \otimes \mathrm{id}) \circ \vee(x) = \sum x_1 \otimes x_2 \otimes \cdots \otimes x_k,$$

and let

$$s_k(x) := \sum (-1)^{(k-1)|x_1|+(k-2)|x_2|+\cdots+|x_{k-1}|} \cdot \Big(\vee_k(\Delta x) - \sum_{i=0}^{k-1} (\mathrm{id}^{\otimes i} \otimes \Delta \otimes \mathrm{id}^{\otimes k-i-1}) \circ \vee_k(x)\Big),$$

for $k = 2, 3, \ldots$. *Then* $V[1]$ *is a DG gravity coalgebra with differential* Δ *and cobrackets* $\{s_n\}$. *In particular*, $(V[1], s_2, \Delta)$ *is a DG Lie coalgebra.*

The proof of the theorem is completely dual to that of Theorem 13.

3. The Batalin–Vilkovisky algebra

Next we recall the proof of Theorem 5 from [Chen 2007].

Lemma 17. *The chain complex* $(CC_*(C)[m], b)$ *is a DG algebra with product* $\bullet$.

Proof. The proof is by direct verification; see [Chen 2007, Lemma 4.1]. □

The product $\bullet$ on $CC_*(C)[m]$ is not commutative, but homotopy commutative:

Lemma 18. *Define a bilinear operator*

$$* : CC_*(C) \otimes CC_*(C) \to CC_*(C)$$

as follows: *for* $\alpha = a_0[a_1 | \cdots | a_n]$, $\beta = b_0[b_1 | \cdots | b_r] \in CC_*(C)$,

$$\alpha * \beta := \sum_{i=1}^{n} (-1)^{|b_0|+(|\beta|-1)|[a_{i+1}|\cdots|a_n]|} \cdot \varepsilon(a_i b_0) a_0[a_1 | \cdots | a_{i-1} | b_1 | \cdots | b_r | a_{i+1} | \cdots | a_n]. \tag{7}$$

Then

$$b(\alpha * \beta) = b\alpha * \beta + (-1)^{|\alpha|+1} \alpha * b\beta + (-1)^{|\alpha|}(\alpha \bullet \beta - (-1)^{|\alpha||\beta|} \beta \bullet \alpha). \tag{8}$$

Proof. The proof is by direct verification; see [Chen 2007, Lemma 5.1]. □

It follows from Lemma 17 and Lemma 18 that $(HH_*(C)[m], \bullet)$ is a graded commutative algebra.

Define the binary operator

$$\{\cdot, \cdot\} : CC_*(C) \otimes CC_*(C) \to CC_*(C)$$

to be the commutator of $*$ above, namely

$$\{\alpha, \beta\} := \alpha * \beta - (-1)^{(|\alpha|+1)(|\beta|+1)} \beta * \alpha \quad \text{for } \alpha, \beta \in CC_*(C).$$

Lemma 19. *The chain complex* $(CC_*(C)[m-1], b)$ *is a DG Lie algebra with the Lie bracket* $\{\cdot,\cdot\}$.

Proof. The proof is direct; see [Chen 2007, Lemma 5.4 and Corollary 5.5]. □

In particular $HH_*(C)[m-1]$ is a graded Lie algebra. Moreover, $\bullet$ and $\{\cdot,\cdot\}$ are compatible in the following sense, which makes $HH_*(C)[m]$ into a Gerstenhaber algebra:

Definition 20 [Gerstenhaber 1963]. Let V be a graded vector space. A *Gerstenhaber algebra* on V is a triple $(V,\cdot,\{\cdot,\cdot\})$ such that

(i) $(V,\cdot)$ is a graded commutative algebra;

(ii) $(V,\{\cdot,\cdot\})$ is a graded Lie algebra whose Lie bracket is of degree 1;

(iii) for any $\alpha,\beta,\gamma\in V$, one has

$$\{\alpha\bullet\beta,\gamma\}=\alpha\bullet\{\beta,\gamma\}+(-1)^{|\beta|(|\gamma|+1)}\{\alpha,\gamma\}\bullet\beta. \tag{9}$$

Theorem 21. *The Hochschild homology* $HH_*(C)[m]$ *is a Gerstenhaber algebra, with product* $\bullet$ *and bracket* $\{\cdot,\cdot\}$.

Proof. From above, $HH_*(C)[m]$ is both a graded commutative algebra and a degree one graded Lie algebra. Equation (9) is immediate from Lemma 22. □

Lemma 22. *For any*

$$\alpha=a_0[a_1|\cdots|a_n],\quad \beta=b_0[b_1|\cdots|b_r],\quad \gamma=c_0[c_1|\cdots|c_l]\in CC_*(C),$$

one has

(i) $(\alpha\bullet\beta)*\gamma=\alpha\bullet(\beta*\gamma)+(-1)^{|\beta|(|\gamma|+1)}(\alpha*\gamma)\bullet\beta$; *and*

(ii) $\gamma*(\alpha\bullet\beta)-(\gamma*\alpha)\bullet\beta-(-1)^{|\alpha|(|\gamma|+1)}\alpha\bullet(\gamma*\beta)=(b\circ\rho-\rho\circ b)(\alpha\otimes\beta\otimes\gamma)$, *where*

$$\rho(\alpha\otimes\beta\otimes\gamma):=\sum_{i<j}(-1)^{\epsilon}\varepsilon(c_ia_0)\varepsilon(c_jb_0)c_0[c_1|\cdots|c_{i-1}|a_1|\cdots|a_n|c_{i+1}| \\ \cdots|c_{j-1}|b_1|\cdots|b_r|c_{j+1}|\cdots|c_l],$$

and $\epsilon=(|\alpha|-1)|[c_{i+1}|\cdots|c_n]|+(|\beta|-1)|[c_{j+1}|\cdots|c_n]|$.

Proof. The proof is by direct verification; see [Chen 2007, Lemma 5.8]. □

Theorem 5 follows from [Getzler 1994a, Proposition 1.2], Theorem 21, and the following:

Lemma 23. *For any* $\alpha,\beta\in HH_*(C)[m]$, *one has*

$$\{\alpha,\beta\}=(-1)^{|\alpha|}B(\alpha\bullet\beta)-(-1)^{|\alpha|}B(\alpha)\bullet\beta-\alpha\bullet B(\beta).$$

More precisely, for $\alpha = x[a_1 \,|\, \cdots \,|\, a_n]$ *and* $\beta = y[b_1 \,|\, \cdots \,|\, b_r] \in CC_*(C)$, *define*

$$\begin{aligned}
&\phi(\alpha, \beta) \\
&\quad := \sum_{i<j} \pm \varepsilon(x)\varepsilon(a_j y) a_i [a_{i+1} \,|\, \cdots \,|\, a_{j-1} \,|\, b_1 \,|\, \cdots \,|\, b_r \,|\, a_{j+1} \,|\, \cdots \,|\, a_n \,|\, a_1 \,|\, \cdots \,|\, a_{i-1}], \\
&\psi(\alpha, \beta) \\
&\quad := \sum_{k<l} \pm \varepsilon(y)\varepsilon(b_l x) b_k [b_{k+1} \,|\, \cdots \,|\, b_{l-1} \,|\, a_1 \,|\, \cdots \,|\, a_n \,|\, b_{l+1} \,|\, \cdots \,|\, b_r \,|\, b_1 \,|\, \cdots \,|\, b_{k-1}],
\end{aligned}$$

and let $\theta := \phi + \psi$. (*The* ± *signs are determined by the Koszul sign rule.*) *Then*

$$\begin{aligned}
&(b \circ \theta + \theta \circ b)(\alpha \otimes \beta) \\
&\qquad = \{\alpha, \beta\} - (-1)^{|\alpha|} B(\alpha \bullet \beta) - (-1)^{(|\beta|+1)(|\alpha|+1)} \beta \bullet B(\alpha) + \alpha \bullet B(\beta).
\end{aligned}$$

Proof. The proof is by a direct verification; see [Chen 2007, Lemma 7.3]. □

4. The gravity algebra

We define the complex $(CC_*(C)[u, u^{-1}], b + u^{-1}B)$ by

$$(b + u^{-1}B)(\alpha \otimes u^n) = b\alpha \otimes u^n + B\alpha \otimes u^{n-1} \quad \text{for all } n.$$

The quotient of $(CC_*(C)[u, u^{-1}], b + u^{-1}B)$ by its subcomplex $CC_*(C)[u^{-1}]u^{-1}$ is the complex $(CC_*(C)[u], b + u^{-1}B)$ in Definition 3. The short exact sequence

$$0 \to CC_*(C) \longrightarrow CC_*(C)[u] \xrightarrow{u^{-1}} CC_*(C)[u] \to 0$$

induces the long exact sequence (2). By diagram chasing, one can see that

$$\mathrm{M} \circ \mathrm{E} = B : HH_*(C) \to HH_{*+1}(C).$$

Corollary 6 is immediate from Theorem 5 and the following general result; see [Chen 2007, Theorem 8.5].

Proposition 24. *Let* $(V, \bullet, \Delta)$ *be a Batalin–Vilkovisky algebra, and let* W *be a graded vector space. Let* $\mathrm{E} : V_* \to W_*$ *and* $\mathrm{M} : W_* \to V_{*+1}$ *be two maps such that* $\mathrm{E} \circ \mathrm{M} = 0$ *and* $\mathrm{M} \circ \mathrm{E} = \Delta$. *Then* $(W[-2], \{c_n\})$ *is a gravity algebra, where*

$$c_n(\alpha_1 \otimes \cdots \otimes \alpha_n) := (-1)^{(n-1)|\alpha_1| + (n-2)|\alpha_2| + \cdots + |\alpha_{n-1}|} \mathrm{E}(\mathrm{M}(\alpha_1) \bullet \cdots \bullet \mathrm{M}(\alpha_n)).$$

Proof. It follows from (3), by induction on n, that

$$\begin{aligned}
(10) \quad \Delta(x_1 \bullet x_2 \bullet \cdots \bullet x_n) = \sum_{i<j} \pm \Delta(x_i \bullet x_j) \bullet x_1 \bullet \cdots \bullet \widehat{x_i} \bullet \cdots \bullet \widehat{x_j} \bullet \cdots \bullet x_n \\
+ (n-2) \sum_i \pm x_1 \bullet \cdots \bullet \Delta x_i \bullet \cdots \bullet x_n.
\end{aligned}$$

Now let $x_i = \mathrm{M}(\alpha_i)$, and apply E to both sides of the above equality; we obtain

$$\begin{aligned}
&\mathrm{E} \circ \Delta(\mathrm{M}(\alpha_1) \bullet \mathrm{M}(\alpha_2) \bullet \cdots \bullet \mathrm{M}(\alpha_n)) \\
&\quad = \sum_{i<j} \pm \mathrm{E} \circ \big(\Delta(\mathrm{M}(\alpha_i) \bullet \mathrm{M}(\alpha_j)) \bullet \mathrm{M}(\alpha_1) \bullet \cdots \bullet \widehat{\mathrm{M}(\alpha_i)} \bullet \cdots \bullet \widehat{\mathrm{M}(\alpha_j)} \bullet \cdots \bullet \mathrm{M}(\alpha_n)\big) \\
&\qquad\qquad + (n-2) \sum_i \pm \mathrm{E}(\mathrm{M}(\alpha_1) \bullet \cdots \bullet \Delta \circ \mathrm{M}(\alpha_i) \bullet \cdots \bullet \mathrm{M}(\alpha_n)).
\end{aligned}$$

Since $\mathrm{E} \circ \Delta = \mathrm{E} \circ \mathrm{M} \circ \mathrm{E} = 0$ and $\Delta \circ \mathrm{M} = \mathrm{M} \circ \mathrm{E} \circ \mathrm{M} = 0$, we have

$$\sum_{1 \le i < j \le n} \pm c_{n-1}(c_2(\alpha_i \otimes \alpha_j) \otimes \alpha_1 \otimes \cdots \otimes \widehat{\alpha_i} \otimes \cdots \otimes \widehat{\alpha_j} \otimes \cdots \otimes \alpha_n) = 0.$$

Similarly, by multiplying $y_1 \bullet \cdots \bullet y_l$ on both sides of (10), letting $y_j = \mathrm{M}(\beta_j)$, and then applying E on both sides, we obtain

$$\begin{aligned}
\sum_{1 \le i < j \le n} \pm c_{n+l-1}(c_2(\alpha_i \otimes \alpha_j) \otimes \alpha_1 \otimes \cdots \otimes \widehat{\alpha_i} \otimes \cdots \otimes \widehat{\alpha_j} \otimes \cdots \otimes \alpha_n \otimes \beta_1 \otimes \cdots \otimes \beta_l) \\
= c_{l+1}(c_n(\alpha_1 \otimes \cdots \otimes \alpha_n) \otimes \beta_1 \otimes \cdots \otimes \beta_l)
\end{aligned}$$

for $l > 0$. This proves the proposition. □

Proposition 24 can also be applied to the Hochschild homology of a Calabi-Yau algebra (see [Ginzburg 2006, Theorem 3.4.3]) to give a gravity algebra structure on its cyclic homology.

5. The Batalin–Vilkovisky coalgebra

The proof of Theorem 8 is similar to the proof of Theorem 5.

Lemma 25. *The chain complex* $(\widetilde{CC_*}(C)[1-m], b)$ *is a DG coalgebra with coproduct* $\vee$.

Proof. It is clear that $\vee$ is coassociative. Therefore we only need to check that b is a derivation with respect to $\vee$. Observe that the expressions $b \circ \vee(\alpha)$ and $\vee \circ b(\alpha)$ have two parts, one contains those terms involving the differentials of the entries in α (which we call the *differential part*), the other contains those terms involving the coproducts of the entries in α (which we call the *diagonal part*). It follows directly from the definition of $\vee$ that the differential parts of $b \circ \vee(\alpha)$ and $\vee \circ b(\alpha)$ are equal. For the diagonal parts, omitting the signs determined by the Koszul sign rule from our notation (see 32), we have

(11) $b \circ \vee(a_0[a_1 | \cdots | a_n])$

(12) $$= \sum_{1<i<n} b\big((a_0a_i)'[a_1 | \cdots | a_{i-1}]\big) \otimes (a_0a_i)''[a_{i+1} | \cdots | a_n]$$

$$\pm \sum_{1<i<n} (a_0a_i)'[a_1\,|\cdots|\,a_{i-1}] \otimes b\Big((a_0a_i)''[a_{i+1}\,|\cdots|\,a_n]\Big) \tag{13}$$

$$= \sum_{1\le j<i<n} \pm(a_0a_i)'[a_1\,|\cdots|\,a_j'\,|\,a_j''\,|\cdots|\,a_{i-1}] \otimes (a_0a_i)''[a_{i+1}\,|\cdots|\,a_n] \tag{14}$$

$$+ \sum_{1<i<n} \pm((a_0a_i)')'[((a_0a_i)')''\,|\,a_1\,|\cdots|\,a_{i-1}] \otimes (a_0a_i)''[a_{i+1}\,|\cdots|\,a_n] \tag{15}$$

$$- \sum_{1<i<n} \pm((a_0a_i)')'[a_1\,|\cdots|\,a_{i-1}\,|\,((a_0a_i)')''] \otimes (a_0a_i)''[a_{i+1}\,|\cdots|\,a_n] \tag{16}$$

$$+ \sum_{1<i<j\le n} \pm(a_0a_i)'[a_1\,|\cdots|\,a_{i-1}] \otimes (a_0a_i)''[a_{i+1}\,|\cdots|\,a_j'\,|\,a_j''\,|\cdots|\,a_n] \tag{17}$$

$$+ \sum_{1<i<n} \pm(a_0a_i)'[a_1\,|\cdots|\,a_{i-1}] \otimes ((a_0a_i)'')'[((a_0a_i)'')''\,|\,a_{i+1}\,|\cdots|\,a_n] \tag{18}$$

$$- \sum_{1<i<n} \pm(a_0a_i)'[a_1\,|\cdots|\,a_{i-1}] \otimes ((a_0a_i)'')'[a_{i+1}\,|\cdots|\,a_n\,|\,((a_0a_i)'')''], \tag{19}$$

while

$$\vee \circ b(a_0[a_1\,|\cdots|\,a_n]) \tag{20}$$

$$= \sum_{1\le j<i<n} \pm(a_0a_i)'[a_1\,|\cdots|\,a_j'\,|\,a_j''\,|\cdots|\,a_{i-1}] \otimes (a_0a_i)''[a_{i+1}\,|\cdots|\,a_n] \tag{21}$$

$$+ \sum_{1<i<j\le n} \pm(a_0a_i)'[a_1\,|\cdots|\,a_{i-1}] \otimes (a_0a_i)''[a_{i+1}\,|\cdots|\,a_j'\,|\,a_j''\,|\cdots|\,a_n] \tag{22}$$

$$+ \sum_{1<i<n} \pm(a_0a_i')'[a_1\,|\cdots|\,a_{i-1}] \otimes (a_0a_i')''[a_i''\,|\,a_{i+1}\,|\cdots|\,a_n] \tag{23}$$

$$\pm\, (a_0a_n')'[a_1\,|\cdots|\,a_{n-1}] \otimes (a_0a_n')''[a_n''] \tag{24}$$

$$\pm\, (a_0a_1'')'[a_1'] \otimes (a_0a_1'')''[a_2\,|\cdots|\,a_n] \tag{25}$$

$$+ \sum_{1<i<n} \pm(a_0a_i'')'[a_1\,|\cdots|\,a_{i-1}\,|\,a_i'] \otimes (a_0a_i'')''[a_{i+1}\,|\cdots|\,a_n] \tag{26}$$

$$\pm\, (a_0'a_1)'[a_0''] \otimes (a_0'a_1)''[a_2\,|\cdots|\,a_n] \tag{27}$$

$$+ \sum_{1<i<n} \pm(a_0'a_i)'[a_0''\,|\,a_1\,|\cdots|\,a_{i-1}] \otimes (a_0'a_i)''[a_{i+1}\,|\cdots|\,a_n] \tag{28}$$

$$- \sum_{1<i<n} \pm(a_0'a_i)'[a_1\,|\cdots|\,a_{i-1}] \otimes (a_0a_i)''[a_{i+1}\,|\cdots|\,a_n\,|\,a_0''] \tag{29}$$

$$- \pm(a_0'a_n)'[a_1\,|\cdots|\,a_{n-1}] \otimes (a_0'a_n)''[a_0'']. \tag{30}$$

Keeping (1) in mind, we see that (14) and (21) are equal; so are (15) and (28), (16) and (26), (17) and (22), (18) and (23), and (19) and (29). Also, (24) and (30) cancel; so do (25) and (27). Hence, (11) = (20). □

Define the permutations τ and σ by

$$\tau : \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C) \to \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C)$$
$$\alpha_1 \otimes \alpha_2 \mapsto \pm\alpha_2 \otimes \alpha_1$$

and

$$\sigma : \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C) \to \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C)$$
$$\alpha_1 \otimes \alpha_2 \otimes \alpha_3 \mapsto \pm\alpha_2 \otimes \alpha_3 \otimes \alpha_1.$$

The following lemma says that $\vee$ is cocommutative up to homotopy, and therefore $(\widetilde{HH}_*(C)[1-m], \vee)$ is a graded cocommutative, coassociative coalgebra.

Lemma 26. *Let $h : \widetilde{CC}_*(C) \to \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C)$ be defined by*

$$h(\alpha) := \sum_{i<j} \pm a_0[a_1| \cdots |a_{i-1}|a_{j+1}| \cdots |a_n] \otimes a_i a_j [a_{i+1}| \cdots |a_{j-1}]$$

for any $\alpha = a_0[a_1| \cdots |a_n] \in \widetilde{CC}_*(C)$. (The $\pm$ sign is determined by the Koszul sign rule on page 32.) Then

$$b \circ h(\alpha) - h \circ b(\alpha) = \tau \circ \vee(\alpha) - \vee(\alpha). \tag{31}$$

Proof. It is easy to see that the differential parts of the left side of (31) cancel each other, so we only need to consider the diagonal parts. In fact, the diagonal parts of $h(b\alpha)$ are equal to

$$\sum_i \pm a_0'[a_{i+1}| \cdots |a_n] \otimes (a_0'' a_i)[a_1| \cdots |a_{i-1}] \tag{32}$$

$$+ \sum_{i<j} \pm a_0'[a_0''|a_1| \cdots |a_{i-1}|a_{j+1}| \cdots |a_n] \otimes (a_i a_j)[a_{i+1}| \cdots |a_{j-1}] \tag{33}$$

$$- \sum_i \pm a_0'[a_1| \cdots |a_{i-1}] \otimes (a_i a_0'')[a_{i+1}| \cdots |a_n] \tag{34}$$

$$- \sum_{i<j} \pm a_0'[a_1| \cdots |a_{i-1}|a_{j+1}| \cdots |a_n|a_0''] \otimes (a_i a_j)[a_{i+1}| \cdots |a_{j-1}] \tag{35}$$

$$+ \sum \pm a_0[a_1| \cdots |a_k'|a_k''| \cdots |a_n] \otimes (a_i a_j)[a_{i+1}| \cdots |a_{j-1}] \tag{36}$$

$$+ \sum_{i<k<j} \pm a_0[a_1| \cdots |a_{i-1}|a_{j+1}| \cdots |a_n] \otimes (a_i a_j)[a_{i+1}| \cdots |a_k'|a_k''| \cdots |a_{j-1}] \tag{37}$$

$$+ \sum_{i<j} \pm a_0[a_1| \cdots |a_{i-1}|a_{j+1}| \cdots |a_n] \otimes (a_i a_j)'[(a_i a_j)''|a_{i+1}| \cdots |a_{j-1}] \tag{38}$$

$$- \sum_{i<j} \pm a_0[a_1| \cdots |a_{i-1}|a_{j+1}| \cdots |a_n] \otimes (a_i a_j)'[a_{i+1}| \cdots |a_{j-1}|(a_i a_j)''], \tag{39}$$

(where the sum in (36) is taken over all $k < i < j$ and $i < j < k$).

Now (33)+(35)+(36)+(37)+(38)+(39) is exactly $b(h\alpha)$, while the remaining terms (32) + (34) are exactly $-\tau \circ \vee(\alpha) + \vee(\alpha)$. □

Lemma 27. *Let h be as in Lemma 26. Define $S : \widetilde{CC}_*(C) \to \widetilde{CC}_*(C) \otimes \widetilde{CC}_*(C)$ by*

$$S(\alpha) := h(\alpha) - \tau \circ h(\alpha) \quad \textit{for any } \alpha \in \widetilde{CC}_*(C).$$

Then the chain complex $(\widetilde{CC}_*(C)[2-m], b)$ is a DG Lie coalgebra with the cobracket S.

Proof. It follows from the definition that S is skew-symmetric, and b commutes with S by (31). Now, for any $\alpha = a_0[a_1|\cdots|a_n]$,

$$\begin{aligned}
&(h\otimes 1)h(\alpha) - (1\otimes h)h(\alpha)\\
&\quad = \sum_{k<l<i<j} \pm a_0[a_1|\cdots|a_{k-1}|a_{l+1}|\cdots|a_{i-1}|a_{j+1}|\cdots|a_n]\\
&\qquad\qquad\qquad \otimes a_k a_l[a_{k+1}|\cdots|a_{l-1}] \otimes a_i a_j[a_{i+1}|\cdots|a_{j-1}]\\
&\qquad + \sum_{i<j<k<l} \pm a_0[a_1|\cdots|a_{i-1}|a_{j+1}|\cdots|a_{k-1}|a_{l+1}|\cdots|a_n]\\
&\qquad\qquad\qquad \otimes a_k a_l[a_{k+1}|\cdots|a_{l-1}] \otimes a_i a_j[a_{i+1}|\cdots|a_{j-1}]\\
&\quad = (1\otimes\tau)((h\otimes 1)h(\alpha) - (1\otimes h)h(\alpha)).
\end{aligned}$$

It follows that

$$\begin{aligned}
&(1+\sigma+\sigma^2)(S\otimes 1)S\\
&\quad = (1+\sigma+\sigma^2)\big((h\otimes 1)h - (1\otimes h)h - (1\otimes\tau)((h\otimes 1)h - (1\otimes h)h)\big) = 0,
\end{aligned}$$

so the co-Jacobi identity holds. □

It follows that $(\widetilde{HH}_*(C)[2-m], S)$ is a graded Lie coalgebra. The Lie cobracket S and the cocommutative coproduct $\vee$ are compatible in the following sense:

Definition 28. Let V be a graded vector space. A *Gerstenhaber coalgebra* on V is a triple $(V, \vee, S)$ such that

(i) $(V, \vee)$ is a graded cocommutative coalgebra;

(ii) (V, S) is a graded Lie coalgebra whose Lie cobracket is of degree 1; and

(iii) $S : V \to V \otimes V$ is a coderivation with respect to $\vee$, that is, the following diagram commutes:

$$\begin{tikzcd}
V \arrow[r, "\vee"] \arrow[d, "S"'] & V\otimes V \arrow[d, "(\mathrm{id}\otimes\tau)\circ(S\otimes\mathrm{id})+\mathrm{id}\otimes S"] \\
V\otimes V \arrow[r, "\vee\otimes\mathrm{id}"] & V\otimes V\otimes V
\end{tikzcd}$$

Theorem 29. *The reduced Hochschild homology* $(\widetilde{HH}_*(C)[1-m], \vee, S)$ *is a Gerstenhaber coalgebra.*

Proof. From the definition of h in Lemma 26, the diagram

$$\begin{tikzcd}
\widetilde{CC} \arrow[r, "\vee"] \arrow[d, "h"'] & \widetilde{CC}\otimes\widetilde{CC} \arrow[d, "(\mathrm{id}\otimes\tau)\circ(h\otimes\mathrm{id})+\mathrm{id}\otimes h"] \\
\widetilde{CC}\otimes\widetilde{CC} \arrow[r, "\vee\otimes\mathrm{id}"] & \widetilde{CC}\otimes\widetilde{CC}\otimes\widetilde{CC}
\end{tikzcd}$$

commutes. We next show that

(40)
$$\begin{tikzcd}
\widetilde{CC} \arrow[r, "\vee"] \arrow[d, "\tau\circ h"'] & \widetilde{CC}\otimes\widetilde{CC} \arrow[d, "(\mathrm{id}\otimes\tau)\circ(\tau\circ h\otimes\mathrm{id})+\mathrm{id}\otimes\tau\circ h"] \\
\widetilde{CC}\otimes\widetilde{CC} \arrow[r, "\vee\otimes\mathrm{id}"] & \widetilde{CC}\otimes\widetilde{CC}\otimes\widetilde{CC},
\end{tikzcd}$$

commutes up to homotopy, and therefore, from $S = h - \tau\circ h$, the diagram

$$\begin{tikzcd}
\widetilde{HH} \arrow[r, "\vee"] \arrow[d, "S"'] & \widetilde{HH}\otimes\widetilde{HH} \arrow[d, "(\mathrm{id}\otimes\tau)\circ(S\otimes\mathrm{id})+\mathrm{id}\otimes S"] \\
\widetilde{HH}\otimes\widetilde{HH} \arrow[r, "\vee\otimes\mathrm{id}"] & \widetilde{HH}\otimes\widetilde{HH}\otimes\widetilde{HH}
\end{tikzcd}$$

commutes. Let $\varrho : \widetilde{CC} \to \widetilde{CC}\otimes\widetilde{CC}\otimes\widetilde{CC}$ be the map defined by

$$\varrho(\alpha) := \sum_{i<j<k<l} \pm a_0[a_1|\cdots|a_{i-1}|a_{j+1}|\cdots|a_{k-1}|a_{l+1}|\cdots|a_n] \otimes a_ia_j[a_{i+1}|\cdots|a_{j-1}] \otimes a_ka_l[a_{k+1}|\cdots|a_{l-1}],$$

for any $\alpha = a_0[a_1|\cdots|a_n]$. (The ± sign is determined by the Koszul sign rule.) Let $\rho := \sigma\circ\varrho$. Then

$$(b\circ\rho-\rho\circ b)(\alpha) = ((\vee\otimes\mathrm{id})\circ(\tau\circ h) - ((\mathrm{id}\otimes\tau)\circ(\tau\circ h\otimes\mathrm{id})+\mathrm{id}\otimes\tau\circ h)\circ\vee)(\alpha) \tag{41}$$

for any $\alpha\in\widetilde{CC}$. Indeed, one has

$$\varrho\circ b(\alpha) - b\circ\varrho(\alpha) =$$

$$\sum_{i<j<k} \pm (a_0a_i)'[a_{i+1}|\cdots|a_{j-1}|a_{k+1}|\cdots|a_n] \otimes (a_0a_i)''[a_1|\cdots|a_{i-1}] \otimes (a_ja_k)[a_{j+1}|\cdots|a_{k-1}] \tag{42}$$

$$+\sum_{j<i<k} \pm a_0[a_1|\cdots|a_{j-1}|a_{k+1}|\cdots|a_n] \otimes (a_ja_ka_i)'[a_{j+1}|\cdots|a_{i-1}] \otimes (a_ja_ka_i)''[a_{i+1}|\cdots|a_{k-1}] \tag{43}$$

$$+\sum_{j<k<i} \pm (a_0a_i)'[a_1|\cdots|a_{j-1}|a_{k+1}|\cdots|a_{i-1}] \otimes (a_ja_k)[a_{j+1}|\cdots|a_{k-1}] \otimes (a_0a_i)''[a_{i+1}|\cdots|a_n]. \tag{44}$$

After applying σ, (42) becomes $(\mathrm{id}\otimes\tau\circ h)\circ\vee(\alpha)$, (43) becomes $(\vee\otimes\mathrm{id})\circ(\tau\circ h)(\alpha)$, and (44) becomes $(\mathrm{id}\otimes\tau)\circ(\tau\circ h\otimes\mathrm{id})\circ\vee(\alpha)$. This proves the identity (41), and hence (40) is proved. □

Theorem 8 follows from the dual version of [Getzler 1994a, Proposition 1.2], Theorem 29, and the following lemma.

Lemma 30. *For any* $\alpha = a_0[a_1|\cdots|a_n]\in\widetilde{CC}_*(C)$, *let*

$$\phi(\alpha):=\sum_{i<j<k}\pm\varepsilon(a_0)a_i[a_{i+1}|\cdots|a_{j-1}|a_{k+1}|\cdots|a_n|a_1|\cdots|a_{i-1}]\otimes a_ja_k[a_{j+1}|\cdots|a_{k-1}], \tag{45}$$

$$\psi(\alpha):=\sum_{j<k<i}\pm\varepsilon(a_0)a_ja_k[a_{j+1}|\cdots|a_{k-1}]\otimes a_i[a_{i+1}|\cdots|a_n|a_1|\cdots|a_{j-1}|a_{k+1}|\cdots|a_{i-1}], \tag{46}$$

and let $\theta=\phi+\psi$. (*The ± signs are determined by the Koszul sign rule.*) *Then*

$$b\circ\theta+\theta\circ b=\vee\circ B-B\circ\vee-S,$$

where S is as defined in Lemma 27.

Proof. The proof is similar to that of Lemma 23. For any $\alpha=a_0[a_1|\cdots|a_n]$, the terms on the right hand side of the desired equation are

$\vee\circ B(\alpha)=$

$$\sum_{i>j}\pm\varepsilon(a_0)(a_ia_j)'[a_{i+1}|\cdots|a_n|a_1|\cdots|a_{j-1}]\otimes(a_ia_j)''[a_{j+1}|\cdots|a_{i-1}] \tag{47}$$

$$+\sum_{i<j}\pm\varepsilon(a_0)(a_ia_j)'[a_{i+1}|\cdots|a_{j-1}]\otimes(a_ia_j)''[a_{j+1}|\cdots|a_n|a_1|\cdots|a_{i-1}], \tag{48}$$

$B\circ\vee(\alpha)=$

$$\sum_{i>k}\pm a_k[a_{k+1}|\cdots|a_{i-1}|a_1|\cdots|a_{k-1}]\otimes a_0a_i[a_{i+1}|\cdots|a_n] \tag{49}$$

$$+\sum_{i<k}\pm a_0a_i[a_1|\cdots|a_{i-1}]\otimes a_k[a_{k+1}|\cdots|a_n|a_{i+1}|\cdots|a_{k-1}], \tag{50}$$

$S(\alpha)=$

$$\sum_{i<j}\pm a_0[a_1|\cdots|a_{i-1}|a_{j+1}|\cdots|a_n]\otimes a_ia_j[a_{i+1}|\cdots|a_{j-1}] \tag{51}$$

$$+\sum_{i<j}\pm a_ia_j[a_{i+1}|\cdots|a_{j-1}]\otimes a_0[a_1|\cdots|a_{i-1}|a_{j+1}|\cdots|a_n]. \tag{52}$$

It follows that

$$\phi \circ b(\alpha) = -b \circ \phi(\alpha) + (47) - (49) - (51),$$

while

$$\psi \circ b(\alpha) = -b \circ \psi(\alpha) + (48) - (50) - (52). \qquad \square$$

6. The gravity coalgebra

Corollary 9 is immediate from Theorem 8 and the following result.

Proposition 31. *Let $(V, \vee, \Delta)$ be a Batalin–Vilkovisky coalgebra, and let W be a graded vector space. Let $\mathrm{E} : V_* \to W_*$ and $\mathrm{M} : W_* \to V_{*+1}$ be two maps such that $\mathrm{E} \circ \mathrm{M} = 0$ and $\mathrm{M} \circ \mathrm{E} = \Delta$. Define $s_n : W \to W^{\otimes n}$ for $n \geq 2$ by*

$$s_n(\alpha) := (\mathrm{E} \otimes \cdots \otimes \mathrm{E}) \circ (\vee \otimes \mathrm{id}^{\otimes n-2}) \circ \cdots \circ \vee \circ \mathrm{M}(\alpha)$$

for any $\alpha \in W$. Then $(W[1], \{s_n\})$ is a gravity coalgebra.

Proof. The proof is analogous to that of Proposition 24. By induction on n, we deduce from the identity in Definition 14 that

$$\vee_n \circ \Delta(x) - (n-2)\left(\sum_{i=1}^{n-1} \mathrm{id}^{\otimes i} \otimes \Delta \otimes \mathrm{id}^{\otimes n-i-1}\right) \circ \vee_n(x) \\ = S_{2,n-2} \circ (\vee \circ \Delta \otimes \mathrm{id}^{\otimes n-2}) \circ \vee_{n-1}(x), \tag{53}$$

for all $x \in V$, where we set $\vee_n := (\vee \otimes \mathrm{id}^{\otimes n-2}) \circ \cdots \circ \vee : V \to V^{\otimes n}$ as before.

Let $x = \mathrm{M}(\alpha)$ where $\alpha \in W$. Applying $\mathrm{E}^{\otimes n}$ to both sides of (53), we get

$$\mathrm{E}^{\otimes n} \circ \left(\vee_n \circ \Delta(\mathrm{M}(\alpha)) - (n-2) \sum_{i=1}^{n-1} (\mathrm{id}^{\otimes i} \otimes \Delta \otimes \mathrm{id}^{\otimes n-i-1}) \circ \vee_n(\mathrm{M}(\alpha))\right) \\ = \mathrm{E}^{\otimes n} \circ S_{2,n-2} \circ (\vee \circ \Delta \otimes \mathrm{id}^{\otimes n-2}) \circ \vee_{n-1}(\mathrm{M}(\alpha)),$$

where the left side vanishes since $\Delta = \mathrm{M} \circ \mathrm{E}$ and $\mathrm{E} \circ \mathrm{M} = 0$. Hence, we have

$$\begin{aligned} 0 &= \mathrm{E}^{\otimes n} \circ S_{2,n-2} \circ (\vee \circ \Delta \otimes \mathrm{id}^{\otimes n-2}) \circ \vee_{n-1}(\mathrm{M}(\alpha)) \\ &= \mathrm{E}^{\otimes n} \circ S_{2,n-2} \circ (\vee \circ \mathrm{M} \circ \mathrm{E} \otimes \mathrm{id}^{\otimes n-2}) \circ \vee_{n-1}(\mathrm{M}(\alpha)) \\ &= S_{2,n-2} \circ (\mathrm{E}^{\otimes 2} \circ \vee \circ (\mathrm{M} \circ \mathrm{E}) \otimes \mathrm{E}^{\otimes n-2}) \circ \vee_{n-1}(\mathrm{M}(\alpha)) \\ &= S_{2,n-2} \circ (s_2 \otimes \mathrm{id}^{\otimes n-2}) \circ s_{n-1}(\alpha). \end{aligned}$$

This proves the identity (6) in the definition of a gravity coalgebra for the case $l = 0$.

Now let $l > 0$. Let $x = \mathrm{M}(\alpha)$ where $\alpha \in W$ and suppose

$$\vee_{l+1}(x) = x_1 \otimes \cdots \otimes x_{l+1}.$$

Applying the identity (53) to the first component on both sides, by the same argument as above, we obtain

$$S_{2,n-2} \circ (s_2 \otimes \mathrm{id}^{\otimes n-2}) \circ s_{n-1+l}(\alpha) = (s_n \otimes \mathrm{id}^{\otimes l}) \circ s_{l+1}(\alpha).$$

This proves the identity (6) for the case $l > 0$. □

Acknowledgments

We thank F. Eshmatov for many discussions.

References

[Chas and Sullivan 1999] M. Chas and D. Sullivan, "String topology", preprint, 1999. arXiv math/9911159

[Chas and Sullivan 2004] M. Chas and D. Sullivan, "Closed string operators in topology leading to Lie bialgebras and higher string algebra", pp. 771–784 in *The legacy of Niels Henrik Abel* (Oslo, 2002), edited by O. A. Laudal and R. Piene, Springer, Berlin, 2004. MR 2005f:55007 Zbl 1068.55009

[Chen 2007] X. Chen, "An algebraic chain model of string topology", preprint, version 3, 2007. arXiv 0708.1197v3

[Cohen and Godin 2004] R. L. Cohen and V. Godin, "A polarized view of string topology", pp. 127–154 in *Topology, geometry and quantum field theory*, edited by U. Tillmann, London Math. Soc. Lecture Note Ser. **308**, Cambridge Univ. Press, 2004. MR 2005m:55014 Zbl 1095.55006

[Cohen and Jones 2002] R. L. Cohen and J. D. S. Jones, "A homotopy theoretic realization of string topology", *Math. Ann.* **324**:4 (2002), 773–798. MR 2004c:55019 Zbl 1025.55005

[Eu and Schedler 2009] C.-H. Eu and T. Schedler, "Calabi–Yau Frobenius algebras", *J. Algebra* **321**:3 (2009), 774–815. MR 2009m:16019 Zbl 05549335

[Félix and Thomas 2008] Y. Félix and J.-C. Thomas, "Rational BV-algebra in string topology", *Bull. Soc. Math. France* **136**:2 (2008), 311–327. MR 2009c:55015 Zbl 1160.55006

[Gerstenhaber 1963] M. Gerstenhaber, "The cohomology structure of an associative ring", *Ann. of Math.* (2) **78** (1963), 267–288. MR 28 #5102 Zbl 0131.27302

[Getzler 1994a] E. Getzler, "Batalin–Vilkovisky algebras and two-dimensional topological field theories", *Comm. Math. Phys.* **159**:2 (1994), 265–285. MR 95h:81099 Zbl 0807.17026

[Getzler 1994b] E. Getzler, "Two-dimensional topological gravity and equivariant cohomology", *Comm. Math. Phys.* **163**:3 (1994), 473–489. MR 95h:81100 Zbl 0806.53073

[Getzler 1995] E. Getzler, "Operads and moduli spaces of genus 0 Riemann surfaces", pp. 199–230 in *The moduli space of curves* (Texel Island, 1994), edited by R. Dijkgraaf et al., Progr. Math. **129**, Birkhäuser, Boston, MA, 1995. MR 96k:18008 Zbl 0851.18005

[Ginzburg 2006] V. Ginzburg, "Calabi–Yau algebras", preprint, 2006. arXiv math/0612139

[Godin 2007] V. Godin, "Higher string topology operations", preprint, 2007. arXiv 0711.4859

[Jones 1987] J. D. S. Jones, "Cyclic homology and equivariant homology", *Invent. Math.* **87**:2 (1987), 403–423. MR 88f:18016 Zbl 0644.55005

[Lambrechts and Stanley 2008] P. Lambrechts and D. Stanley, "Poincaré duality and commutative differential graded algebras", *Ann. Sci. Éc. Norm. Supér.* (4) **41** (2008), 495–509. MR 2009k:55022 Zbl 1172.13009

[Menichi 2004] L. Menichi, “Batalin–Vilkovisky algebras and cyclic cohomology of Hopf algebras”, *K-Theory* **32**:3 (2004), 231–251. MR 2006c:16018 Zbl 1101.19003

[Sullivan 2004] D. Sullivan, “Open and closed string field theory interpreted in classical algebraic topology”, pp. 344–357 in *Topology, geometry and quantum field theory*, edited by U. Tillmann, London Math. Soc. Lecture Note Ser. **308**, Cambridge Univ. Press, 2004. MR 2005g:81289 Zbl 1088.81082

[Tradler 2008] T. Tradler, “The Batalin–Vilkovisky algebra on Hochschild cohomology induced by infinity inner products”, *Ann. Inst. Fourier* (*Grenoble*) **58**:7 (2008), 2351–2379. MR 2010a:16020 Zbl 05505486

[Westerland 2008] C. Westerland, “Equivariant operads, string topology, and Tate cohomology”, *Math. Ann.* **340**:1 (2008), 97–142. MR 2008k:55014 Zbl 1141.55002

Received June 25, 2009. Revised September 7, 2009.

XIAOJUN CHEN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MI 48109
UNITED STATES

xch@umich.edu

WEE LIANG GAN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
RIVERSIDE, CA 92521
UNITED STATES

wlgan@math.ucr.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

INVARIANT FINSLER METRICS ON POLAR HOMOGENEOUS SPACES

SHAOQIANG DENG

We study invariant Finsler metrics on polar homogeneous manifolds. After establishing existence results, we prove that an invariant Finsler metric on a nonsymmetric polar homogeneous manifold of a simply connected compact simple Lie group is Berwaldian if and only if it is Riemannian. As an application, we prove that on each such manifold with generalized rank of at least 2, there exist infinitely many invariant Finsler metrics that are reversible, non-Berwaldian and of vanishing S-curvature; this kind of space is sought after in an open problem of Shen. Finally, using one type of polar homogeneous manifold, we give a classification of homogeneous Randers spaces with positive constant flag curvature.

Introduction

A fundamental problem in Riemann–Finsler geometry is that of classifying the Finsler metrics on a given manifold. In full generality, this problem is intractable, so we must focus on metrics with certain special properties (particularly curvature properties), such as spaces of constant flag curvature and spaces with isotropic S-curvature. One of the most important advances has been the classification of Randers metrics with constant flag curvature obtained by Bao et al. [2004]. Also important is the work of Szabó [1981; 2006] on symmetric Berwald spaces.

Here we consider this problem for invariant Finsler metrics on homogeneous manifolds. More precisely, let G be a Lie group and H be a closed subgroup of G. Then the coset space G/H admits a (unique) differentiable structure such that the action of G on G/H is smooth; that is, G can be viewed as a Lie transformation group on the manifold G/H. Our goal is to classify the G-invariant Finsler metrics on G/H and study the geometrical properties of such metrics. In previous work, we have obtained some partial results. For example, in [Deng and Hou 2004a], we proved that there exist invariant non-Riemannian Finsler metrics on G/H, provided

MSC2000: 22E46, 53C12, 53C60.

Keywords: polar actions, Minkowski representations, generalized rank, Berwald spaces, Randers metrics.

Supported by NSFC of China (grant 10971104).

H is compact and the action of H on the tangent space of G/H at the origin $o = eH$ (that is, the action of the linear isotropic representation) is not irreducible; in other words, we assumed that there exist nontrivial invariant subspaces of H. However, when the linear isotropic representation is irreducible, the situation is very complicated. For example, if (G, H) is a Riemannian symmetric pair and the symmetric space G/H is irreducible of rank 1, then any G-invariant Finsler metric on G/H must be Riemannian. However, if the rank is at least 2, then there exist infinitely many G-invariant Finsler metrics on G/H that are non-Riemannian [Szabó 1981]. Therefore it is interesting to find the conditions under which a coset space with irreducible linear isotropic representation has invariant non-Riemannian Finsler metrics and to classify those metrics.

Coset spaces with irreducible linear isotropic representation are called isotropic irreducible homogeneous spaces. Wolf [1968; 1977] has studied the interesting geometry of these manifolds extensively. It is known that a connected, simply connected, noncompact, isotropic irreducible homogeneous space is either flat or a Riemannian symmetric space [Besse 1987]. Therefore we are only interested in the compact case. In this case, a classification of a special type of such spaces (strongly isotropic irreducible homogeneous spaces) was obtained independently by Manturov [1961a; 1961b; 1966], Wolf [1968], and Krámer [1975]. Wang and Ziller [1991] studied a more generalized class of Riemannian spaces: isotropy irreducible Riemannian spaces. It turns out that many such spaces are nonsymmetric. So, our first step is to classify the invariant Finsler metrics on compact isotropic irreducible homogeneous manifolds.

A deep analysis of this problem shows that a general classification is unreachable even if we confine ourselves to the isotropic irreducible homogeneous spaces. However, the situation simplifies if the isotropic representation is polar, meaning that there exists a subspace of the tangent space that intersects every orbit of the isotropic group and does so perpendicularly at any intersection. Then the algebraic methods of representation theory are available, and we can obtain a satisfactory classification theorem.

Our main results can be summarized as follows: We first use the notion of Minkowski representations of Lie groups to refine Szabó's result on the existence of invariant non-Riemannian Finsler metrics on Riemannian symmetric spaces to establish a bijection between the invariant Finsler metrics on a polar homogeneous space and the Weyl-invariant Minkowski norms on a generalized Cartan space. In the compact case, we study the geometric properties of such metrics. In particular, we prove that an invariant Finsler metric on a nonsymmetric polar homogeneous manifold of a simply connected compact simple Lie group is Berwaldian if and only if it is Riemannian. As an application, we prove that on each nonsymmetric polar homogeneous manifold of a compact simple Lie group with generalized rank of at least 2, there exist infinitely many invariant Finsler metrics that are reversible, non-Berwaldian and have vanishing S-curvature. Finally, using one type of polar homogeneous manifold, we obtain a classification of homogeneous Randers spaces with positive constant flag curvature. The bijection established here is new, and can be viewed as a way to classify invariant Finsler metrics on polar homogeneous spaces. We also mention that Szabó gave a classification of invariant Berwald metrics on Riemannian symmetric spaces by explicit construction via Chevalley polynomials. It is an interesting problem to consider the generalization of his method to the polar cases.

This paper is organized as follows. In Sections 1 and 2, we recall the general properties of Minkowski representations and polar representations. Section 3 gives our classification of invariant Finsler metrics on polar homogeneous manifolds, while Section 4 classifies invariant Finsler metrics in general. In Section 5, we study the geometrical properties of such metrics. In Section 6, we give a complete classification of invariant Randers metrics on polar homogeneous manifolds.

1. Minkowski representations of Lie groups

Definition 1.1. Let V be an n-dimensional real vector space. A Minkowski norm on V is a functional F on V that is smooth on $V\setminus\{0\}$ and satisfies these conditions:

- $F(u) \geq 0$ for all $u \in V$.
- $F(\lambda u) = \lambda F(u)$ for all $\lambda > 0$.
- For any basis $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ of V, write $F(y) = F(y^1, y^2, \dots, y^n)$, where $y = y^j \varepsilon_j$. Then the Hessian matrix
$$(g_{ij}) := \left(\left[\tfrac{1}{2}F^2\right]_{y^i y^j}\right)$$
is positive definite at any point of $V \setminus \{0\}$.

If V is a real vector space endowed with a Minkowski norm F, then (V, F) is called a Minkowski space. Minkowski spaces play a role in Finsler geometry analogous to the role Euclidean spaces play in Riemannian geometry. In fact, a Finsler space is just a smooth manifold endowed with a smoothly varying family of Minkowski norms on its tangent spaces. Unlike in Riemannian manifolds, in a Finsler space the Minkowski norms in different tangent spaces may not be linearly isomorphic to each other.

Definition 1.2. Let G be a Lie group, and (V, ρ) a (real) representation of G. If F is a Minkowski norm on V such that
$$F(\rho(g)v) = F(v) \quad \text{for all } g \in G \text{ and } v \in V,$$
then we call (V, ρ, F) a Minkowski representation of G.

The notion of Minkowski representations of Lie groups is a natural and obvious generalization of orthogonal representations.

Proposition 1.3 [Deng and Hou 2004a]. *Suppose G/H is a coset space of a Lie group G. Then there exists a bijection between the invariant Finsler metric F on G/H and the Minkowski norm F_0 on $\mathfrak{g}/\mathfrak{h}$, such that* $(\mathfrak{g}/\mathfrak{h}, \mathrm{Ad}, F_0)$ *is a Minkowski representation of H, where $\mathfrak{g}$ and $\mathfrak{h}$ are the Lie algebras of G and H and* Ad *is the adjoint action of H on $\mathfrak{g}/\mathfrak{h}$.*

Does a given homogeneous manifold admit an invariant non-Riemannian Finsler metric? In [Deng and Hou 2004a], we proved that it does if the adjoint action of H on $\mathfrak{g}/\mathfrak{h}$ is not irreducible. When the adjoint action is irreducible, the situation is more complicated, and in general, a necessary and sufficient condition seems to be unattainable. However, we can give a complete answer when the action of H on $\mathfrak{g}/\mathfrak{h}$ is polar. Let us first recall some definitions in the next section.

2. Polar actions of Lie groups

Let G be a compact Lie group with Lie algebra $\mathfrak{g}$ and a real representation (ρ, V). By Weyl's unitary trick, there is an inner product $\langle \cdot, \cdot \rangle$ that is invariant under $\rho(g)$ for all $g \in G$. Therefore, we get a continuous homomorphism ρ from G to $O(V)$, where $O(V)$ is the orthogonal group with respect to $\langle \cdot, \cdot \rangle$. In many situations, we hope to find a linear subspace of V that intersects every orbit of the G-action and is of minimal possible dimension. In studying G-invariant differential equations or differential operators, such a subspace can be used for reduction of variables. It is also useful in analyzing orbit structure, which is important in geometry.

The existence of such a cross section comes from a simple fact: As pointed out in [Dadok 1985], if for $v \in V$ we let $a_v = \{u \in V \mid \langle u, \mathfrak{g} \cdot v \rangle = 0\} = (\mathfrak{g} \cdot v)^{\perp}$, then the linear space a_v intersects every G-orbit. The action of $\mathfrak{g}$ on V is the differential of that of G, and hence $\mathfrak{g} \cdot v$ is just the tangent space to the G-orbit through v. To obtain a cross section of minimal dimension, we only need to choose v on an orbit of maximal dimension. A vector v in V is called regular if $\mathfrak{g} \cdot v$ is of maximal possible dimension. In some special cases, we may choose a cross section that intersects each orbit orthogonally.

Definition 2.1 [Dadok 1985]. A representation $\rho : G \to O(V)$ is called polar if it satisfies any of the following equivalent conditions:

- For any regular elements v_1 and v_2, we have $\mathfrak{g} \cdot v_1 = k \cdot (\mathfrak{g} \cdot v_2)$ for some $k \in G$.
- For any regular elements v_1 and v_2, we have $a_{v_1} = k \cdot a_{v_2}$ for some $k \in G$.
- For any regular element $v \in V$ and $u \in a_v$, the scalar product $\langle \mathfrak{g} \cdot u, a_v \rangle$ vanishes.

For a polar representation (ρ, V) of G, any minimal linear cross section a_v is called a Cartan subspace. The dimension of $\mathfrak{a}$ is called the (generalized) rank of the representation.

The notion of a polar representation is closely related to that of a polar action of a compact Lie group on a Riemannian manifold. An isometric action of a compact Lie group G on a Riemannian manifold M is called polar if there exists a closed, connected submanifold Σ of M that meets all G-orbits and meets these orbits orthogonally. Any such Σ is called a section of the action. A section is necessarily totally geodesic in M. If the section is flat in the induced Riemannian metric, then the action is called hyperpolar. Therefore, a polar representation is just a hyperpolar action consisting of linear isometries on a Euclidean space.

Now we recall some known results and terminology of polar representations. Let $\mathfrak{a}$ be a fixed Cartan subspace, and let $N_G(\mathfrak{a})$ and $Z_G(\mathfrak{a})$ be respectively the normalizer and centralizer of $\mathfrak{a}$ in G:

$$N_G(\mathfrak{a}) = \{g \in G \mid \rho(g)(\mathfrak{a}) \subset \mathfrak{a}\}, \quad Z_G(\mathfrak{a}) = \{g \in G \mid \rho(g)(X) = X \text{ for } X \in \mathfrak{a}\}.$$

Then the Lie algebras of $N_G(\mathfrak{a})$ and $Z_G(\mathfrak{a})$ coincide by [Dadok 1985]. Hence $W = N_G(\mathfrak{a})/Z_G(\mathfrak{a})$ is a finite group acting on $\mathfrak{a}$. W is called the Weyl group of the representation.

Theorem 2.2 [Dadok 1985]. *Let $\rho : G \to O(V)$ be a polar representation, and let $\mathfrak{a}$ be a Cartan subspace. Then for any $x \in V$, the orbit $G \cdot x$ intersects $\mathfrak{a}$ at finitely many points and the set of intersections comprises a single W-orbit.*

Definition 2.3. A symmetric space representation of a connected compact Lie group G (with Lie algebra $\mathfrak{g}$) is an orthogonal representation $\rho : G \to \mathrm{SO}(V)$ such that there exists a noncompact real Lie algebra $\mathfrak{g}_1$ with a Cartan decomposition $\mathfrak{g}_1 = \mathfrak{k}_1 + \mathfrak{m}_1$, a Lie algebra isomorphism $A : \mathfrak{g} \to \mathfrak{k}_1$, and a real vector space isomorphism $L : V \to \mathfrak{m}_1$ such that $L \circ \rho(X)(y) = [A(X), L(y)]$ for all $X \in \mathfrak{g}$ and $y \in V$.

Theorem 2.4 [Dadok 1985, Proposition 6]. *Let $\rho : G \to \mathrm{SO}(V)$ be a polar representation of a connected compact Lie group G. Then there exist a connected compact Lie group G_1 and a symmetric space representation $\rho_1 : G_1 \to \mathrm{SO}(V)$ such that the G- and G_1-orbits coincide.*

Remark 2.5. The fact that the G- and G_1-orbits coincide does not mean that they are equivalent representations. In fact, the classification of polar representations implies there are irreducible polar representations that are not equivalent to the isotropic representation of a Riemannian symmetric space; see Section 6.

The following result will be useful in proving the main results of Section 3.

Proposition 2.6. *Let $\rho : G \to \mathrm{SO}(V)$ be a symmetric space representation of a connected compact Lie group G. On the vector space $\bar{\mathfrak{g}} = \mathfrak{g} + V$ (direct sum), there is a Lie algebra structure that makes $\bar{\mathfrak{g}}$ into a noncompact semisimple Lie algebra and such that $\bar{\mathfrak{g}} = \mathfrak{g} + V$ is a Cartan decomposition. Further, there exists a Riemannian globally symmetric space $\overline{G}/\overline{K}$ of noncompact type such that $\bar{\mathfrak{g}} = \mathfrak{g}+V$ is the canonical decomposition of the corresponding Lie algebra, and the G-orbit in V coincides with the $\overline{K}$-orbit in V of the adjoint representation of $\overline{K}$ on V.*

Proof. Let $\mathfrak{g}_1$, $\mathfrak{k}_1$, $\mathfrak{m}_1$, A, L be as in Definition 2.3. In the vector space $\bar{\mathfrak{g}} = \mathfrak{g} + V$ (where the addition is direct sum of subspaces), we introduce a bracket: On $\mathfrak{g}$ we take the same bracket operation as that of the Lie algebra. For $x, y \in V$, the bracket is equal to $A^{-1}([L(x), L(y)])$, where the Lie bracket is the same as that of $\mathfrak{g}_1$. (Note that $L(x), L(y)$ are contained in $\mathfrak{m}_1$, so $[L(x), L(y)]$ lies in $\mathfrak{k}_1$ and $A^{-1}([L(x), L(y)])$ is contained in $\mathfrak{g}$.) For $X \in \mathfrak{g}, x \in V$, we define $[X, x] = L^{-1}([A(X), L(x)])$, which is an element in V. It can be checked directly (albeit nontrivially) that $\bar{\mathfrak{g}}$ with the brackets above forms a Lie algebra that is isomorphic to $\mathfrak{g}_1$. Since $\mathfrak{g}_1$ is a noncompact semisimple Lie algebra, so is $\bar{\mathfrak{g}}$, and since $\mathfrak{g}_1 = \mathfrak{k}_1 + \mathfrak{m}_1$ is a Cartan decomposition, so is $\bar{\mathfrak{g}} = \mathfrak{g} + V$. Now, according to the theory of orthogonal symmetric Lie algebras [Helgason 1978], there exists a Riemannian symmetric pair $(\overline{G}, \overline{K})$ of noncompact type with $\operatorname{Lie}\overline{G} = \bar{\mathfrak{g}}$ and $\operatorname{Lie}\overline{K} = \mathfrak{g}$. Moreover, by the definition of the brackets, the differential of the adjoint representation of $\overline{K}$ on V is just the induced action of $\mathfrak{g}$ on V of the representation ρ of G. Since ρ is an orthogonal representation, the inner product $\langle\cdot,\cdot\rangle$ on V is invariant under the action of G. Hence, $\rho(X)$ is skew-symmetric with respect to $\langle\cdot,\cdot\rangle$ for any $X \in \mathfrak{g}$. Since in the Riemannian symmetric pair of noncompact type, the subgroup must be compact and connected [Helgason 1978], $\overline{K}$ must be a connected compact Lie group. This implies that the exponential map of $\mathfrak{g}$ to $\overline{K}$ must be surjective [Kobayashi and Nomizu 1963]. Thus $\langle\cdot,\cdot\rangle$ is also invariant under the action of $\overline{K}$. Hence, there is a G_2-invariant Riemannian metric Q on G_2/K_2 whose restriction on $T_o(\overline{G}/\overline{K}) = V$ is $\langle\cdot,\cdot\rangle$, where o is the origin of $\overline{G}/\overline{K}$. Then $(\overline{G}/\overline{K}, Q)$ is a Riemannian globally symmetric space [Helgason 1978]. Now the proposition follows from the facts that the exponential map of G is surjective, and that the differentials of ρ and of the adjoint representation of K_2 on V coincide. □

3. Classification of Minkowski representations associated with a polar representation

In the two theorems below, let G be a compact connected Lie group with a polar representation (V, ρ).

Theorem 3.1. *Let $\mathfrak{a} \subset V$ be a Cartan subspace and W be the corresponding Weyl group. Then there exists a bijection between the set of Minkowski norms on V that* *make* (V, ρ, F) *a Minkowski representation, and the set of* W*-invariant Minkowski norms on* $\mathfrak{a}$.

Theorem 3.2.

- *If the generalized rank of* (V, ρ) *is* 1, *then there does not exist a non-Euclidean Minkowski norm* F *on* V *such that* (V, ρ, F) *is a Minkowski representation of* G.
- *If the generalized rank of* (V, ρ) *is at least* 2, *then there exist infinitely many non-Euclidean Minkowski norms* F *on* V *such that* (V, ρ, F) *is a Minkowski representation of* G.

We remark here that Szabó's argument [1981] on the existence of invariant non-Riemannian Finsler metrics on a Riemannian symmetric space is also valid for polar homogeneous space, since on a polar homogeneous space the isotropic representation has the same orbit as that of a Riemannian symmetric space. Hence Theorem 3.2 should not be viewed as a new result. The main point here is that our refinement of Szabó's argument can be used to establish the bijection stated in Theorem 3.1. This will lead to a classification of all the invariant Finsler metrics on a polar homogeneous space; see Section 4.

To prove the two theorems above, we need several lemmas.

Lemma 3.3. *Let* G, (V, ρ), $\mathfrak{a}$ *and* W *be as in Theorem 3.1. Then any* W*-invariant Minkowski norm on* $\mathfrak{a}$ *can be uniquely extended to a* G*-invariant functional on* V *that is smooth on* $V \setminus \{0\}$.

Proof. We use the same argument as in [Szabó 2006]. Since $\mathfrak{a}$ is a Cartan subspace, according to Definition 2.1, $\mathfrak{a}$ intersects every orbit of the action of G on V. Thus for any $y \in V$, there exist a $y_a \in \mathfrak{a}$ and $g_y \in G$ (not necessarily unique) such that $g_y(y_a) = y$. We now define a functional L on V by $L(y) = F(y_a)$. Since F is W-invariant, it is easy to check that L is well defined. To prove that L is smooth on the slit space $V \setminus \{0\}$, we need a result of [Dadok 1982], which says that the extension of a smooth W-invariant function on $\mathfrak{a}$ to V is also smooth. The Minkowski norm F is only smooth on $\mathfrak{a} \setminus \{0\}$ but is continuous on the whole space $\mathfrak{a}$. Define a functional F_1 by

$$F_1(y) = e^{-1/\langle y, y\rangle} \cdot F(y) \quad \text{for } y \in V \setminus \{0\}; \quad F_1(0) = 0,$$

where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathfrak{a}$. Then F_1 is smooth on all of $\mathfrak{a}$. Since ρ is an orthogonal representation, $\langle \cdot, \cdot \rangle$ is G-invariant. Thus F_1 is also W-invariant. Also, the extension of F_1 to V is equal to $e^{-1/\langle X, X\rangle} F(X)$ on $\mathfrak{g} \setminus \{0\}$. From this the smoothness of F on $\mathfrak{g} \setminus \{0\}$ follows. The uniqueness of the extension is obvious. □

Lemma 3.3 establishes the smoothness of the extension of the W-invariant Minkowski norms. Next we consider the strong convexity of the extension. Since it is very difficult to obtain strong convexity directly, we first prove a lemma about strict convexity. For this we need some results related to Kostant's celebrated convexity theorem, whose theory we now sketch.

Let $(G/H, Q)$ be a globally Riemannian manifold, and let $(\mathfrak{g}, \sigma)$ be the corresponding orthogonal Lie algebra. Let $\mathfrak{g} = \mathfrak{h} + \mathfrak{p}$ be the canonical decomposition of the orthogonal Lie algebra. Then we can identify the tangent space $T_{eH}(G/H)$ of G/H at the origin with the space $\mathfrak{p}$. The isotropic representation of H on $T_{eH}(G/H)$ then corresponds to the adjoint representation of H on $\mathfrak{p}$. Let $\mathfrak{t}$ be a maximal commutative subspace in $\mathfrak{p}$. Then $\mathfrak{t}$ is a cross section of the action of H and it intersects every orbit orthogonally. Therefore the isotropic representation is polar. Let W be the corresponding Weyl group and π be the orthogonal projection of $\mathfrak{p}$ onto $\mathfrak{t}$.

Kostant's convexity theorem [1973]. *For any point $x \in \mathfrak{t}$, the subset $\pi(H \cdot x)$ is equal to the convex hull of the points $W \cdot x$, where $H \cdot x$ is the orbit of the point x.*

To prove the strong convexity of the extension, we still need a lemma on the convexity of the orbit of a convex domain in V. Let $\mathfrak{g}$ be a noncompact semisimple Lie algebra, let $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$ be a Cartan decomposition of $\mathfrak{g}$, and let $\mathfrak{t}$ be a maximal commutative subspace of $\mathfrak{p}$. Let $W_{\mathfrak{t}}$ be the corresponding Weyl group and $\overline{C}$ be a fixed Weyl chamber in $\mathfrak{t}$. The Cartan–Killing form B of $\mathfrak{g}$ is positive definite on $\mathfrak{p}$, so $\langle x, y\rangle = B(x, y)$ defines an inner product on $\mathfrak{p}$. The restriction of this inner product to $\mathfrak{t}$, which we still denote by $\langle\cdot,\cdot\rangle$, is $W_{\mathfrak{t}}$-invariant. The dual cone of $\overline{C}$, denoted by $\overline{C}^*$, is defined by $x \in \overline{C}^*$ if and only if $\langle x, y\rangle \geq 0$ for all $y \in \overline{C}$. A partial order can be defined on $\mathfrak{t}$ such that $x \geq y$ if and only if $x - y \in \overline{C}^*$. Then Kostant [1973] proved that $x \leq y$ for $x, y \in \overline{C}$ if and only if x lies in the convex hull of the W-orbit of y. Let H be the maximal compact subgroup of the adjoint group $\operatorname{Int}\mathfrak{g}$ of $\mathfrak{g}$. It is well known that each H-orbit in $\mathfrak{g}$ intersects $\overline{C}$ at exactly one point [Helgason 1978]. For $x \in \mathfrak{p}$, we denote by $\overline{C}(x)$ the unique element of the intersection of the orbit $G \cdot x = \{g \cdot x \mid g \in G\}$ and $\overline{C}$.

The following result is proved using Kostant's convexity theorem.

Theorem 3.4 [Tam 1998]. *For any $x, y \in \mathfrak{p}$, we have $\overline{C}(x + y) \leq \overline{C}(x) + \overline{C}(y)$.*

The following lemma will be useful in proving the main result of this paper.

Lemma 3.5. *Let D be a strictly convex domain in $\mathfrak{t}$ containing the origin, with smooth boundary S, and let D be invariant under the action of $W_{\mathfrak{t}}$. Then the orbit of D under the action of H forms a strictly convex domain in $\mathfrak{p}$.*

Proof. Since D is W-invariant, the boundary S is also $W_{\mathfrak{t}}$-invariant. Define a nonnegative function h_1 on $\mathfrak{t}$ by $h_1(y) = 1/t$ where $t > 0$ is such that $ty \in S$. Then h_1 is smooth on $\mathfrak{t} \setminus \{0\}$, and $h_1(\lambda y) = \lambda h_1(y)$ for any $\lambda > 0$. Also, h_1 satisfies the triangle inequality: $h_1(x+y) \leq h_1(x)+h_1(y)$ with the equality holding if and only if $x = \alpha y$ or $y = \alpha x$ for some $\alpha \geq 0$ [Bao et al. 2000]. Moreover, the function h_1

is obviously $W_{\mathfrak{t}}$-invariant. Hence h_1 can be extended to a well-defined function h_2 on $\mathfrak{p}$ by defining $h_2(g \cdot y) = y$ for $g \in H$ and $y \in \mathfrak{t}$. Then it is easily seen that h_2 is H-invariant and the orbit of D forms the set $D_1 = \{y \in \mathfrak{g} \mid h_2(y) < 1\}$. Now suppose $y_1, y_2 \in \overline{D}_1$ and $0 \le \lambda \le 1$. Let $g \in H$ be such that

$$g \cdot (\lambda y_1 + (1-\lambda)y_2) = \overline{C}(\lambda y_1 + (1-\lambda)y_2).$$

Then we have

$$h_2(\lambda y_1 + (1-\lambda)y_2) = h_2(g \cdot (\lambda y_1 + (1-\lambda)y_2)) = h_1(\overline{C}(\lambda y_1 + (1-\lambda)y_2)).$$

Suppose $W_{\mathfrak{t}} = \{w_1, w_2, \dots, w_s\}$, where $s = |W_{\mathfrak{t}}|$. Then by Theorem 3.4 and Kostant's convexity theorem, there exist nonnegative numbers a_i for $i = 1, 2, \dots, s$ that sum to one and satisfy

$$\overline{C}(\lambda y_1 + (1-\lambda)y_2) = \sum_{i=1}^{s} a_i w_i(\overline{C}(\lambda y_1) + \overline{C}((1-\lambda)y_2)).$$

Hence

$$\begin{aligned}
&h_1(\overline{C}(\lambda y_1 + (1-\lambda)y_2))\\
&= h_1\Bigl(\sum_{i=1}^{s} a_i w_i(\overline{C}(\lambda y_1) + \overline{C}((1-\lambda)y_2))\Bigr) \le \sum_{i=1}^{s} h_1(a_i w_i(\overline{C}(\lambda y_1) + \overline{C}((1-\lambda)y_2)))\\
&= \sum_{i=1}^{s} a_i h_1(w_i(\overline{C}(\lambda y_1) + \overline{C}((1-\lambda)y_2))) = \sum_{i=1}^{s} a_i h_1((\overline{C}(\lambda y_1) + \overline{C}((1-\lambda)y_2)))\\
&\le \sum_{i=1}^{s} a_i (h_1(\overline{C}((\lambda y_1)) + h_1(\overline{C}(1-\lambda)y_2))) = \sum_{i=1}^{s} a_i (h_2(\lambda y_1) + h_2((1-\lambda)y_2))\\
&= \sum_{i=1}^{s} a_i (\lambda h_2(y_1) + (1-\lambda)h_2(y_2)) = \lambda h_2(y_1) + (1-\lambda)h_2(y_2) \le \lambda + (1-\lambda) = 1.
\end{aligned}$$

Thus $\lambda y_1 + (1-\lambda)y_2 \in \overline{D}_1$. Further, if $h_2(\lambda y_1 + (1-\lambda)y_2) = 1$, then from the above equation we see that either $h_2(y_1) = 1$ with $\lambda = 1$ or $h_2(y_2) = 1$ with $\lambda = 0$; that is, either $\lambda y_1 + (1-\lambda)y_2 = y_1$ or $\lambda y_1 + (1-\lambda)y_2 = y_2$. Hence the interior of the line segment joining y_1 and y_2 is contained in D_1. This proves the lemma. □

Corollary 3.6. *Let G, V, ρ and $\mathfrak{a}$ be as in Theorem 3.1. Let D be a strictly convex domain in $\mathfrak{a}$ containing the origin and invariant under the action of the Weyl group. Then the orbit of D under the action of G forms a strictly convex domain in V.*

Proof. By Dadok's result, the G-orbit coincides with that of a symmetric space representation $\rho_1 : G \to \mathrm{SO}(V)$. Then by Proposition 2.6, we can assume that $\mathfrak{a}$ is a Cartan subspace of a Riemannian symmetric space of the noncompact type, and that the action ρ_1 of G on V is exactly the isotropic representation of the Riemannian symmetric space. Now the corollary follows from Lemma 3.5. □

Proof of Theorem 3.1. If F is a Minkowski norm on V such that (V, ρ) is a Minkowski representation of G, then $F|_{\mathfrak{a}}$ is a W-invariant function and obviously a Minkowski norm on $\mathfrak{a}$. It is a direct consequence of the definition of the Cartan subspace that this correspondence is one-to-one. To prove it is surjective, let F_1 be a W-invariant Minkowski norm on $\mathfrak{a}$. For any $x \in V$, there exists $g \in G$ such that $\rho(g)(x) \in V$. We then define a function F on V by $F(x) = F_1(\rho(g)(x))$. Since F_1 is W-invariant, F is well defined. By Lemma 3.3, F is smooth on $V \setminus \{0\}$. Next we prove that F is a Minkowski norm on V. Let $\alpha_1, \dots, \alpha_m$ be an orthonormal basis of $\mathfrak{a}$ with respect to the inner product restricted to $\mathfrak{a}$, and write $F_1(z) = F_1(z^1, z^2, \dots, z^n)$ for $z = \sum_{i=1}^m z^i \alpha_i$. Then we define the Hessian matrix of F_1 by [Bao et al. 2000]

$$(a_{ij}) = \big(\big[\tfrac{1}{2} F_1^2\big]_{z^i z^j}\big).$$

For any $y \neq 0$, denote the minimal eigenvalue of the matrix $(a_{ij}(y))$ by $\mu(y)$. Let

$$\mu = \inf_{\{y \in \mathfrak{a} \mid \langle y, y\rangle = 1\}} \mu(y).$$

Since F_1 is a Minkowski norm, $\mu(y) > 0$ for any $y \in \mathfrak{a} \setminus \{0\}$. Since $\{y \in \mathfrak{a} \mid \langle y, y\rangle = 1\}$ is compact and the function $\mu(y)$ is continuous, we have $\mu > 0$. Now on $\mathfrak{a}$ we write

$$F_1(x) = \sqrt{\big(F_1^2(x) - \tfrac{1}{2}\mu\langle x, x\rangle\big) + \tfrac{1}{2}\mu\langle x, x\rangle} = \sqrt{L_*(x) + \tfrac{1}{2}\mu\langle x, x\rangle},$$

where $L_*(x) = F_1^2(x) - \frac{1}{2}\mu\langle x, x\rangle$. Since $F_1^2(x) = \sum_{i,j=1}^n a_{ij}(x) x^i x^j$ for $x \in V \setminus \{0\}$, we have

$$F_1^2(x) \geq \sum_{i=1}^n \mu(x) x^i x^i = \mu(x)\langle x, x\rangle.$$

Hence $L_*(x) > 0$ for any $x \neq 0$. Further, the Hessian matrix of $\sqrt{L_*(x)}$ is positive definite at any $x \in \mathfrak{a} \setminus \{0\}$. Hence $\sqrt{L_*}$ is also a Minkowski norm on $\mathfrak{a}$. In particular, the domain $\{y \in \mathfrak{a} \mid L_*(y) < 1\}$ is strictly convex [Bao et al. 2000]. Denote its boundary by S. Since $\langle x, x\rangle$ is W-invariant, $L_*(x)$ is also W-invariant. Therefore $L_*(x)$ can be uniquely extended to a functional L on V that is smooth on $V \setminus \{0\}$. By Lemma 3.5, the orbit of S under the action of G is the boundary of a strictly convex domain D_1 in V. But it is obvious that $D_1 = \{y \in V \mid L(y) < 1\}$. Therefore, the Hessian matrix of $\sqrt{L}$ (with respect to certain basis of V) is positive semidefinite [Bao et al. 2000]. Therefore the Hessian matrix of F is positive definite at any $x \neq 0$. Thus F is a Minkowski norm. This proves that the correspondence above is surjective. Therefore it is a bijection. □

Proof of Theorem 3.2. By Theorem 3.1, we only need to consider the W-invariant Minkowski norms on a Cartan subspace $\mathfrak{a}$.

If $\dim V = 1$, the conclusion is obvious. Therefore, we suppose that $\dim V \geq 2$.

If the generalized rank of the polar representation is 1, then we assert that the Weyl group W consists of two elements: $W = \{1, -1\}$. In fact, since the Weyl group is generated by reflections, we have only two possibilities: either $W = \{1\}$ or $W = \{1, -1\}$. If $W = \{1\}$, each W-orbit consists of only one point. Consider the unit sphere S of V, and let u and $-u$ be the unit element in $\mathfrak{a}$. Since each G-orbit intersects $\mathfrak{a}$ at one point, we have $S = G \cdot u \cup G \cdot (-u)$ and $G \cdot u \cap G \cdot (-u) = \varnothing$. Now a contradiction arises, because by the theory of Lie transformation groups, the orbits $G \cdot u$ and $G \cdot (-u)$ are connected closed submanifolds of S [Helgason 1978]. Hence $W = \{1, -1\}$. Suppose F is a W-invariant Minkowski norm on $\mathfrak{a}$. Then we have $F(x) = F(-x)$. Since $\dim \mathfrak{a} = 1$, we see that F is a Euclidean norm on $\mathfrak{a}$. Suppose $F(x) = c\sqrt{\langle x, x\rangle}$ for $x \in \mathfrak{a}$, where $\langle \cdot, \cdot \rangle$ is the G-invariant inner product on V and c is a positive constant. Then by Theorem 3.1, the extension of F to V must be equal to $c\sqrt{\langle \cdot, \cdot \rangle}$ on V. This proves the first conclusion of the theorem.

Now we suppose that the generalized rank of the polar action is at least 2. By Dadok's result and Proposition 2.6, there exists a Riemannian symmetric space G_1/H_1 of noncompact type with canonical decomposition $\mathfrak{g}_1 = \mathfrak{h}_1 + \mathfrak{p}_1$ such that V is linearly isometric to $\mathfrak{p}_1$ through a linear isometry τ, and such that the G-orbit corresponds to the H_1-orbit of the isotropic representation on $\mathfrak{p}_1$ through τ. It is easily seen that $\mathfrak{a}_1 = \tau(\mathfrak{a})$ is a maximal abelian subalgebra (that is, a Cartan subspace of the polar action of H_1 on $\mathfrak{p}_1$) of $\mathfrak{p}_1$. Let W_1 be the corresponding Weyl group. Then W-orbits correspond to W_1-orbits. Hence, to prove the theorem in this case, we only need to prove that there are infinitely many W_1-invariant Minkowski norms on $\mathfrak{a}_1$. Fix one Weyl chamber, say $\mathscr{C}$. It is known that the closure $\overline{\mathscr{C}}$ of $\mathscr{C}$ is a fundamental domain of the action of W_1. That is, every orbit intersects $\overline{\mathscr{C}}$ at exactly one point. Now we assert that there exist infinitely many functions f on $\mathscr{C}$ such that

- $f(\lambda x) = \lambda f(x)$ for all $\lambda > 0$,
- f can be extended to a W-invariant smooth function f_1 on $\mathfrak{a}_1 \setminus \{0\}$, and
- the domain $\{x \in \mathfrak{a}_1 \mid f_1(x) < 1\}$ is strictly convex.

For example, we first choose a sphere (centered at the origin) with respect to the inner product on $\mathfrak{a}_1$. Then we choose a hyperplane whose the intersection with the sphere is contained in $\mathscr{C}$. In this way, we get a hypersurface S_1, which, together with the Weyl walls of $\mathscr{C}$, bounds a strictly convex domain. The hypersurface S_1 is of course not smooth, but it is easily seen that we can make it smooth, while keeping the bounded domain strictly convex. We denote one such hypersurface by S and define a function $f(x)$ on $\mathscr{C}$ by $f(x) = \lambda$ if $x/\lambda \in S$. Then $f(x)$ satisfies the conditions above. It is obvious that there exist infinitely many functions of this type. Now, each such f can be extended to a function F_1 on $\mathfrak{a}_1$ and, similarly to the proof of Theorem 3.1, we see that the function

$$F(u)=\sqrt{F_1^2(u)+\langle u,u\rangle},$$

with $u\in\mathfrak{a}_1$, defines a W_1-invariant Minkowski norm on $\mathfrak{a}_1$. □

4. Invariant Finsler metrics: A classification

Let G be a Lie group and H be a closed subgroup of G. Then on the coset space G/H there exists a smooth structure such that G becomes a Lie transformation group of G/H. A fundamental problem in geometry to study the G-invariant geometric structures on G/H. In [Deng and Hou 2004a], we considered this problem for Finsler metrics. We proved that there is a one-to-one correspondence between the G-invariant Finsler metrics on G/H and the H-invariant Minkowski norms on the tangent space $T_o(G/H)$ of G/H at the origin $o=eH$. Therefore, to classify the G-invariant Finsler metrics, we only need to classify the Minkowski representations of H on $T_o(G/H)$.

Without losing generality, we can assume that H is a compact subgroup of G. In fact, if (M,F) is a connected homogeneous Finsler space, then the isotropic subgroup (at a fixed point $x\in M$) of the full group $I_x(M,F)$ of isometries $I(M,F)$ must be a compact subgroup of $I(M,F)$ by [Deng and Hou 2002]. Hence $M=I(M,F)/I_0(M,F)$ and F can be viewed as an $I(M,F)$-invariant Finsler metric on M. The compactness of H implies that there exist H-invariant inner products on the tangent space $T_o(G/H)$. Fix one such inner product, and denote it by $\langle\cdot,\cdot\rangle$. Then $\langle\cdot,\cdot\rangle$ can be extended to $\mathfrak{g}=\operatorname{Lie}G$ so that

$$\langle\operatorname{Ad}(h)(x),\operatorname{Ad}(h)(y)\rangle=\langle x,y\rangle\quad\text{for all } h\in H \text{ and } x,y\in\mathfrak{g}.$$

Let $\mathfrak{m}$ be the orthogonal complement of $\mathfrak{h}=\operatorname{Lie}H$. Then $\mathfrak{m}$ satisfies

$$\operatorname{Ad}(h)\mathfrak{m}\subset\mathfrak{m}\quad\text{and}\quad\mathfrak{g}=\mathfrak{h}+\mathfrak{m}. \tag{4-1}$$

Hence the tangent space $T_o(G/H)$ can be identified with $\mathfrak{m}$, and the isotropic representation of H on $T_o(G/H)$ corresponds to the adjoint representation of H on $\mathfrak{m}$.

Our goal is to a classify all the G-invariant Finsler metrics on G/H. We stress here that this problem for Riemannian metrics is easy. In fact, since H is compact, $\mathfrak{m}$ can be decomposed into a direct sum of subspaces

$$\mathfrak{m}=\mathfrak{m}_0+\mathfrak{m}_1+\cdots+\mathfrak{m}_k,$$

where $\mathfrak{m}_0$ consists of H-fixed vectors in $\mathfrak{m}$, and $\mathfrak{m}_i$ for $i=1,2,\dots,k$ are invariant, irreducible subspaces of H. For simplicity, we assume that the submodules m_i are not equivalent to each other. If there are two H-invariant inner products on $\mathfrak{m}$, then by Schur's lemma, for each i with $1 \leq i \leq k$, the restrictions of the two inner products to $\mathfrak{m}_i$ must differ only by a positive multiple. On the other hand, any inner product is H-invariant on $\mathfrak{m}_0$. Fix an H-invariant inner product $\langle\cdot,\cdot\rangle_i$ on $\mathfrak{m}_i$ for $i = 1, 2, \ldots, k$. Then any H-invariant inner product must be of the form

$$\langle\cdot,\cdot\rangle_0 + c_1\langle\cdot,\cdot\rangle_1 + \cdots + c_k\langle\cdot,\cdot\rangle_k,$$

where $\langle\cdot,\cdot\rangle_0$ is an arbitrary inner product on $\mathfrak{m}_0$, and $c_1, c_2, \ldots, c_k$ are arbitrary positive real numbers. This classifies H-invariant inner products on $\mathfrak{m}$, and hence G-invariant Riemannian metrics on G/H.

Therefore, the difficult case is that of non-Riemannian Finsler metrics. Since the general problem seems to be unsolvable, we restrict to the special case where the isotropic representation is polar. Theorems 3.1 and 3.2 give this:

Theorem 4.1. *Let G be a Lie group and H be a compact subgroup of G. Suppose that the isotropic representation of G/H at the origin $o = eH$ is polar and $\mathfrak{m}$ is as in Equation* (4-1).

- *If the generalized rank of the adjoint action of H on $\mathfrak{m}$ is* 1, *then there does not exist any non-Riemannian invariant Finsler metric on G/H.*
- *If the generalized rank of the action of H on $\mathfrak{m}$ is at least* 2, *then there exist infinitely many non-Riemannian invariant Finsler metrics on G/H* (*even if we do not distinguish those metrics that are differ only by a positive multiple*). *In this case, there is a one-to-one correspondence between the G-invariant Finsler metrics on G/H and the W-invariant Minkowski norms on $\mathfrak{a}$, where $\mathfrak{a}$ is a Cartan subspace of $\mathfrak{m}$ and W is the corresponding Weyl group.*

Now we consider some special cases of Theorem 4.1. Suppose (G, H) is a Riemannian symmetric pair. That is, G is a connected Lie group and H is a closed subgroup of G such that there exists an involutive automorphism σ such that $(G_\sigma)_e \subset H \subset G_\sigma$, where G_σ denotes the fixed points of σ and $(G_\sigma)_e$ denotes its unit components. Suppose also that the isotropic action of H at $T_o(G/H)$ leaves certain inner products invariant. We consider the classification of all G-invariant Finsler metrics on G/H. Note that in this case the isotropic action is polar.

Theorem 4.2. *Let (G, H) be a Riemannian symmetric pair. Let $\mathfrak{g} = \mathfrak{h} + \mathfrak{m}$ be the canonical decomposition of the Lie algebra of G. Fix one maximal subspace $\mathfrak{a}$ of $\mathfrak{m}$, and let W be the corresponding Weyl group. Then there exists a bijection between the G-invariant Finsler metrics on G/H and the W-invariant Minkowski norms on $\mathfrak{a}$. In particular, if the rank of G/H is* 1, *there does not exist any G-invariant non-Riemannian Finsler metric on G/H; if the rank of G/H is at least* 2, *there exist infinitely many G-invariant non-Riemannian Finsler metrics on G/H.*

The existence part of this theorem was established in [Szabó 1981], but the one-to-one part was not given there.

By [Deng and Hou 2007], any G-invariant Finsler metric on G/H is a globally symmetric Berwald metric; and by [Deng and Hou 2005a], any globally symmetric Berwald metric can be constructed in this way. Therefore, Theorem 4.2 can be viewed as a classification of all globally symmetric Berwald metrics.

Example 4.3. Consider the unit sphere S^n in Euclidean $\mathbb{R}^{n+1}$. It is a Riemannian symmetric space of rank 1. The special orthogonal group $\mathrm{SO}(n+1)$ acts transitively on S^n, and the isotropic subgroup at $(1, 0, \dots, 0)$ can be identified with the subgroup $\mathrm{SO}(n)$ of $\mathrm{SO}(n+1)$. Thus S^n can be viewed as the coset space $\mathrm{SO}(n+1)/\mathrm{SO}(n)$, and the induced Riemannian metric on S^n can be viewed as an $\mathrm{SO}(n+1)$-invariant metric on $\mathrm{SO}(n+1)/\mathrm{SO}(n)$. By Theorem 4.2, there is no $\mathrm{SO}(n+1)$-invariant non-Riemannian Finsler metric on S^n. Now we consider the product manifold $S^n \times S^m$, where $m, n \geq 1$. Write $S^n \times S^m$ as

$$G/H = (\mathrm{SO}(n+1) \times \mathrm{SO}(m+1))/(\mathrm{SO}(n) \times \mathrm{SO}(m)).$$

The rank of G/H is 2. Therefore there exist infinitely many G-invariant non-Riemannian Finsler metrics on G/H. All these metrics are globally symmetric and of Berwald type. Now we can give an explicit description of these metrics. Let $\mathfrak{a}_1$ and $\mathfrak{a}_2$ be Cartan subspaces of $\mathrm{SO}(n+1)/\mathrm{SO}(n)$ and $\mathrm{SO}(m+1)/\mathrm{SO}(m)$, respectively. The corresponding Weyl groups on $\mathfrak{a}_1$ and $\mathfrak{a}_2$ are $W_1 = W_2 = \{1, -1\}$. The direct sum $\mathfrak{a} = \mathfrak{a}_1 \oplus \mathfrak{a}_2$ is a Cartan subspace of G/H, and the corresponding Weyl group W consists of four elements: $W = \{1, \sigma_1, \sigma_2, \sigma_3\}$, where

$$\sigma_1(x+y) = -x+y, \quad \sigma_2(x+y) = x-y, \quad \sigma_3(x+y) = -x-y \quad \text{for } x \in \mathfrak{a}_1,\ y \in \mathfrak{a}_2.$$

Therefore a Minkowski norm F on $\mathfrak{a}$ is W-invariant if and only if F satisfies

$$F(\pm x \pm y) = F(x+y) \quad \text{for } x \in \mathfrak{a}_1,\ y \in \mathfrak{a}_2. \tag{4-2}$$

By Theorem 4.2, there is a one-to-one correspondence between the G-invariant Finsler metrics on $S^n \times S^m$ and the Minkowski norms on $\mathfrak{a}$ satisfying (4-2). It is easily seen that there actually exist infinitely many Minkowski norms on $\mathfrak{a}$ that satisfy (4-2). For example, identifying $\mathfrak{a}_1$ and $\mathfrak{a}_2$ with $\mathbb{R}^1$, we can define a set of Minkowski norms by

$$F_\mu(x+y) = \sqrt{x^2 + y^2 + \mu\sqrt{x^4 + y^4}} \quad \text{for } x, y \in \mathbb{R}^1,$$

where μ is an arbitrary positive real number. These norms satisfy (4-2) and are pairwise not mutually linearly isometric [Bao et al. 2000].

Similarly, we can consider other Riemannian symmetric spaces of rank 1 and their product. Next we give an irreducible example of rank at least 2.

Example 4.4. Consider the Riemannian symmetric pair $(\mathrm{SL}(n,\mathbb{R}), \mathrm{SO}(n))$, where $n \geq 3$. The rank is $n-1$. So there exist infinitely many $\mathrm{SL}(n,\mathbb{R})$-invariant non-Riemannian Finsler metrics on $\mathrm{SL}(n,\mathbb{R})/\mathrm{SO}(n)$. Now we explicitly describe these metrics. The canonical decomposition of the Lie algebra is $\mathfrak{sl}(n,\mathbb{R}) = \mathfrak{so}(n) + \mathfrak{p}$, where $\mathfrak{p}$ consists of all $n \times n$ traceless symmetric matrices. A Cartan space can be taken as the space of all diagonal matrices in $\mathfrak{p}$, denoted by $\mathfrak{a}$. The corresponding Weyl group is isomorphic to the full permutation group of n indices, which acts on $\mathfrak{a}$ by permuting the entries along the diagonal [Helgason 1978]. Therefore, if we write the elements in $\mathfrak{a}$ as

$$\mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n), \quad \text{where } \textstyle\sum_{i=1}^n \lambda_i = 0,$$

then a Minkowski norm F on $\mathfrak{a}$ is W-invariant if and only if $F(\lambda_1, \dots, \lambda_n)$ is a symmetric function of $\lambda_1, \lambda_2, \dots, \lambda_n$. An explicit series of such norms can be constructed as follows:

$$F_\mu(\lambda_1, \dots, \lambda_n) = \sqrt{\sum_{i=1}^n \lambda_i^2 + \mu \sqrt{\sum_{i=1}^n \lambda_i^4}},$$

where μ is an arbitrary positive real number. As in Example 4.3, the Minkowski norms above define infinitely many $\mathrm{SL}(n,\mathbb{R})$-invariant non-Riemannian Finsler metrics on $\mathrm{SL}(n,\mathbb{R})/\mathrm{SO}(n)$. These metrics are all of the Berwald type.

Next we consider nonsymmetric polar homogeneous manifolds. By Kollross and Podestà [2003] classified the polar homogeneous spaces of a connected, simply connected, simple Lie group. Combining their list and Theorem 4.1 gives this:

Theorem 4.5. *Let G/H be a connected, simply connected, isotropic polar homogeneous manifold, where G is a simply connected simple compact Lie group and H is a closed subgroup of G. Then the pair (G, H) must be either a Riemannian symmetric pair or one of the pairs in Table 1. Among the manifolds in Table 1, any G-invariant Finsler metric on G/H must be Riemannian in types VIII and IX. In any of the other types, however, there exist infinitely many G-invariant non-Riemannian Finsler metrics on G/H.*

Proof. Because Kollross and Podestà [2003] listed the isotropic polar homogeneous manifolds, we only need to find which type is of rank 1. Since H is compact, the action of H on the tangent space $T_o(G/H)$ leaves the inner product invariant. Therefore, we can view H as a subgroup of $\mathrm{SO}(m)$ (note that H is connected in all the types in Table 1), where $m = \dim T_o(G/H)$. Therefore, to find out which type is of rank 1, we only need to find in which case H acts transitively on S^{m-1}. The compact connected subgroups of $\mathrm{SO}(m)$ that act transitively on S^{m-1} were classified by Montgomery, Samelson, and Borel [Besse 1987]. Therefore the

Type	G	H	Rank
I	$\mathrm{SU}(n+1)$	$\mathrm{SU}(n)$	≥ 2
II	$\mathrm{Sp}(n+1)$	$\mathrm{Sp}(n)$	≥ 2
III	$\mathrm{Sp}(n+1)$	$\mathrm{U}(1) \times \mathrm{Sp}(n)$	≥ 2
IV	$\mathrm{SU}(p+q)$	$\mathrm{SU}(p) \times \mathrm{SU}(q),\ p < q$	≥ 2
V	$\mathrm{Spin}(2n)$	$\mathrm{SU}(n),\ n$ odd	≥ 2
VI	E_6	$\mathrm{Spin}(10)$	≥ 2
VII	$\mathrm{Spin}(9)$	$\mathrm{Spin}(7)$	≥ 2
VIII	$\mathrm{Spin}(7)$	G_2	1
IX	G_2	$\mathrm{SU}(3)$	1

Table 1. Pairs (G, H) occurring in Theorem 4.5.

theorem can be proved through a case-by-case computation of the dimensions of the manifolds G/H. □

Now we give a description of the invariant Finsler metrics on the homogeneous manifolds of type I in Table 1.

Example 4.6. Consider a homogeneous manifold

$$M = G/H = \mathrm{SU}(n+1)/\mathrm{SU}(n), \quad \text{with } n \geq 2.$$

Now we give a realization of the manifold M. The group $G = \mathrm{SU}(n+1)$ acts in the standard way on the standard Hermitian space $\mathbb{C}^{n+1} = \mathbb{R}^{2n+2}$. The action keeps the sphere S^{2n+1} invariant, and the restriction of the action to the sphere is transitive. The subgroup $H = \mathrm{SU}(n)$ can be identified with the isotropic subgroup of G at the point $o = (1, 0, \dots, 0)' \in S^{2n+1}$, that is,

$$A \hookrightarrow \begin{pmatrix} 1 & 0 \\ 0 & A \end{pmatrix} \quad \text{for } A \in H.$$

Therefore M is just the sphere S^{2n+1}. Now we make some observations on the isotropic representation. By selecting a certain local coordinate system, we can identify the tangent space of G/H at o with the hyperplane

$$P = \{(1, b_1, b_2, \dots, b_{2n+1})' \in \mathbb{R}^{2n+1}\}$$

through the mapping

$$b = (b_1, b_2, \dots, b_{2n+1}) \hookrightarrow (1, b_1, b_2, \dots, b_{2n+1})'.$$

Then the isotropic representation can be described as follows. For

$$b = (b_1, b_2, \dots, b_{2n+1}) \in T_o(G/H) \quad \text{and} \quad A \in H,$$

define $\tilde{b} = (1 + b_1\sqrt{-1}, b_2 + b_3\sqrt{-1}, \ldots, b_{2n} + b_{2n+1}\sqrt{-1})' \in \mathbb{C}^{n+1}$. Let

$$c = \begin{pmatrix} 1 & 0 \\ 0 & A \end{pmatrix} \cdot \tilde{b} = \bigl(1 + b_1\sqrt{-1}, c_1 + c_2\sqrt{-1}, \ldots, c_{2n-1} + c_{2n}\sqrt{-1}\bigr)' \in \mathbb{C}^{n+1},$$

where c_j, $j = 1, 2, \ldots, 2n$ are real numbers. Then

$$A \cdot b = (b_1, c_1, c_2, \ldots, c_{2n}).$$

Therefore, the isotropic representation is equal to the identity transformation on the subspace $V_1 = \{(b_1, 0, \ldots, 0) \in T_o(M)\}$. On the other hand, the action of H on the subspace $V_2 = \{(0, b_2, \ldots, b_{2n+1}) \in T_o(M))\}$ is just the standard action of the group $\mathrm{SU}(n)$ on $\mathbb{C}^n$. Therefore, H is transitive on the unit sphere in V_2. From this, we see that the action of H on V_2 is polar of rank 1. Since $T_o(M)$ is the orthogonal sum of V_1 and V_2, the action of H on $T_o(M)$ is polar, and a Cartan space can be chosen to be $\mathfrak{a} = V_1 + \mathfrak{a}_2$, where $\mathfrak{a}_2$ is an arbitrary one-dimensional subspace in V_2. As in Example 4.3 any Minkowski norm F on $\mathfrak{a}$ satisfying

$$F(\pm x \pm y) = F(x + y), \quad x \in V_1, y \in \mathfrak{a}_2$$

can be extended uniquely to a H-invariant Minkowski norm on $T_o(M)$, and hence corresponds to a G-invariant Finsler metric on M.

5. General geometric properties

Let (G, K) be a Riemannian symmetric pair. Then by the results of [Szabó 1981] it is easily seen that any G-invariant Finsler metric on G/K must be a (reversible or nonreversible) affine symmetric Berwald space. By Dadok's results, we have seen that a polar representation must have the same orbits as the isotropic representation of a certain Riemannian symmetric space. It is therefore natural to ask whether the result above holds for polar homogeneous spaces, that is, whether any invariant Finsler metric on a polar homogeneous space must be Berwaldian. We will investigate this problem in this section. It is somehow surprising that the answer is negative. In fact we can prove that in any polar homogeneous manifold in Table 1, an invariant Finsler metric F on G/H is Berwaldian if and only if it is Riemannian. As an application, we show that on any polar homogeneous space of rank at least 2 in Table 1, there exist infinitely many invariant Finsler metrics that are reversible, non-Berwaldian and of vanishing S-curvature. The problem of the existence of such spaces was posed by Shen [2009], and in our paper [Deng and Hou $\geq$ 2010], we constructed some low-dimensional examples of Finsler spaces with the above properties.

We begin with the notions of weakly symmetric Finsler spaces and geodesic orbit Finsler spaces. Let (M, F) be a connected Finsler space. Then (M, F) is called weakly symmetric if for any two points p and q there exists an isometry σ

of (M, F) that interchanges them, that is, $\sigma(p) = q$ and $\sigma(q) = p$. It is called a geodesic orbit Finsler space if any geodesic γ is the orbit of a one-parameter subgroup of the full group of isometries, that is, if there exists a vector X in the Lie algebra $\mathfrak{g}$ of the full group G of isometries such that $\gamma(t) = \exp(tX) \cdot o$, where $o = \gamma(0)$ and exp is the exponential mapping of G. It is obvious that a weakly symmetric space must be reversible and homogeneous and that a geodesic orbit Finsler space must be homogeneous. Berndt et al. [1997] proved that a connected weakly symmetric Riemannian manifold must be a geodesic orbit space. Their proof is also valid for the Finslerian case, so a weakly symmetric Finsler must be a geodesic orbit space. Also it is easy to prove that a geodesic orbit space must have vanishing S-curvature [Deng and Hou $\geq$ 2010].

Nguyen [2000] introduced a way to construct weakly symmetric Riemannian manifolds. Let G be a connected Lie group and θ be an involutive automorphism of G. Suppose H is a θ-stable compact subgroup of G. Select a complement subspace $\mathfrak{m}$ of $\mathfrak{h}$ in $\mathfrak{g}$ that is also invariant under $\mathrm{Ad}_{\mathfrak{g}/\mathfrak{h}}(H)$. Then (G, H, θ) is called a weakly symmetric triple if, given any element $X \in \mathfrak{m}$, there exists an element $h \in H$ such that $(\mathrm{Ad}(h)) \circ d\theta(X) = -X$. Nguyen proved that if (G, H, θ) is weakly symmetric pair, then any G-invariant Riemannian metric on G/H is weakly symmetric. Using this method we can also construct weakly symmetric Finsler spaces.

Proposition 5.1. *If (G, H, θ) is a weakly symmetric triple, then any G-invariant reversible Finsler metric on G/H is weakly symmetric.*

The proof is similar to the Riemannian case, so we omit it [Nguyen 2000].

Theorem 5.2. *Let G/H be one of the polar homogeneous spaces in Table 1 (with nontrivial subgroup H) that is not of type II (that is, not* $\mathrm{Sp}(n+1)/\mathrm{Sp}(n)$*). Then any reversible G-invariant Finsler metric on G/H must be weakly symmetric. In the coset space* $\mathrm{Sp}(n)/\mathrm{Sp}(n-1)$ *for $n \geq 2$, there exist infinitely many invariant weakly symmetric non-Riemannian Finsler metrics. In particular, in any of the polar homogeneous manifolds of rank at least* 2 *in Table 1, there exist infinitely many invariant weakly symmetric non-Riemannian Finsler metrics.*

Proof. The first claim follows from the classification of compact weakly symmetric Riemannian spaces by Nguyen [2000] and Yakimova [2004]. Also some of these homogeneous manifolds are known to be weakly symmetric [Ziller 1996]. We now give a case-by-case clarification. The manifolds $\mathrm{SU}(n)/\mathrm{SU}(n-1)$, with $n \geq 3$, are known to be weakly symmetric [Ziller 1996]. The involutive automorphism θ of $\mathrm{SU}(n)$ can be defined in the following way: Let $\mathrm{SU}(n)$ act in the standard way on the unit sphere in $\mathbb{C}^n$, and let μ be the transformation taking the complex conjugation on each coordinate. Define $\theta(g) = \mu g \mu^{-1}$. It is easy to check that $(\mathrm{SU}(n), \mathrm{SU}(n-1), \theta)$ is a weakly symmetric triple. Therefore by Proposition 5.1,

$\mathrm{SU}(n)/\mathrm{SU}(n-1)$ endowed with any invariant reversible non-Riemannian Finsler metric must be a weakly symmetric Finsler space. This argument is also valid for the homogeneous space $\mathrm{SU}(p+q)/\mathrm{SU}(p) \times \mathrm{SU}(q)$, with $p < q$. Therefore $\mathrm{SU}(p+q)/\mathrm{SU}(p) \times \mathrm{SU}(q)$ endowed with any reversible Finsler spaces must be weakly symmetric. According to Nguyen [2000], the homogeneous space $G/H = \mathrm{Sp}(n)/\mathrm{Sp}(n-1) \cdot U(1)$ is weakly symmetric with respect to G, meaning that for any $X \in T_o(G/H)$, there exists an $h \in H$ such that $dh(X) = -X$. Similarly to the Riemannian case, this means that any invariant reversible Finsler metric on G/H must be weakly symmetric [Ziller 1996]. Now by [Nguyen 2000], the space $\mathrm{SO}(2n)/\mathrm{SU}(n)$, with n odd, is weakly symmetric. Since the space $\mathrm{Spin}(2n)/\mathrm{SU}(n)$ is the universal covering of $\mathrm{SO}(2n)/\mathrm{SU}(n)$, we see that $\mathrm{Spin}(2n)/\mathrm{SU}(n)$ endowed with any reversible must weakly symmetric [Yakimova 2004]. The situation is the same for the space $\mathrm{Spin}(9)/\mathrm{Spin}(7)$. Finally, the space $E_6/\mathrm{Spin}(10)$ appears in the list in [Nguyen 2000] (note that D_5 is exactly $\mathrm{Spin}(10)$). Therefore, if G/H is a homogeneous manifold of rank at least 2 in Table 1 not of type II, then any invariant reversible Finsler metric on G/H must be weakly symmetric.

Now we consider spaces of type II, that is, $\mathrm{Sp}(n)/\mathrm{Sp}(n-1)$ with $n \geq 2$. Note that $\mathrm{Sp}(n)$ acts transitively on the sphere S^{4n-1} in the standard way, and that any isotropy subgroup is isomorphic to $\mathrm{Sp}(n-1)$. Hence $\mathrm{Sp}(n)/\mathrm{Sp}(n-1) = S^{4n-1}$. In this way, we can also view $\mathrm{Sp}(n)$ as a subgroup of $\mathrm{SU}(2n)$. Since we have $\mathrm{SU}(2n)/\mathrm{SU}(2n-1) = S^{4n-1}$, we see that any $\mathrm{SU}(2n)$ invariant Finsler metric on the sphere S^{4n-1} must be an invariant metric on the coset space $\mathrm{Sp}(n)/\mathrm{Sp}(n-1)$. From this the theorem follows. □

Theorem 5.3. *Let G be a connected simply connected compact simple Lie group and H be a closed subgroup of G. Let G/H be a nonsymmetric polar homogeneous manifold. Then a G-invariant Finsler metric on G/H is Berwaldian if and only if it is Riemannian.*

Lemma 5.4. *Let G be a compact connected Lie group and H be a closed subgroup of G. Suppose G/H is diffeomorphic to the n-sphere, with $n \geq 2$. Then any G-invariant Riemannian metric on G/H is holonomy irreducible.*

Proof. Suppose conversely that a G-invariant Riemannian metric Q on G/H is holonomy reducible. Then by the de Rham decomposition theorem, we have a Riemannian manifold decomposition

$$G/H = M_1 \times M_2 \times \cdots \times M_s,$$

where $M_1, M_2, \ldots, M_s$ are holonomy irreducible. By Hano's result [1955], the full group K of isometries of $(G/H, Q)$ is isomorphic to the product of the groups of isometries of M_i for $1 \leq i \leq s$. This fact combined with the homogeneity of $(G/H, Q)$ implies that for each i, the group of isometries of M_i must be transitive on M_i; in particular, it has dimension at least 1. Thus the identity component K^0 of K is not a simple Lie group. Now by the complete list of compact connected Lie groups that act transitively on the spheres [Besse 1987], K^0 must be one of $\mathrm{U}(n)$ with $n \geq 2$, $\mathrm{Sp}(n)\,\mathrm{Sp}(1)$ with $n \geq 1$, or $\mathrm{Sp}(n)\,\mathrm{U}(1)$ with $n \geq 1$. This means that $s = 2$ and M_1 can be chosen to be a coset space of $\mathrm{U}(1) = S^1$ or $\mathrm{Sp}(1) = \mathrm{SU}(2) = S^3$. But it is easily seen that the coset spaces of $\mathrm{U}(1)$ or $\mathrm{Sp}(1)$ must be diffeomorphic to S^1, S^2 or S^3. Therefore we have $S^n = G/H = S^j \times M_2$ for $n > j$. But this decomposition is impossible since the j-th homotopy group of S^n (where $j = 1, 2$, or 3 and $n > j$) is the identity group and $\pi_j(S^j) = \mathbb{Z}$. □

Oniscik [1963] proved that the only compact connected groups that act transitively on the projective complex spaces are $\mathrm{SU}(n)$ (on $\mathbb{C}P^{n-1}$), and $\mathrm{Sp}(n)$ (on $\mathbb{C}P^{2n-1}$). Similar to Lemma 5.4, we have this:

Lemma 5.5. *Let G be a compact connected Lie group and H be a closed subgroup of G. Suppose the coset space G/H is diffeomorphic to the projective complex space $\mathbb{C}P^n$. Then any G-invariant Riemannian metric on G/H must be holonomy irreducible.*

To state the next lemma, we need some definitions about Hermitian symmetric spaces. Let (G, H) be an irreducible Riemannian metric of compact or noncompact type. Then it is well known that G/H can be made into a Hermitian symmetric space if and only if the (connected) compact subgroup H has nondiscrete center, in which case the center of H is a one-dimensional Lie group [Helgason 1978]. Let $\mathfrak{g} = \mathfrak{h} + \mathfrak{m}$ be the canonical decomposition of the Lie algebra. Let $\mathfrak{a}$ be a maximal abelian subspace of $\mathfrak{m}$ and extend $\mathfrak{a}$ to a Cartan subalgebra $\mathfrak{t}$ of $\mathfrak{g}$. Then on the subspace $\mathfrak{m}$ there is a complex structure J corresponding to the root system of $(\mathfrak{g}, \mathfrak{t})$ (not necessary the complex structure induced by that of the manifold G/H) [Korányi and Wolf 1965]. Let Z^J be the element in the center $\mathfrak{z}_{\mathfrak{h}}$ of $\mathfrak{h}$ that corresponds to the complex structure above. Then we have a decomposition $Z^J = Z^0 + Z'$, where Z^0 defines the complex structure on the polydisc or the polysphere inside G/H when G/H is realized as a generalized half-plane [Korányi and Wolf 1965] and Z' is an element in $\mathfrak{h}$ that centralizes $\mathfrak{a}$. The Hermitian symmetric space G/H is said to be of tube type if in the decomposition above we have $Z' = 0$. Let $H = Z_H \cdot H_s$ be the decomposition of H, where Z_H is the one-dimensional center of H and H_s is the semisimple part of H. Let $S' = \{\exp(tZ') \mid t \in \mathbb{R}\}$. It is known that if G/H is not of the tube type, then Z' is not in the center $\mathfrak{z}_{\mathfrak{h}}$ and we have $H = S'H_s = H_sS'$. Further,

$$\mathfrak{m} = \mathrm{Ad}(H_s)(\mathfrak{a}). \tag{5-1}$$

Lemma 5.6. *Let G be a compact connected simply connected simple Lie group and H be a closed subgroup of G such that (G, H) is an irreducible Hermitian* *symmetric pair of nontube type. Let $H = Z_H H_s$ be the decomposition of H, where Z_H is the one-dimensional center of H and H_s is the semisimple part of H. Then any G-invariant Riemannian metric on the coset space G/H_s must be holonomy irreducible.*

Proof. Let $\mathfrak{g} = \mathfrak{h} + \mathfrak{m}$ be the canonical decomposition of the symmetric pair (G, H). Then we can identify the tangent space $T_o(G/H)$ with $\mathfrak{m}$. Since (G, H) is irreducible, the action of H on $\mathfrak{m}$ is irreducible. Now we claim that the action of the semisimple part H_s of H on $\mathfrak{m}$ is also irreducible. In fact, if this is not true, then we can find a nontrivial subspace V_1 of $\mathfrak{m}$ that is invariant under H_s. Let $X \in V_1$. Then by (5-1) there exists $k \in H_s$ and $X_{\mathfrak{a}} \in \mathfrak{a}$ such that $X = \mathrm{Ad}(k)(X_{\mathfrak{a}})$. Now for any $s \in S'$, select $s_1 \in S'$, $k_1 \in H_s$ such that $sk = k_1 s_1$. Then

$$\begin{aligned}\mathrm{Ad}(s)(X) &= \mathrm{Ad}(s)\,\mathrm{Ad}(k)(X_{\mathfrak{a}}) = \mathrm{Ad}(k_1 s_1)(X_{\mathfrak{a}}) = \mathrm{Ad}(k_1)(X_{\mathfrak{a}}) \\ &= \mathrm{Ad}(k_1 k^{-1})\,\mathrm{Ad}(k)(X_{\mathfrak{a}}) = \mathrm{Ad}(k_1 k^{-1})(X) \in V_1,\end{aligned} \tag{5-2}$$

where we have used the fact that Z' centralizes $\mathfrak{a}$. Now (5-2) means that V_1 is also invariant under the action of S'. Hence it is invariant under H. This contradicts the assumption that G/H is irreducible, and proves our claim. The claim means that the action of H_s on the tangent space $T_o(G/H_s)$, which can be identified with $\mathfrak{s} + \mathfrak{m}$ (direct sum), where $\mathfrak{s}$ is the one-dimensional Lie algebra of S^1, decomposes as the sum of irreducible subspaces $\mathfrak{s}$ and $\mathfrak{m}$. By Schur's lemma, this implies that any H_s-invariant inner product on $T_o(G/H')$ must be of the form

$$Q_1 + \alpha(-B)|_{\mathfrak{m}\times\mathfrak{m}}, \tag{5-3}$$

where Q_1 is an arbitrary inner product on $\mathfrak{a}$ (which is unique up to a positive scalar), B is the Cartan–Killing form of the Lie algebra $\mathfrak{g}$, and α is an arbitrary positive number. Now by the construction of D'Atri and Ziller [1979], any inner product of the form (5-3) induces a naturally reductive G-invariant Riemannian metric on G/H'. Such a Riemannian metric on a simply connected coset space of a connected simple Lie group must be holonomy irreducible [Kobayashi and Nomizu 1969, page 215]. □

Proof of Theorem 5.3. We need only prove the "only if" part. We divide the homogeneous manifolds in Table 1 into three groups. Group 1 consists of types I, II, VII, VIII, and IX. Homogeneous manifolds in this group are diffeomorphic to a sphere. Group 2 consists of type III, the manifold $\mathrm{Sp}(n)/(\mathrm{U}(1)\times\mathrm{Sp}(n-1))$. It has a symmetric extension $((\mathrm{SU}(2n), \mathrm{SU}(2n-1)\times\mathrm{U}(1))$; in other words, the quotient $\mathrm{Sp}(n)/(\mathrm{U}(1)\times\mathrm{Sp}(n-1))$ is diffeomorphic to the projective complex space

$$\mathbb{C}P^{2n-1} = \mathrm{SU}(2n)/(\mathrm{SU}(2n-1)\times\mathrm{U}(1)),$$

where we consider $\mathrm{Sp}(n)$ as a subgroup of $\mathrm{SU}(2n)$. Group 3 consists of types IV, V, and VI. Homogeneous manifolds in this type are S^1-bundles over Hermitian symmetric spaces of nontube type [Nguyen 2000]. Let G/H be one of the homogeneous manifolds in Table 1. Then by Lemmas 5.4, 5.5, and 5.6, we have seen that any G-invariant Riemannian metric on G/H must be holonomy irreducible. Suppose F is an invariant Finsler metric on G/H of the Berwald type. Then there exists a Riemannian metric Q on G/H whose Levi-Civita connection coincides with the Chern connection of F [Szabó 1981]. Let $A(Q)$ and $I(Q)$ be the group of affine transformations and the group of isometries of Q. Then any isometry of F must be contained in $A(Q)$ [Deng and Hou 2005b]. In particular, $G \subset A(Q)$. On the other hand, since G/H is compact, we have $A(Q)^0 = I(Q)^0$ [Kobayashi and Nomizu 1963, page 244]. Since G is connected, we have $G \subset A(Q)^0 = I(Q)^0$. That is, any element of G is an isometry of Q, or in other words, Q is a G-invariant Riemannian metric and hence must be holonomy irreducible. If F is not Riemannian, then according to [Szabó 1981], $(G/H, Q)$ must be an irreducible Riemannian symmetric space of rank at least 2. In particular, let K be the full group of isometries of Q, let K_0 be the identity component of K, and let N be the isotropic subgroup of K_0 at a fixed point. Then $G \subset K^0$, and (K_0, N) is a Riemannian symmetric pair [Helgason 1978]. This means that the pair (G, H) has a symmetric extension with rank at least 2. The symmetric extension of weakly symmetric homogeneous manifolds has been classified by Yakimova [2004]. The list shows that the manifolds of types I, III, VII, VIII, and IX in Table 1 have symmetric extension of rank 1, and that the manifolds in other types do not admit any symmetric extension. On the other hand, $\mathrm{Sp}(n)/\mathrm{Sp}(n-1)$ is the only homogeneous manifold in Table 1 that is not weakly symmetric. But $\mathrm{Sp}(n)/\mathrm{Sp}(n-1)$ is diffeomorphic to the sphere S^{4n-1}. Hence $(\mathrm{Sp}(n), \mathrm{Sp}(n-1))$ has as its only symmetric extension $(\mathrm{SO}(4n), \mathrm{SO}(4n-1))$, which is of rank 1. This contradiction proves the theorem. □

Since a weakly symmetric Finsler space has vanishing S-curvature, combining Theorems 5.2 and 5.3 gives this corollary:

Corollary 5.7. *Let G be a connected simply connected compact Lie group, and let H be a closed subgroup of G. Let G/H be a nonsymmetric polar homogeneous manifold of rank 2. Then there exist infinitely many invariant Finsler metrics on G/H that are reversible, non-Berwaldian, and of vanishing S-curvature.*

6. Randers metrics

We now consider invariant Randers metrics on the homogeneous manifolds in Table 1, and give a (global) complete classification of such metrics. As pointed out in Theorem 5.3, all the non-Riemannian Randers metrics we find are non-Berwaldian. As an application, we give a classification of homogeneous Randers spaces with positive constant flag curvature.

We first recall some known results. Let G be a Lie group, and let H be a closed subgroup of G. Suppose Q is an invariant Riemannian metric on G/H. Then by the results of [Deng and Hou 2004b], there exists a bijection between the invariant Randers metrics on G/H with underlying Riemannian metric Q and the invariant vector fields on G/H with length < 1. Suppose the coset space G/H is reductive, that is, that there exists a subspace $\mathfrak{m}$ of $\mathfrak{g}$ (the Lie algebra of G) such that

$$\mathfrak{g} = \mathfrak{h} + \mathfrak{m} \quad \text{(direct sum of subspaces)},$$

and $\mathrm{Ad}(h)\mathfrak{m} \subset \mathfrak{m}$, for all $h \in H$. Then invariant vector fields are in bijection with the set

$$V = \{X \in \mathfrak{m} \mid \mathrm{Ad}(h)(X) = X \text{ for all } h \in H\}.$$

Therefore, to find all the invariant Randers metrics on G/H, we need first to find all the invariant Riemannian metrics with respect to which all vectors in V have length less than 1.

Now we consider $\mathrm{SU}(n)/\mathrm{SU}(n-1)$. In this case we have a decomposition

$$\mathfrak{su}(n) = \mathfrak{su}(n-1) + \mathfrak{m}, \tag{6-1}$$

where

$$\mathfrak{m} = \left\{ \begin{pmatrix} a\sqrt{-1} & \alpha \\ -\bar{\alpha}' & -\frac{1}{n-1}a\sqrt{-1}I_{n-1} \end{pmatrix} \,\middle|\, a \in \mathbb{R},\ \alpha \in \mathbb{C}^{n-1} \right\}. \tag{6-2}$$

A direct computation shows that

$$V = \left\{ \mathrm{diag}\left(-a\sqrt{-1}, \tfrac{1}{n-1}a\sqrt{-1}, \dots, \tfrac{1}{n-1}a\sqrt{-1}\right) \,\middle|\, a \in \mathbb{R} \right\}.$$

Theorem 6.1. *There is a one-to-one correspondence between the invariant Randers metrics on* $\mathrm{SU}(n)/\mathrm{SU}(n-1)$ *and the Minkowski norms on* $\mathfrak{m}$ *in* (6-2)*. This bijection is defined by*

$$F_o(X) = \sqrt{c_1 a^2 + c_2 \alpha\bar{\alpha}'} + c_1 c a, \qquad X = \begin{pmatrix} a\sqrt{-1} & \alpha \\ -\bar{\alpha}' & -\frac{1}{n-1}a\sqrt{-1}I_{n-1} \end{pmatrix} \in \mathfrak{m},$$

where c_1, c_2 *are positive real parameters and* c *is a real number with* $|c| < 1/\sqrt{c_1}$.

Among these Randers metrics, given any positive number k, *the family with parameters*

$$c_1 = \frac{d^2+k^2-kd^2}{(k-d^2)^2}, \quad c_2 = \left(\frac{k}{k-d^2}\right)^2, \quad c = \frac{d(d^2-k)}{d^2+k^2-kd^2},$$

has constant flag curvature k, where d is a real parameter satisfying $|d| < \sqrt{k}$. Moreover, up to isometry these are all the connected simply connected homogeneous Randers spaces with positive constant flag curvature.

Moreover, any non-Riemannian Randers metric above is not of the Douglas type, and hence is not projectively flat.

Proof. As stated above, to obtain the classification of invariant Randers metrics on $\mathrm{SU}(n)/\mathrm{SU}(n-1)$, we first need to determine all invariant Riemannian metrics thereon. These metrics are in bijection with the $\mathrm{SU}(n-1)$-invariant inner products on $\mathfrak{m}$. Now the vector space $\mathfrak{m}$, as a representation space of $\mathrm{SU}(n-1)$, has the decomposition $\mathfrak{m} = V + \mathfrak{m}_1$, where

$$\mathfrak{m}_1 = \left\{ \begin{pmatrix} 0 & \alpha \\ -\bar{\alpha}' & 0 \end{pmatrix} \,\middle|\, \alpha \in \mathbb{C}^{n-1} \right\}.$$

Moreover, the subrepresentations V and $\mathfrak{m}_1$ are irreducible. By Schur's lemma, it is easily seen that any $\mathrm{SU}(n-1)$-invariant inner product on $\mathfrak{m}$ must be of the form

$$\langle X_1, X_2 \rangle = c_1 a_1 a_2 + c_2 \operatorname{Re}(\alpha_1 \bar{\alpha}_2') \quad \text{for } c_1, c_2 > 0,$$

where

$$X_i = \begin{pmatrix} a_i\sqrt{-1} & \alpha_i \\ -\bar{\alpha}_i' & -\frac{1}{n-1} a_i \sqrt{-1} \end{pmatrix} \quad \text{for } i = 1, 2.$$

Therefore, the $\mathrm{SU}(n)$-invariant Randers metric on $\mathrm{SU}(n)/\mathrm{SU}(n-1)$ determined by $\langle\,\cdot\,,\cdot\,\rangle$ and

$$X_0 = \begin{pmatrix} c\sqrt{-1} & 0 \\ 0 & -\frac{1}{n-1} c \sqrt{-1} \end{pmatrix} \in V, \quad \text{with } \sqrt{c_1} c < 1,$$

must be the one stated in the theorem.

Now we prove the theorem's second claim. Note that $\mathrm{SU}(n)$ is a closed subgroup of $\mathrm{SO}(2n)$ and is transitive on the sphere $S^{2n-1} = \mathrm{SO}(2n)/\mathrm{SO}(2n-1)$. This means that any $\mathrm{SO}(2n)$-invariant Riemannian metric must be $\mathrm{SU}(n)$-invariant. Thus for any $k > 0$ there is one $\mathrm{SU}(n)$-invariant Riemannian metric on $\mathrm{SU}(n)/\mathrm{SU}(n-1)$ with constant sectional curvature k. We denote this Riemannian metric by Q_k, and next determine it explicitly. For two orthogonal unit vectors $X, Y \in \mathfrak{m}$, the sectional curvature of the plane spanned by X, Y is given by [Besse 1987, p. 183]

$$\begin{aligned} K(X,Y) = -\tfrac{3}{4}\bigl|[X,Y]_{\mathfrak{m}}\bigr|^2 - \tfrac{1}{2}\langle [X,[X,Y]]_{\mathfrak{m}}, Y\rangle \\ -\tfrac{1}{2}\langle [Y,[Y,X]]_{\mathfrak{m}}, X\rangle + |U(X,Y)|^2 - \langle U(X,X), U(Y,Y)\rangle, \end{aligned}$$

where $\langle\,\cdot\,,\cdot\,\rangle$ is the inner product determined by Q_k, $|\cdot|$ is the length function of $\langle\,\cdot\,,\cdot\,\rangle$, $[\,\cdot\,,\cdot\,]$ is the Lie bracket of $\mathfrak{su}(n)$, $[X,Y]_{\mathfrak{m}}$ denotes the projection of

$[X, Y]$ to $\mathfrak{m}$ corresponding to the decomposition (6-1), and U is a symmetric bilinear mapping from $\mathfrak{m} \times \mathfrak{m}$ to $\mathfrak{m}$ defined by

$$\langle U(X, Y), Z\rangle = \tfrac{1}{2}\big(\langle [Z, X]_{\mathfrak{m}}, Y\rangle + \langle [Z, Y]_{\mathfrak{m}}, X\rangle\big).$$

Up to homotheties, the set of invariant Riemannian metrics on $\mathrm{SU}(n)/\mathrm{SU}(n-1)$ has dimension 1, so there must be a $c_1 > 0$ such that the inner product $c_1 a_1 a_2 + \mathrm{Re}(\alpha_1 \bar{\alpha}_2')$ defines a Riemannian metric with constant sectional curvature. To find this c_1, we select three vectors in $\mathfrak{m}$:

$$X_1 = \operatorname{diag}\Big(\frac{\sqrt{-1}}{\sqrt{c_1}}, -\frac{\sqrt{-1}}{(n-1)\sqrt{c_1}} I_{n-1}\Big),$$

$$X_2 = \begin{pmatrix} 0 & \alpha_2 \\ -\alpha_2' & 0 \end{pmatrix}, \quad \alpha_2 = (1, 0, \dots, 0) \in \mathbb{R}^{n-1}, \quad X_3 = \sqrt{-1} X_2.$$

Then a direct (albeit somewhat tedious) computation shows that

$$\begin{aligned} U(X_1, X_1) &= U(X_2, X_2) = U(X_3, X_3) = 0, \\ U(X_1, X_2) &= \frac{1}{2}\Big(\frac{n}{(n-1)\sqrt{c_1}} - 2\sqrt{c_1}\Big) X_3, \\ U(X_1, X_3) &= \frac{1}{2}\Big(-\frac{n}{(n-1)\sqrt{c_1}} + 2\sqrt{c_1}\Big) X_2, \\ U(X_2, X_3) &= 0. \end{aligned}$$

Substituting the above into the formula of sectional curvature, we get

$$K(X_1, X_2) = c_1 + \frac{n}{n-1}(1 - \sqrt{c_1}) \quad \text{and} \quad K(X_2, X_3) = 4 - 3c_1.$$

Now the equation $K(X_1, X_2) = K(X_2, X_3)$ has a unique positive solution $c_1 = 1$, and in this case the sectional curvature is equal to 1. Therefore the inner product

$$(1/k) a_1 a_2 + (1/k) \mathrm{Re}(\alpha_1 \bar{\alpha}_2')$$

defines the invariant Riemannian metric Q_k on $\mathrm{SU}(n)/\mathrm{SU}(n-1)$ with constant sectional curvature k.

Now for any $X \in V$ with $Q_k(X, X) < 1$, we can construct an invariant Randers metric $F_{k,X}$ on $\mathrm{SU}(n)/\mathrm{SU}(n-1)$. Since Q_k is $\mathrm{SU}(n)$-invariant, the one parameter group $\{\exp(tX) \mid t \in \mathbb{R}\}$ consists of isometries of Q_k. In particular, for any $U, V \in \mathfrak{m}$, we have $Q_k\big(\mathrm{Ad}(\exp(tX))(U), \mathrm{Ad}(\exp(tX))(V)\big) = Q_k(U, V)$ for all t. Taking the derivative and considering the value at $t = 0$, we get

$$Q_k([X, U]_{\mathfrak{m}}) + Q_k(U, [X, V]_{\mathfrak{m}}) = 0.$$

In other words, $\mathscr{L}_{\tilde{X}} Q_k$ is equal to 0 at the origin. Since both the Riemannian metric and the vector field $\tilde{X}$ are $\mathrm{SU}(n)$-invariant, we have $\mathscr{L}_{\tilde{X}} Q_k = 0$ everywhere. By the criterion for a Randers metric to have constant flag curvature [Bao et al. 2004], the Randers metric with navigation data $(Q_k, \tilde{X})$, where

$$X = \begin{pmatrix} d\sqrt{-1} & 0 \\ 0 & -\frac{1}{n-1}d\sqrt{-1} \end{pmatrix}, \quad \text{with } |d| < \sqrt{k},$$

has constant flag curvature k. By the transformation formulas between the defining data and navigation data of Randers metrics [Chern and Shen 2005], we see that the Randers metrics with parameters as described in the theorem have constant flag curvature k.

Now we prove the converse of the conclusion above. Suppose (M, F) is a connected simply connected homogeneous non-Riemannian Randers metric with constant positive flag curvature k. Suppose the underlying Riemannian metric is Q and the corresponding vector field is $\tilde{X}$. Then (M, Q) is a connected simply connected Riemannian metric with constant positive sectional curvature k. Thus M is diffeomorphic to a sphere and $\tilde{X}$ is invariant under the full group of isometries of (M, Q). In particular, $\tilde{X}$ has constant length with respect to Q. This means that the Randers metric is in the corrected Yasuda–Shimada family. That is, it satisfies $\theta = 0$, and up to isometry there is only one family of such Randers metrics on odd-dimensional spheres [Bao et al. 2004]. Therefore they must be exactly the metrics constructed above.

Finally, by our previous result [An and Deng 2008], the Randers metric $F_{k,X}$ is of Douglas type if and only if $Q_k(X, [U, V]_{\mathfrak{m}}) = 0 \quad \text{for all } U, V \in \mathfrak{m}$. A simple direct computation shows that this holds only if $X = 0$. Thus any non-Riemannian Randers metric constructed above is not of the Douglas type. □

Other cases can be treated similarly; we omit the details here. The conclusion is that on each of the homogeneous manifolds other than the types VII, VIII, and IX, there exist invariant non-Riemannian Randers metrics. Any such metric is not of the Douglas type. On $\mathrm{Sp}(n)/\mathrm{Sp}(n-1)$ with $n \geq 2$, and given any $k > 0$, there is also a family of invariant Randers metrics with constant flag curvature k. By the arguments above, this family must be isometric to the corresponding family on $\mathrm{SU}(2n)/\mathrm{SU}(2n-1) = S^{4n-1}$ that has constant flag curvature k. Moreover, on $\mathrm{Spin}(9)/\mathrm{Spin}(7)$, since the isotropy representation has no fixed nonzero points [Ziller 1996], there are no $\mathrm{Spin}(9)$-invariant non-Riemannian Randers metrics, although there do exist invariant non-Riemannian metrics. Of course, there are no invariant non-Riemannian metrics on the homogeneous manifolds of types VIII and IX.

References

[An and Deng 2008] H. An and S. Deng, "Invariant (α, β)-metrics on homogeneous manifolds", *Monatsh. Math.* **154**:2 (2008), 89–102. MR 2009b:53077 Zbl 05553530

[Bao et al. 2000] D. Bao, S.-S. Chern, and Z. Shen, *An introduction to Riemann–Finsler geometry*, Graduate Texts in Mathematics **200**, Springer, New York, 2000. MR 2001g:53130 Zbl 0954.53001

[Bao et al. 2004] D. Bao, C. Robles, and Z. Shen, "Zermelo navigation on Riemannian manifolds", *J. Differential Geom.* **66**:3 (2004), 377–435. MR 2005k:58023 Zbl 1078.53073

[Berndt et al. 1997] J. Berndt, O. Kowalski, and L. Vanhecke, "Geodesics in weakly symmetric spaces", *Ann. Global Anal. Geom.* **15**:2 (1997), 153–156. MR 98f:53037 Zbl 0880.53044

[Besse 1987] A. L. Besse, *Einstein manifolds*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) **10**, Springer, Berlin, 1987. MR 88f:53087 Zbl 0613.53001

[Chern and Shen 2005] S.-S. Chern and Z. Shen, *Riemann–Finsler geometry*, Nankai Tracts in Mathematics **6**, World Scientific, Hackensack, NJ, 2005. MR 2006d:53094 Zbl 1085.53066

[Dadok 1982] J. Dadok, "On the C^∞ Chevalley's theorem", *Adv. in Math.* **44**:2 (1982), 121–131. MR 83m:53073 Zbl 0521.22009

[Dadok 1985] J. Dadok, "Polar coordinates induced by actions of compact Lie groups", *Trans. Amer. Math. Soc.* **288**:1 (1985), 125–137. MR 86k:22019 Zbl 0565.22010

[D'Atri and Ziller 1979] J. E. D'Atri and W. Ziller, *Naturally reductive metrics and Einstein metrics on compact Lie groups*, Mem. Amer. Math. Soc. **18**:215, Amer. Math. Soc., Providence, RI, 1979. MR 80i:53023 Zbl 0404.53044

[Deng and Hou 2002] S. Deng and Z. Hou, "The group of isometries of a Finsler space", *Pacific J. Math.* **207**:1 (2002), 149–155. MR 2003m:53127 Zbl 1055.53055

[Deng and Hou 2004a] S. Deng and Z. Hou, "Invariant Finsler metrics on homogeneous manifolds", *J. Phys. A* **37**:34 (2004), 8245–8253. MR 2005e:53112 Zbl 1062.58007

[Deng and Hou 2004b] S. Deng and Z. Hou, "Invariant Randers metrics on homogeneous Riemannian manifolds", *J. Phys. A* **37**:15 (2004), 4353–4360. Corrected in **39**:18 (2006), 5249–5250. MR 2005f:53125 Zbl 1049.83005

[Deng and Hou 2005a] S. Deng and Z. Hou, "Minkowski symmetric Lie algebras and symmetric Berwald spaces", *Geom. Dedicata* **113** (2005), 95–105. MR 2006e:53131 Zbl 1084.53059

[Deng and Hou 2005b] S. Deng and Z. Hou, "On locally and globally symmetric Berwald spaces", *J. Phys. A* **38**:8 (2005), 1691–1697. MR 2005j:53028

[Deng and Hou 2007] S. Deng and Z. Hou, "On symmetric Finsler spaces", *Israel J. Math.* **162** (2007), 197–219. MR 2008k:53161 Zbl 1141.53071

[Deng and Hou $\geq$ 2010] S. Deng and Z. Hou, "Weakly symmetric Finsler spaces", preprint. To appear in Comm. Contemp. Math.

[Hano 1955] J.-i. Hano, "On affine transformations of a Riemannian manifold", *Nagoya Math. J.* **9** (1955), 99–109. MR 17,891d Zbl 0067.14601

[Helgason 1978] S. Helgason, *Differential geometry, Lie groups, and symmetric spaces*, Pure and Applied Mathematics **80**, Academic Press, New York, 1978. MR 80k:53081 Zbl 0451.53038

[Kobayashi and Nomizu 1963] S. Kobayashi and K. Nomizu, *Foundations of differential geometry*, vol. 1, Wiley, New York, 1963. MR 27 #2945 Zbl 0119.37502

[Kobayashi and Nomizu 1969] S. Kobayashi and K. Nomizu, *Foundations of differential geometry*, vol. 2, Interscience Tracts in Pure and Applied Mathematics **15**, Wiley, New York, 1969. MR 38 #6501 Zbl 0175.48504

[Kollross and Podestà 2003] A. Kollross and F. Podestà, "Homogeneous spaces with polar isotropy", *Manuscripta Math.* **110**:4 (2003), 487–503. MR 2004b:53074 Zbl 1048.53035

[Korányi and Wolf 1965] A. Korányi and J. A. Wolf, "Realization of hermitian symmetric spaces as generalized half-planes", *Ann. of Math.* (2) **81** (1965), 265–288. MR 30 #4980 Zbl 0137.27402

[Kostant 1973] B. Kostant, "On convexity, the Weyl group and the Iwasawa decomposition", *Ann. Sci. École Norm. Sup.* (4) **6** (1973), 413–455. MR 51 #806 Zbl 0293.22019

[Krämer 1975] M. Krämer, "Eine Klassifikation bestimmter Untergruppen kompakter zusammenhängender Liegruppen", *Comm. Algebra* **3**:8 (1975), 691–737. MR 51 #13140 Zbl 0309.22013

[Manturov 1961a] O. V. Manturov, "Homogeneous, non-symmetric Riemannian spaces with an irreducible rotation group", *Dokl. Akad. Nauk SSSR* **141** (1961), 792–795. In Russian; translated in *Sov. Math., Dokl.* **2** (1961), 1550–1554. MR 25 #2552 Zbl Zbl 0188.26502

[Manturov 1961b] O. V. Manturov, "Riemannian spaces with orthogonal and symplectic motion groups and an irreducible rotation group", *Dokl. Akad. Nauk SSSR* **141** (1961), 1034–1037. In Russian; translated in *Sov. Math., Dokl* **2** (1961), 1034–1037. MR 25 #2551

[Manturov 1966] O. V. Manturov, "Homogeneous Riemannian spaces with an irreducible rotation group", *Trudy Sem. Vektor. Tenzor. Anal.* **13** (1966), 68–145. In Russian. MR 35 #926 Zbl 0173.24102

[Nguyen 2000] H. D. Nguyễñ, "Compact weakly symmetric spaces and spherical pairs", *Proc. Amer. Math. Soc.* **128**:11 (2000), 3425–3433. MR 2001b:53060 Zbl 0976.53056

[Oniščik 1963] A. L. Oniščik, "Transitive compact transformation groups", *Mat. Sb.* (*N.S.*) **60 (102)** (1963), 447–485. In Russian; translated in *Amer. Math. Soc. Trans.*, **55**:2 (1966), 153–194. MR 27 #5868

[Shen 2009] Z. Shen, "Some open problems in Finsler geometry", preprint, 2009, available at www.math.iupui.edu/∼zshen/Research/papers/Problem.pdf.

[Szabó 1981] Z. I. Szabó, "Positive definite Berwald spaces: Structure theorems on Berwald spaces", *Tensor* (*N.S.*) **35**:1 (1981), 25–39. MR 82f:53043 Zbl 0464.53025

[Szabó 2006] Z. I. Szabó, "Berwald metrics constructed by Chevalley polynomials", preprint, 2006. arXiv math/0601522

[Tam 1998] T.-Y. Tam, "A unified extension of two results of Ky Fan on the sum of matrices", *Proc. Amer. Math. Soc.* **126**:9 (1998), 2607–2614. MR 98m:15032 Zbl 0912.15018

[Wang and Ziller 1991] M. Wang and W. Ziller, "On isotropy irreducible Riemannian manifolds", *Acta Math.* **166**:3-4 (1991), 223–261. MR 92b:53078 Zbl 0732.53040

[Wolf 1968] J. A. Wolf, "The goemetry and structure of isotropy irreducible homogeneous spaces", *Acta Math.* **120** (1968), 59–148. Correction in **152** (1984), 141–142. MR 36 #6549 Zbl 0157.52102

[Wolf 1977] J. A. Wolf, *Spaces of constant curvature*, 4th ed., Publish or Perish, Houston, 1977. MR 88k:53002 Zbl 0281.53034

[Yakimova 2004] O. S. Yakimova, "Weakly symmetric Riemannian manifolds with a reductive isometry group", *Mat. Sb.* **195**:4 (2004), 143–160. In Russian; translated in *Sb. Math.* **195**:3-4 (2004), 143–160. MR 2005f:53081 Zbl 1078.53043

[Ziller 1996] W. Ziller, "Weakly symmetric spaces", pp. 355–368 in *Topics in geometry*, edited by S. Gindikin, Progr. Nonlinear Differential Equations Appl. **20**, Birkhäuser, Boston, MA, 1996. MR 97c:53081 Zbl 0860.53030

Received April 30, 2009.

SHAOQIANG DENG
SCHOOL OF MATHEMATICAL SCIENCES AND LPMC
NANKAI UNIVERSITY
TIANJIN 300071
CHINA

dengsq@nankai.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

A PROOF OF THE CONCUS–FINN CONJECTURE

KIRK E. LANCASTER

To Paul Concus and Robert Finn

Consider a nonparametric capillary or prescribed mean curvature surface $z = f(x, y)$ defined in a cylinder $\Omega \times \mathbb{R}$ over a two-dimensional region Ω that has a boundary corner point at O with an opening angle of 2α. Suppose $2\alpha \leq \pi$ and the contact angle approaches limiting values γ_1 and γ_2 in $(0, \pi)$ as O is approached along each side of the opening angle. Our results yield a proof of the Concus–Finn conjecture, which provides the last piece of the puzzle of determining the qualitative behavior of a capillary surface at a convex corner. We find that

- **if (γ_1, γ_2) satisfies $2\alpha + |\gamma_1 - \gamma_2| > \pi$, then f is bounded but discontinuous at O and has radial limits at O from all directions in Ω and, these radial limits behave in a prescribed way;**
- **if (γ_1, γ_2) satisfies $|\gamma_1 + \gamma_2 - \pi| > 2\alpha$, then f is unbounded in every neighborhood of O; and**
- **otherwise f is continuous at O.**

1. Introduction and statement of theorems

Let $\Omega \subset \mathbb{R}^2$ be a connected, open set. Consider the prescribed mean curvature problem

$$Nf = H(\cdot, f(\cdot)) \quad \text{in } \Omega, \tag{1}$$

$$Tf \cdot \boldsymbol{\nu} = \cos\gamma \qquad \text{almost everywhere on } \partial\Omega, \tag{2}$$

where $Tf = \nabla f/\sqrt{1+|\nabla f|^2}$, $Nf = \nabla \cdot Tf$, $\boldsymbol{\nu}$ is the exterior unit normal on $\partial\Omega$, $H(x, t)$ is a weakly increasing function of t for each $x \in \Omega$ and $\gamma = \gamma(x) \in [0, \pi]$. If (1) specifically is

$$Nf = \kappa f + \lambda \quad \text{in } \Omega \tag{3}$$

(that is, $H(x, t) = \kappa t + \lambda$), where κ and λ are constants with $\kappa \geq 0$, then the surface $z = f(x)$ for $x \in \Omega$ represents the stationary liquid-gas interface formed

MSC2000: primary 76B45; secondary 35J60, 53A10.
Keywords: Concus–Finn conjecture, capillary graph.

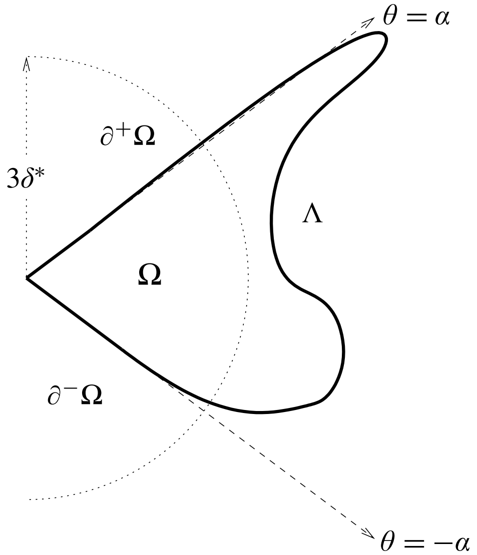

Figure 1. The domain Ω.

by an incompressible fluid in a vertical cylindrical tube with cross section Ω in a microgravity environment or in a downward oriented gravitational field; here the subgraph $U=\{(x,t)\in\Omega\times\mathbb{R}: t<f(x)\}$ represents the fluid-filled portion of the cylinder and $\gamma(x)$ is the angle at which the liquid-gas interface meets the vertical cylinder at $(x, f(x))$ [Finn 1986].

Since 1970, Paul Concus and Robert Finn have made fundamental contributions to the mathematical theory of capillary surfaces and have discovered that these surfaces can behave in very peculiar and unexpected ways; see for example [Finn 1999; 2002b; 2002a]. Of particular interest, to both the mathematical and physical theories in vertical cylinders, are domains Ω whose boundaries contain corners.

Suppose $O=(0,0)\in\partial\Omega$ and Ω has a corner of size $2\alpha\le\pi$ at O. With Ω as illustrated in Figure 1, suppose there exist $\gamma_1,\gamma_2\in(0,\pi)$ such that

$$\lim_{\partial^+\Omega\ni x\to(0,0)}\gamma(x)=\gamma_1 \quad\text{and}\quad \lim_{\partial^-\Omega\ni x\to(0,0)}\gamma(x)=\gamma_2. \tag{4}$$

Then Figure 2 can be used to illustrate our knowledge of the behavior of a solution f of (3) and (2) at the corner O; here let R, $D_1^\pm$ and $D_2^\pm$ be the indicated open regions in the (open) square $(0,\pi)\times(0,\pi)$. If $(\gamma_1,\gamma_2)\in\overline{R}\cap(0,\pi)\times(0,\pi)$, then f is continuous at O; see [Concus and Finn 1996, Theorem 1; Lancaster and Siegel 1996a, Corollary 4]. If $(\gamma_1,\gamma_2)\in D_1^\pm$, then f is unbounded in any neighborhood of O and the capillary problem has no solution if $\kappa=0$ [Concus and Finn 1996; Finn 1996]. If $(\gamma_1,\gamma_2)\in D_2^\pm$, then f is bounded [Lancaster and Siegel 1996a, Proposition 1] but its continuity at O is unknown. Concus and Finn

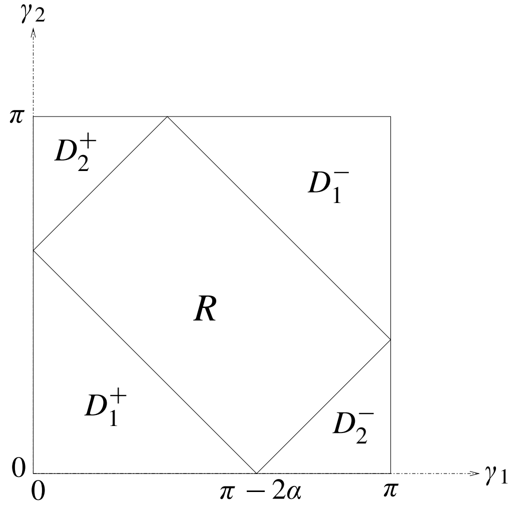

Figure 2. The Concus–Finn rectangle.

discovered bounded solutions of (3) and (2) in domains with corners whose unit normals (that is, Gauss maps) cannot extend continuously as functions of (x, y) to a corner on the boundary of the domain (for example [Finn 1988b; Finn 1988a, page 15; Concus and Finn 1996, Example 2; Finn 1996]). In 1992, as a result of computational experiments, they formulated a conjecture on the continuity of such surfaces [Concus et al. 1992; Concus and Finn 1996, page 67]; additional numerical experiments in 1994 by Concus and Finn and in 1996 by Mittelmann and Zhu found evidence to support the conjecture, which says that if $(\gamma_1, \gamma_2) \in D_2^{\pm}$, then f has a jump discontinuity at O [Finn 1999, page 776]. Writing the conditions for a pair of angles to be in $D_2^{\pm}$ yields the following formulation of the conjecture:

Concus–Finn conjecture. *Suppose that* $0 < \alpha < \pi/2$, *that the limits* (4) *exist and that* $0 < \gamma_1, \gamma_2 < \pi$. *If* $2\alpha + |\gamma_1 - \gamma_2| > \pi$, *then any solution of* (1) *and* (2), *with* $H(x, z) = \kappa z + \lambda$ *and* κ *nonnegative*, *has a jump discontinuity at* O.

We will prove this conjecture when $\partial\Omega \setminus \{(0, 0)\}$ is locally Hölder continuously differentiable and γ is locally Hölder continuous on $\partial\Omega \setminus \{(0, 0)\}$ in a neighborhood of the origin. For convenience, we will adopt the following notation throughout this paper. We will write points of $\mathbb{R}^2$ as lower case letters (for example, x) and points of $\mathbb{R}^3$ as upper case letters (for example, X). For $m \in \mathbb{N}$ with $m \geq 2$, we will write O_m as the origin in $\mathbb{R}^m$; however, we will write O for $O_2 = (0, 0)$. We denote by $B^m(P, r)$ the open ball in $\mathbb{R}^m$ of radius $r > 0$ centered at $P \in \mathbb{R}^m$ and by $B(x, r)$ the ball $B^2(x, r)$ for $x \in \mathbb{R}^2$. We will fix $\rho^* \in (0, 1)$ and $\alpha \in (0, \pi]$; later we will assume $\alpha \leq \pi/2$. We will write $\omega(\theta)$ for $(\cos(\theta), \sin(\theta))$ for $\theta \in \mathbb{R}$.

Our domain Ω will be a connected, simply connected open set in $\mathbb{R}^2$ such that $O \in \partial\Omega$, $\partial\Omega \setminus \{O\}$ is a piecewise C^1 curve, $\partial\Omega$ has a corner of size 2α at O, and the tangent cone to $\partial\Omega$ at O is $L^+ \cup L^-$, where polar coordinates relative to O are denoted by r and θ, and $L^+ = \{\theta = \alpha\}$ and $L^- = \{\theta = -\alpha\}$. We will assume there exists $\delta^* > 0$ such that $\partial^+\Omega = \partial\Omega \cap \overline{B(O, 3\delta^*)} \cap T^+$ and $\partial^-\Omega = \partial\Omega \cap \overline{B(O, 3\delta^*)} \cap T^-$ are connected, C^{1,ρ^*} arcs such that the tangent rays to $\partial^+\Omega$ and $\partial^-\Omega$ at O are L^+ and L^- respectively; here $T^+ = \{x \in \mathbb{R}^2 : x_2 \geq 0\}$ and $T^- = \{x \in \mathbb{R}^2 : x_2 \leq 0\}$. We set $\Lambda = \partial\Omega \setminus (\partial^+\Omega \cup \partial^-\Omega)$ and obtain

$$\partial\Omega = \partial^+\Omega \cup \partial^-\Omega \cup \Lambda \quad \text{with } O \in \partial^+\Omega \cap \partial^-\Omega \text{ and } B(O, 3\delta^*) \cap \overline{\Lambda} = \varnothing.$$

We will assume $\Omega \subset \{r\omega(\theta) : r > 0, -\pi < \theta < \pi\}$. Let us define $\tau^+ \in C^{0,\rho^*}(\partial^+\Omega)$ and $\tau^- \in C^{0,\rho^*}(\partial^-\Omega)$ such that $\tau^+(O) = \alpha$, $\tau^-(O) = -\alpha$,

$$(\cos(\tau^+(x)), \sin(\tau^+(x)), 0) \quad \text{is a unit tangent to } \partial^+\Omega \times \mathbb{R} \text{ for } x \in \partial^+\Omega$$

and

$$(\cos(\tau^-(x)), \sin(\tau^-(x)), 0) \quad \text{is a unit tangent to } \partial^-\Omega \times \mathbb{R} \text{ for } x \in \partial^-\Omega.$$

We will assume (4) holds and that $\gamma \in C^{0,\rho^*}(\overline{\partial^+\Omega})$ (when $\gamma(O)$ is set equal to γ_1) and $\gamma \in C^{0,\rho^*}(\overline{\partial^-\Omega})$ (when $\gamma(O)$ is set equal to γ_2.) If $\gamma_1 = \pi/2$ or $\gamma_2 = \pi/2$, we will need to be able to use slicing [Allard 1972, 4.10] and so we will assume

$$(5) \qquad |D\gamma| \in L^1(\partial^+\Omega) \quad \text{if } \gamma_1 = \tfrac{1}{2}\pi \quad \text{and} \quad |D\gamma| \in L^1(\partial^-\Omega) \quad \text{if } \gamma_2 = \tfrac{1}{2}\pi.$$

We will also assume (γ, Ω, O) is admissible as defined in Definition 3.4 (which essentially says Emmer's (boundary) condition holds at each point of $\partial\Omega \setminus \{O\}$). For a solution $f \in C^2(\Omega) \cap C^{1,\rho^*}(\overline{\Omega} \setminus \{O\})$ of (1) and (2), we let

$$\vec{n}(X) = \vec{n}_f(X) = \frac{(\nabla f(x), -1)}{\sqrt{1 + |\nabla f(x)|^2}}, \quad \text{where } X = (x, t) \in (\overline{\Omega} \setminus \{O\}) \times \mathbb{R},$$

denote the downward unit normal to the graph of f; in the capillary interpretation, $\vec{n}$ represents the inward unit normal with respect to the fluid region. Using comparison theorems (for example, [Finn 1986, Theorem 5.1]) and existence and regularity theorems for variational solutions (for example, [Finn 1986, Theorem 7.5 together with Lemma 4.1]), we see that we may assume f is a variational ($BV(\Omega)$) solution. Since our interest will be in the local behavior of solutions of (1) and (2) near the corner O, we sometimes think of Ω as the intersection of a larger domain with an appropriate neighborhood of O and a solution f of (1) and (2) as the restriction to $\overline{\Omega} \setminus \{O\}$ of a function F that is a solution of a boundary value problem, perhaps like (1) and (2), in this larger domain; in this case, restricting the problem to a subdomain Ω for which (γ, Ω, O) is admissible is straightforward.

The following theorem will establish the validity of the Concus–Finn conjecture.

Theorem 1.1. *Let Ω and γ be as above with $\alpha \in (0, \pi/2]$, and suppose that $f \in C^2(\Omega) \cap C^{1,\rho^*}(\overline{\Omega}\backslash\{O\})$ is a bounded solution to* (1) *satisfying* (2) *on $\partial^{\pm}\Omega\backslash\{O\}$ with $|H|_\infty = \sup_{\boldsymbol{x}\in\Omega} |H(\boldsymbol{x}, f(\boldsymbol{x}))| < \infty$. Suppose* (4) *holds and $\gamma_1, \gamma_2 \in (0, \pi)$. Then f is discontinuous at O whenever (γ_1, γ_2) satisfies*

$$2\alpha + |\gamma_1 - \gamma_2| > \pi. \tag{6}$$

Notice that we exclude cases in which γ_1 or γ_2 equals 0 or π. It seems likely that an argument in this exceptional situation might use ideas from [Finn 1988b], and it would be interesting to see the details of a proof.

For linear elliptic partial differential equations, especially uniformly elliptic equations, the qualitative behavior "at" a boundary point of the solution f of a boundary value problem can be determined by local information such as the prescribed boundary information and bounds on the maximum rate at which $|f|$ can go to infinity "at" the boundary point (for example, [Bear and Hile 1983]). However this is usually not true for quasilinear equations. The Concus–Finn conjecture, if true, represents one of the rare situations when the qualitative behavior of a solution (that is, its continuity at a convex corner) is determined by the boundary information (that is, α, γ_1 and γ_2) in an arbitrarily small neighborhood of the boundary point. At a nonconvex corner O (that is, $\alpha > \pi/2$), [Shi and Finn 2004] shows that information about $\partial\Omega \cap B_\epsilon(O)$ and γ in $B_\epsilon(O)$ for some $\epsilon > 0$ need not be sufficient to determine the continuity at O of a solution of (3) and (2).

Lancaster and Siegel [1996a] investigated the behavior of bounded solutions of (3) and (2) at corners, both convex and nonconvex corners, and they noted in [1996a; 1996b] that the conclusions in [1996a] carry over to solutions of (1) and (2) when H satisfies some minor restrictions (that is, $H(\boldsymbol{x}, z)$ is either real-analytic or strictly increasing in z); in this case, a bounded solution $f \in C^2(\Omega) \cap C^1(\overline{\Omega}\backslash\{O\})$ of (1) satisfying (2) on $\partial^{\pm}\Omega\backslash\{O\}$ is in $C^0(\overline{\Omega})$ when $(\gamma_1, \gamma_2) \in \overline{R} \cap (0, \pi) \times (0, \pi)$. The arguments in [Concus and Finn 1996] and [Finn 1996] continue to show that if $(\gamma_1, \gamma_2) \in D_1^{\pm}$, then either (1) and (2) has no solution in a neighborhood of O or f is unbounded in any neighborhood of O when H satisfies some extremely minor restrictions. Thus, under mild restrictions on H, Figure 2 continues to illustrate the behavior at O of solutions of (1) and (2). (See Remark 3.1 for a comment about [Lancaster and Siegel 1996a].)

Once we know that a solution of (1) and (2) is discontinuous at a convex corner $O = (0, 0)$, it is natural to ask about its behavior nearby. In [Lancaster and Siegel 1996a, Theorem 1], it is proven that if $\epsilon \le \gamma \le \pi - \epsilon$ for some $\epsilon > 0$, then the radial limits of f,

$$Rf(\theta) = \lim_{r\downarrow 0} f(r\cos(\theta), r\sin(\theta)),$$

exist for all $\theta \in (-\alpha, \alpha)$ and $Rf \in C^0([-\alpha, \alpha])$, where $Rf(-\alpha)$ and $Rf(\alpha)$ are the limits of the trace of f on $\partial^-\Omega$ and $\partial^+\Omega$ respectively; the continuity of the trace of f on $\overline{\partial^-\Omega}$ and on $\overline{\partial^+\Omega}$ is a conclusion of this theorem.

Now suppose (4) holds and $2\alpha + |\gamma_1 - \gamma_2| > \pi$. Then Theorem 1.1 above and [Lancaster and Siegel 1996a, Theorems 1 and 2] imply that there exist α_1 and α_2 with $-\alpha < \alpha_1 < \alpha_2 < \alpha$ such that

$$Rf(\theta) = \begin{cases} \text{constant} & \text{if } -\alpha \le \theta \le \alpha_1, \\ \text{strictly monotonic} & \text{if } \alpha_1 \le \theta \le \alpha_2, \\ \text{constant} & \text{if } \alpha_2 \le \theta \le \alpha \end{cases}$$

and $\alpha_1 - (-\alpha) \ge \pi - \gamma_2$ and $\alpha - \alpha_2 \ge \gamma_1$ if Rf is increasing on (α_1, α_2), while $\alpha_1 - (-\alpha) \ge \gamma_2$ and $\alpha - \alpha_2 \ge \pi - \gamma_1$ if Rf is decreasing on (α_1, α_2). Lancaster and Siegel [1996a] call the intervals $[-\alpha, \alpha_1]$ and $[\alpha_2, \alpha]$ *fans* (of constant radial limits), due to the shape of a region $\{(r\cos(\theta), r\sin(\theta)) : r > 0, \alpha_2 \le \theta \le \alpha\}$ on whose closure f is continuous; for nonconvex corners, a *central fan* (of constant radial limits) with size π can also exist. In particular, we see that Theorem 1.1 implies f has a jump discontinuity at O.

This work arose as a consequence of the *Summer School on Capillarity* held at the Max-Planck-Institut für Mathematik in Leipzig in June and July of 2003. While the Concus–Finn conjecture was discussed at meetings prior to 2003 (for example, the *International Conference on Differential Equations and Dynamic Systems*, University of Waterloo, Waterloo, Canada, August, 1997), the 2003 summer school brought together experts such as Maria Athanassenas, Robert Finn, Kirk Lancaster, John McCuan, Erich Miersemann, David Siegel, Tom Vogel and Henry Wente. In particular, Athanassenas and I worked (unsuccessfully) to find a counterexample to the Concus–Finn conjecture while others attempted to find a proof; our failure to find a counterexample together with the strong confidence in the correctness of the conjecture by others, especially John McCuan, inspired me to attempt to prove the conjecture. After the idea for a proof in the zero mean curvature case was obtained in 2004, Robert Finn strongly encouraged me to find a proof in the general case. In 2005, I did discover the idea of a proof; modulo some essentially minor technical modifications, this idea forms the basis for this work. This discovery may not have happened without the contributions of Athanassenas, Finn and McCuan. On the other hand, the absence of a subsequent summer school on capillarity, perhaps in the United States, may have delayed progress on important questions in capillarity (for example, [Athanassenas and Lancaster 2008; Finn 1999, 2002b; 2002a].)

2. Image of the Gauss map

In this section, we characterize in Theorem 2.1 the behavior of the limits at points of $\{O\} \times \mathbb{R}$ of the Gauss map for the graph of f. The proof involves the use of a 1975 result by Massari and Pepe [1975], generalized solutions (for example, [Giusti 1980]) and Leon Simon's capillarity paper [1980].

The following proposition is [Massari and Pepe 1975, Theorem 3], provided in translation for the convenience of the reader; the author thanks Professor Giuseppe Tenti of the Department of Applied Mathematics of the University of Waterloo for a translation of that paper. Here $\partial^* A$ denotes the reduced boundary of a Caccioppoli set A,

$$\nu_A(x) = \lim_{\rho \to 0} \frac{\int_{B(x,\rho)} D\phi_A}{\int_{B(x,\rho)} |D\phi_A|}$$

and $|\nu_A(x)| = 1$ for $x \in \partial^* A$; if ∂A is a C^1 hypersurface, $x \in \partial A$, and $\nu(x)$ is the interior unit normal to ∂A, then $\nu_A(x) = \nu(x)$; see for example [Giusti 1984, Chapter 3]. In the proposition, $\nu_h(x)$ denotes $\nu_{E_h}(x)$, $\nu(x)$ denotes $\nu_E(x)$, and Ω denotes an open set in $\mathbb{R}^n$; in the context used in this paper, such an open set might be $B^3(X, r)$ for $X \in \mathbb{R}^3$ and $r > 0$, or $\Omega_\infty \times \mathbb{R}$.

Proposition 2.1. *Let $\{E_h\}_h$ be a sequence of Caccioppoli sets of mean curvature $A_h \in L^p_{\text{loc}}(\Omega)$ with $p > n$. If*

$$\phi_{E_h}(x) \to \phi_E(x) \quad \textit{in } L^1_{\text{loc}}(\Omega), \tag{7}$$

$$\partial E_h \cap \Omega \ni x_h \to x \in \partial^* E \cap \Omega, \tag{8}$$

$$A_h(x) \to A(x) \quad \textit{in } L^1_{\text{loc}}(\Omega) \tag{9}$$

and if for every compact K of Ω there exists a constant $\gamma(K)$ such that

$$\|A_h\|_{L^p(K)} < \gamma(K) \quad \textit{for all } h \in \mathbb{N}, \tag{10}$$

then there exists $h_0 \in \mathbb{N}$, such that, for every $h > h_0$, we have

$$x_h \in \partial^* E_h \cap \Omega, \tag{11}$$

$$\lim_{h \to \infty} \nu_h(x_h) = \nu(x). \tag{12}$$

Remark 2.1. We define densities in the usual manner. If μ is a measure on $\mathbb{R}^n$ and $a \in \mathbb{R}^n$, we define the m-dimensional upper density $\Theta^{*m}(\mu, a)$, lower density $\Theta^m_*(\mu, a)$ and density $\Theta^m(\mu, a)$ of μ at a as in [Allard 1972]. For example,

$$\Theta^{*m}(\mu, x) = \limsup_{r \downarrow 0} \frac{\mu(B^n(x, r))}{\alpha_m r^m}.$$

If $A \subset \mathbb{R}^n$, $x \in \mathbb{R}^n$ and $m \leq n$, we define the m-dimensional upper (mass) density $\Theta^{*m}(A, x)$, the m-dimensional lower (mass) density $\Theta^m_*(A, x)$ and the m-dimensional (mass) density $\Theta^m(A, x)$ of A at x in the usual way. For example,

$$\Theta^m_*(A, x) = \liminf_{r \downarrow 0} \frac{H^m(B^n(x, r) \cap A)}{\alpha_m r^m};$$

here $\alpha_m = H^m(B^m(O_m, 1))$ denotes the m-dimensional volume of the unit ball in $\mathbb{R}^m$.

Recall that a m-dimensional varifold in $\mathbb{R}^n$ is a Radon measure on $\mathbb{R}^n \times \mathrm{G}(n, m)$. We denote the space of m-dimensional varifolds in $\mathbb{R}^n$ (with the weak topology) by $\mathbf{V}_m(\mathbb{R}^n)$. To each $\mathcal{H}^m$ measurable and $(\mathcal{H}^m, m)$ rectifiable set S in $\mathbb{R}^n$ is associated a varifold (for example [Allard 1972, Sections 3.5 and 4.7; Taylor 1976, Section I]); we adopt the notation $\mathbf{v}(S)$ of [Allard 1972] for this varifold, whereas [Taylor 1976] uses the notation $|S|$. We denote the first variation of $V \in \mathbf{V}_m(\mathbb{R}^n)$ by δV, as in [Allard 1972, Chapter 4].

For $r > 0$, let $\mu_r : \mathbb{R}^n \to \mathbb{R}^n$ be defined by $\mu_r(X) = rX$ for $X \in \mathbb{R}^n$. Let $V \in \mathbf{V}_m(\mathbb{R}^n)$. We set $V_r = \mu_{r\#}V$ (for example [Allard 1972, Section 3.2; Taylor 1976, Section I]); then

$$\|V_r\| = r^m \mu_{r\#}\|V\| \quad \text{and} \quad \|\delta V_r\| = r^{m-1}\mu_{r\#}\|\delta V\| \tag{13}$$

by [Allard 1972, 3.2(2) and 4.12(1)], respectively. Notice that if $L > 0$, then

$$\begin{aligned}\|\mu_{r\#}V\|(B(O_n, L)) &= r^m \mu_{r\#}\|V\|(B(O_n, L))\\ &= r^m\|V\|(B(O_n, L/r)) = L^m \frac{\|V\|(B(O_n, L/r))}{(L/r)^m}.\end{aligned}$$

Thus, if $\Theta^{*m}(\|V\|, O_n) < \infty$,

$$\limsup_{r\to\infty} \|\mu_{r\#}V\|(B(O_n, L)) \le L^m \alpha(m)\Theta^{*m}(\|V\|, O_n). \tag{14}$$

Similarly, if $k = m - 1$ and $\Theta^{*k}(\|\delta V\|, O_n) < \infty$, then

$$\limsup_{r\to\infty} \|\delta(\mu_{r\#}V)\|(B((O_n, L) \le L^k \alpha(k)\Theta^{*k}(\|\delta V\|, O_n). \tag{15}$$

Theorem 2.1. *Suppose Ω and γ are as in Theorem 1.1 such that* (4) *holds with $\gamma_1, \gamma_2 \in (0, \pi)$ and $\gamma_2 - \gamma_1 > \pi - 2\alpha$, that is, $(\gamma_1, \gamma_2) \in D_2^+$. Let $f \in C^2(\Omega) \cap C^{1,\rho^*}(\overline{\Omega} \setminus \{O\})$ be a bounded solution of* (1) *and* (2) *and suppose there exists $J \in (0, \infty)$ such that $|H(x, f(x))| \le J$ on $\Omega \times \mathbb{R}$. Let $\beta \in (-\alpha, \alpha)$ and let (x_j) be a sequence in Ω satisfying* $\lim_{j\to\infty} x_j = O$ *and*

$$\lim_{j\to\infty} x_j/|x_j| = (\cos(\beta), \sin(\beta)). \tag{16}$$

(i) *If $\beta \in [-\alpha + \pi - \gamma_2, \alpha - \gamma_1]$, then* $\lim_{j\to\infty} \vec{n}(x_j) = (-\sin(\beta), \cos(\beta), 0)$.

(ii) *If $\beta \in (-\alpha, -\alpha + \pi - \gamma_2]$, then*

$$\lim_{j\to\infty} \vec{n}(x_j) = (-\sin(-\alpha + \pi - \gamma_2), \cos(-\alpha + \pi - \gamma_2), 0).$$

(iii) *If $\beta \in [\alpha - \gamma_1, \alpha)$, then* $\lim_{j\to\infty} \vec{n}(x_j) = (-\sin(\alpha - \gamma_1), \cos(\alpha - \gamma_1), 0)$.

The proof consists of minor modifications of the proof of [Simon 1980] and the use of generalized solutions [Giusti 1980; Jeffres and Lancaster 2007]. The rationale for using results from [Simon 1980] and [Giusti 1980] is essentially the same as that used in [Tam 1986c]. (See Remark 2.3.) Simon's technique is the standard one (for example, [Federer 1969, Sections 3.1 and 5.4]) of blowing up the graph of a solution of (1) and (2) about the origin $O_3 \in \mathbb{R}^3$; Simon obtains a plane through the origin, and we modify that proof to show that the limit of a blow-up about O_3 of the graph of $f - Rf(\beta)$ is a vertical half-plane π_1. Unfortunately, the third component of the image $(x_{j_k}/\epsilon_{j_k}, [f(\epsilon_{j_k}x_{j_k}) - Rf(\beta)]/\epsilon_{j_k})$ of the blow-up sequence being used might diverge to infinity. We therefore consider a type of sequence introduced in [Tam 1986c] and use the result above, Proposition 2.1 and $BV(\Omega \times \mathbb{R})$ techniques (for example, [Jeffres and Lancaster 2007]) to determine the unit normal to π_1. One might wish to read Remark 2.4 before examining the proof of this theorem.

It will be convenient to define some quantities and state an assumption. Set

$$\begin{aligned}
\epsilon_0 &= \tfrac{1}{8}\min\{\gamma_1, \pi - \gamma_1, \gamma_2, \pi - \gamma_2\}, & \zeta &= \tfrac{1}{2}\pi - 2\epsilon_0,\\
c_1 &= \tfrac{1}{4}(\cos(2\epsilon_0) - |\cos(\gamma_1)|), & \lambda_1 &= (\cos(\alpha - \zeta), \sin(\alpha - \zeta), 0),\\
c_2 &= \tfrac{1}{4}(\cos(2\epsilon_0) - |\cos(\gamma_1)|), & \lambda_2 &= (\cos(-\alpha + \zeta), \sin(-\alpha + \zeta), 0),\\
C &= (\min\{\sin(\epsilon_0), c_1, c_2\})^{-1}.
\end{aligned}$$

A quick calculation shows $\liminf_{\partial^+\Omega \ni x \to 0}(-\nu(x)\cdot\lambda_1 + \cos(\gamma(x))\vec{n}(x)\cdot\lambda_1) \geq 4c_1$ and $\liminf_{\partial^-\Omega \ni x \to 0}(-\nu(x)\cdot\lambda_2 + \cos(\gamma(x))\vec{n}(x)\cdot\lambda_2) \geq 4c_2$. We will assume $\delta^* > 0$ was chosen small enough that

(a) $|\tau^+(x) - \alpha| < \alpha/4$ and $|\tau^-(x) + \alpha| < \alpha/4$ if $|x| \leq 3\delta^*$.

(b) $\Omega \cap B(O, 3\delta^*) \subset \{r\omega(\theta) : r > 0,\ \theta \in [-\alpha - \epsilon_0, \alpha + \epsilon_0]\}$.

(c) $-\nu(x)\cdot\lambda_1 + \cos(\gamma(x))\vec{n}(x)\cdot\lambda_1 \geq 2c_1$ if $x \in \partial^+\Omega$ and $|x| \leq 3\delta^*$.

(d) $-\nu(x)\cdot\lambda_2 + \cos(\gamma(x))\vec{n}(x)\cdot\lambda_2 \geq 2c_2$ if $x \in \partial^-\Omega$ and $|x| \leq 3\delta^*$.

Notice that (a) and (b) imply there exist $x^\pm : [0, 3\delta^*] \to \mathbb{R}^2$ that are parametrizations of $\partial^\pm\Omega$ such that $x = x^+(|x|)$ for $x \in \partial^+\Omega$ and $x = x^-(|x|)$ for $x \in \partial^-\Omega$. Let $\Omega_\lambda = \Omega \cap B(O, \lambda)$ for $\lambda > 0$.

Proof. Consider $\beta \in (-\alpha, \alpha)$ fixed and set $u(x) = f(x) - Rf(\beta)$, as in [Lancaster and Siegel 1996a]. Set $\delta_0 = 2\delta^*$. Let

$$U = \{(x, t) : x \in \Omega, t < u(x)\} \quad \text{be the subgraph of } u$$

and

$$\begin{aligned}\mathcal{M}_0 &= \{(x,u(x)) : x \in \Omega \cap B(O,3\delta^*)\},\\ \mathcal{M} &= \{(x,u(x)) : x \in \overline{\Omega \cap B(O,3\delta^*)} \setminus \{O\}\},\\ \partial^+\mathcal{M} &= \{(x,u(x)) : x \in \partial^+\Omega \setminus \{O\}\},\\ \partial^-\mathcal{M} &= \{(x,u(x)) : x \in \partial^-\Omega \setminus \{O\}\}.\end{aligned}$$

Notice that $\partial U \cap (\Omega_{3\delta^*} \times \mathbb{R}) = \mathcal{M}_0$. Let $V = \mathbf{v}(\mathcal{M})$ and $V_0 = \mathbf{v}(\mathcal{M}_0)$ and note that these are both two-dimensional integral varifolds; see for example [Allard 1972, Section 3.5].

We will first use a variation of the argument in [Simon 1980, Section 1][1] to show that

$$\mathcal{H}^1(\partial^+\mathcal{M} \cup \partial^-\mathcal{M}) < \infty. \tag{17}$$

As in [S], let η denote the unit vector normal to $\partial\mathcal{M} = \partial^+\mathcal{M} \cup \partial^-\mathcal{M}$ that is tangent to $\mathcal{M}$ and points into $\Omega \times \mathbb{R}$; in the notation here,

$$\eta(X) = \frac{-\nu(X) + (\vec{n}(X)\cdot\nu(X))\vec{n}(X)}{|-\nu(X) + (\vec{n}(X)\cdot\nu(X))\vec{n}(X)|} = \frac{-\nu(X) + \cos(\gamma)\vec{n}(X)}{|-\nu(X) + \cos(\gamma)\vec{n}(X)|}.$$

Let $h_1, h_2, s \in C^\infty(\mathbb{R})$ with $0 \le h_1(t), h_2(t), s(t) \le 1$ for $t \in \mathbb{R}$, such that $h_1 = 0$ on $(-\infty, -\alpha/2]$ and $h_1 = 1$ on $[\alpha/2, \infty)$, with $h_2 = 1 - h_1$ and $s(t) = 1$ if $|t| \le 2\delta^*$ and $s(t) = 0$ if $|t| \ge 3\delta^*$. Define $\phi_1, \phi_2 \in C^\infty((\overline{\Omega} \setminus \{O\}) \times \mathbb{R})$ such that

$$\phi_1(r\omega(\theta), z) = h_1(\theta)s(r)\lambda_1 \quad \text{and} \quad \phi_2(r\omega(\theta), z) = h_2(\theta)s(r)\lambda_2$$

for $0 < r < \infty$ and $\theta \in (-\pi, \pi)$ that satisfy $r\omega(\theta) \in \overline{\Omega} \setminus \{O\}$. Notice that $\sup r|D\phi_1| < \infty$ and $\sup r|D\phi_2| < \infty$. As in [S, (1.4)], we obtain

$$\rho^{-1}\int_{\mathcal{M}\cap[B(O,\rho)\times\mathbb{R}]} (\phi_1 \cdot \delta^M r)\, d\mathcal{H}^2 + \int_{\partial^+\mathcal{M}} \min\{r/\rho, 1\}\phi_1 \cdot \eta\, d\mathcal{H}^1 \\ = -\int_{\mathcal{M}} \min\{r/\rho, 1\}(\delta^M \cdot \phi_1 + H\nu \cdot \phi_1)\, d\mathcal{H}^2,$$

since $h_1(t) = 0$ if $t \le -\alpha/2$, and

$$\rho^{-1}\int_{\mathcal{M}\cap[B(O,\rho)\times\mathbb{R}]} (\phi_2 \cdot \delta^M r)\, d\mathcal{H}^2 + \int_{\partial^-\mathcal{M}} \min\{r/\rho, 1\}\phi_2 \cdot \eta\, d\mathcal{H}^1 \\ = -\int_{\mathcal{M}} \min\{r/\rho, 1\}(\delta^M \cdot \phi_2 + H\nu \cdot \phi_2)\, d\mathcal{H}^2,$$

since $h_2(t) = 0$ if $t \ge \alpha/2$. From (b) and (c), we see that

$$\phi_1(X) \cdot \eta(X) \ge c_1 h_1(\theta)s(r) \quad \text{if } X = (r\omega(\theta), z) \in \partial^+\mathcal{M} \text{ with } \theta \in (0, \alpha + \epsilon_0)$$

[1] In this proof, we refer to [Simon 1980] as [S].

and, from (b) and (d), that

$$\phi_2(X)\cdot\eta(X)\geq c_2h_2(\theta)s(r)\quad\text{if } X=(r\omega(\theta),z)\in\partial^-\mathcal{M}\text{ with }\theta\in(-\alpha-\epsilon_0,0).$$

Using the argument on [S, page 367], we obtain

$$\mathcal{H}^1(\partial^+\mathcal{M}\cap(B(O,\delta_0)\times\mathbb{R}))<\infty\quad\text{and}\quad\mathcal{H}^1(\partial^-\mathcal{M}\cap(B(O,\delta_0)\times\mathbb{R}))<\infty.$$

Since f, and so u, is in $C^{1,\rho^*}(\overline{\Omega}\setminus\{O\})$, we see that (17) holds.

As in the proof of [S, (1.8)], we see using [Allard 1972, 4.2, 4.3(5), 4.7] that (1), (17) and [S, (1.1)] imply

$$\|\delta V\|(B(O,r)\times\mathbb{R})\leq J\mathcal{H}^2(\mathcal{M}\cap(B(O,r)\times\mathbb{R}))+\mathcal{H}^1(\partial\mathcal{M}\cap(B(O,r)\times\mathbb{R}))$$

and therefore

$$\|\delta V\|(B(O,r)\times\mathbb{R})<\infty\quad\text{for } 0<r<R_2. \tag{18}$$

Set $K=\max\{\sup_{\Omega\times\mathbb{R}}|\delta^M\cdot\phi_1|,\sup_{\Omega\times\mathbb{R}}|\delta^M\cdot\phi_2|\}$.

Now let us substitute in [S, (1.4)] successively $\phi=\phi_1\psi$ and $\phi=\phi_2\psi$, where $\psi\in C_0^1(B(O,3\delta^*)\times\mathbb{R})$. If we argue as in [S], we obtain the following analogues for $k=1,2$ of [S, (1.10)]:

$$\rho^{-1}\int_{\mathcal{M}\cap(B(O,\rho)\times\mathbb{R})}\psi(\phi_k\cdot Dr)d\mathcal{H}^2+\int_{\partial\mathcal{M}}\psi(\phi_k\cdot\eta)d\mathcal{H}^1$$
$$\leq(K+J))\int_{\mathcal{M}}(\psi+|\delta^M\psi|)d\mathcal{H}^2+o(1)\quad\text{as }\rho\to0.$$

Now (b) implies $\lambda_1\cdot Dr\geq\sin(\epsilon_0)$ on the support of ϕ_1 and $\lambda_2\cdot Dr\geq\sin(\epsilon_0)$ on the support of ϕ_2. Therefore, if $0<\rho<\delta_0$, then

$$\limsup_{\rho\downarrow0}\rho^{-1}\int_{\mathcal{M}\cap(B(O,\rho)\times\mathbb{R})}h_1\psi d\mathcal{H}^2+\int_{\partial\mathcal{M}}h_1\psi d\mathcal{H}^1$$
$$\leq C(K+J)\int_{\mathcal{M}}\big(h_1\psi+|\delta^M(h_1\psi)|\big)d\mathcal{H}^2$$

and

$$\limsup_{\rho\downarrow0}\rho^{-1}\int_{\mathcal{M}\cap(B(O,\rho)\times\mathbb{R})}h_2\psi d\mathcal{H}^2+\int_{\partial\mathcal{M}}h_2\psi d\mathcal{H}^1$$
$$\leq C(K+J)\int_{\mathcal{M}}\big(h_2\psi+|\delta^M(h_2\psi)|\big)d\mathcal{H}^2.$$

By adding these inequalities, we see that if $0 < \rho < \delta_0$ then

$$\begin{aligned}\limsup_{\rho\downarrow 0}\rho^{-1}\int_{\mathcal{M}\cap(B(O,\rho)\times\mathbb{R})}\psi\, d\mathcal{H}^2+\int_{\partial\mathcal{M}}\psi\, d\mathcal{H}^1 &\le C(K+J)\int_{\mathcal{M}}\big(\psi+|\delta^M(h_1\psi)|+|\delta^M(h_2\psi)|\big)d\mathcal{H}^2\\ &\le C(K+J)\int_{\mathcal{M}}\big(\psi(1+|\delta^M(h_1)|+|\delta^M(h_2)|)+|\delta^M(\psi)|\big)d\mathcal{H}^2.\end{aligned}$$

From the first part of [Allard 1972, 3.1(2)], we see this implies for the varifold $V=\mathbf{v}(\mathcal{M})$ that

$$\|\delta V\|(\psi)\le C(K+J)\int\big(\psi(1+|\delta^M(h_1)|+|\delta^M(h_2)|)+|\delta^M(\psi)|\big)d\|V\|, \tag{19}$$

which is an analogue of [S, (1.11)]. As in [S], this implies

$$\mathcal{H}^2(\mathcal{M}\cap B^3(Y,\rho))\ge \bar{C}\rho^2(1+\delta_0), \tag{20}$$

for some constant $\bar{C}>0$, and therefore

$$\Theta^2_*(\|V\|,Y)\ge\bar{C}>0 \tag{21}$$

if $0<\rho<\delta_0$ and $Y\in\bar{\mathcal{M}}\cap(B(O,\sigma)\times\mathbb{R})$. (These two conclusions can be obtained independently using $BV(\Omega)$ techniques and Lemma 3.1.)

Let

$$\begin{aligned}F_1&=\{(x,t): x\in\partial^+\Omega\setminus\{O\}, t\le u(x)\},\\ \tilde{F}_1&=\{(x,t): x\in\partial^+\Omega\setminus\{O\}, t\ge u(x)\},\\ F_2&=\{(x,t): x\in\partial^-\Omega\setminus\{O\}, t\le u(x)\}, \text{ and}\\ \tilde{F}_2&=\{(x,t): x\in\partial^-\Omega\setminus\{O\}, t\ge u(x)\}.\end{aligned}$$

Let $W_1=\mathbf{v}(F_1)$, $\tilde{W}_1=\mathbf{v}(\tilde{F}_1)$, $W_2=\mathbf{v}(F_2)$ and $\tilde{W}_2=\mathbf{v}(\tilde{F}_2)$, be the two-dimensional varifolds associated with F_1, $\tilde{F}_1$, F_2 and $\tilde{F}_2$, respectively (for example, [Allard 1972, Sections 3.5 and 4.7] and [Taylor 1976, Section 1]). Set

$$E_1=\{x\in\partial\Omega:\gamma(x)<\tfrac{1}{2}\pi\}\times\mathbb{R}\quad\text{and}\quad E_2=\{x\in\partial\Omega:\gamma(x)>\tfrac{1}{2}\pi\}\times\mathbb{R}.$$

Define Z to be the two-dimensional varifold given by

$$Z=V-W_1\llcorner\cos(\gamma)\chi_{E_2}+\tilde{W}_1\llcorner\cos(\gamma)\chi_{E_1}-W_2\llcorner\cos(\gamma)\chi_{E_2}+\tilde{W}_2\llcorner\cos(\gamma)\chi_{E_1}.$$

The monotonicity formula [S, (2.6)] holds for Z; that is, there exists $c\ge 0$ such that

$$\exp(cr^\beta)\frac{\|Z\|(B^3(O_3,r))}{r^2}\quad\text{is increasing in } r \text{ for } 0<r<R, \tag{22}$$

and, in conjunction with (21), we see that the two-dimensional density of Z at O_3 exists and

$$\Theta^2(\|Z\|, O_3) \in (0, \infty). \tag{23}$$

We note, for example, that if $\gamma_1 = \pi/2$, then (5) is used in a slicing argument (that is, [Allard 1972, 4.10(1)]) to show that [S, (2.3)] (with $\partial\Omega$ replaced by $\partial^+\Omega$) holds.

Suppose (x_j) is a sequence in Ω converging to O as $j \to \infty$ and satisfying (16). For each $j \in \mathbb{N}$, set $\epsilon_j = |x_j|$ and $\Omega_j = \{x \in \mathbb{R}^2 : \epsilon_j x \in \Omega\}$, and define $f_j, u_j \in C^2(\Omega_j) \cap C^1(\overline{\Omega_j} \setminus \{O\})$ by

$$f_j(x) = \frac{f(\epsilon_j x) - f(x_j)}{\epsilon_j} \quad \text{and} \quad u_j(x) = \frac{f(\epsilon_j x) - Rf(\beta)}{\epsilon_j} = \frac{u(\epsilon_j x)}{\epsilon_j};$$

notice that $\nabla u_j = \nabla f_j$ on Ω_j and $u_j(x) = f_j(x) + c_j$ if $c_j = (f(x_j) - Rf(\beta))/\epsilon_j$. Let $\vec{n}_j$ be the downward unit normal to the graph of f_j (and the graph of u_j), so that

$$\vec{n}_j(x) = \vec{n}(\epsilon_j x) = \left(Tf_j(x), \frac{-1}{\sqrt{1+|\nabla f_j(x)|^2}}\right) \quad \text{for } x \in \Omega_j.$$

Let $U_j = \{(x,t) \in \Omega_j \times \mathbb{R} : t < u_j(x)\}$ be the subgraph of u_j for each $j \in \mathbb{N}$. Notice that $\mu_{1/\epsilon_j}(\mathcal{M}_0) = \partial U_j \cap (\Omega_j \times \mathbb{R})$ and $\mu_{1/\epsilon_j}(\mathcal{M}) \subset \partial U_j \cap ((\overline{\Omega_j} \setminus \{O\}) \times \mathbb{R})$.

From [Allard 1972, 2.6(2)(a)] with $\mathcal{G} = \{B(O_3, L) \times G(3,2) : L > 0\}$, we see that (14) implies that there is a subsequence (ϵ_{j_k}) of (ϵ_j) and a varifold $C \in \mathbf{V}_2(\mathbb{R}^3)$ in the varifold tangent of Z at O such that

$$C = \lim_{k\to\infty} Z_{1/\epsilon_{j_k}},$$

where $Z_{1/\epsilon_{j_k}} = \mu_{1/\epsilon_{j_k}\#}(Z)$. By (14) and [Allard 1972, 2.6(2)(c)],

$$\|C\|(B(O_3, L)) = C(B(O_3, L) \times G(3,2)) \le L^2\alpha(2)\Theta^2(\|Z\|, O_3);$$

from (22), we see that $\mu_{r\#}\|C\| = \|C\|$ for all $r > 0$ (as observed in [Simon 1980], p. 576). Since $\|V\|(\Omega \times \mathbb{R}) = \|Z\|(\Omega \times \mathbb{R})$ and (17) holds (hence $\mathcal{H}^2(\mathcal{M} \cap (\partial\Omega \times \mathbb{R})) = 0$), we see that

$$\Theta^{*2}(\|V\|, O_3) \le \Theta^2(\|Z\|, O_3) < \infty.$$

Using (14) and [Allard 1972, 2.6(2)(a)], we notice that there is a subsequence of (ϵ_{j_k}), still denoted (ϵ_{j_k}), and a varifold $V_\infty \in \mathbf{V}_2(\mathbb{R}^3)$ in the varifold tangent of V at O_3 such that

$$V_\infty = \lim_{k\to\infty} V_{1/\epsilon_{j_k}}$$

and, by (14) and [Allard 1972, 2.6(2)(c)],

$$\|V_\infty\|(B(O_3, L)) = V_\infty(B(O_3, L) \times G(3,2)) \le L^2\alpha(2)\Theta^2(\|Z\|, O_3).$$

In a similar manner (as in [S, page 370]), we see that

$$\begin{aligned}
W_{1,\infty} &= \lim_{k\to\infty} \mu_{1/\epsilon_{j_k}}{}_{\#}(W_1 \llcorner a(-\cos(\gamma_1))\cos(\gamma_1)\chi_{F_1}),\\
\tilde{W}_{1,\infty} &= \lim_{k\to\infty} \mu_{1/\epsilon_{j_k}}{}_{\#}(\tilde{W}_1 \llcorner a(\cos(\gamma_1))\cos(\gamma_1)\chi_{\tilde{F}_1}),\\
W_{2,\infty} &= \lim_{k\to\infty} \mu_{1/\epsilon_{j_k}}{}_{\#}(W_2 \llcorner a(-\cos(\gamma_2))\cos(\gamma_2)\chi_{F_2}),\\
\tilde{W}_{2,\infty} &= \lim_{k\to\infty} \mu_{1/\epsilon_{j_k}}{}_{\#}(\tilde{W}_2 \llcorner a(\cos(\gamma_2))\cos(\gamma_2)\chi_{\tilde{F}_2})
\end{aligned}$$

all exist and

$$C = V_\infty - \cos(\gamma_1)W_{1,\infty} + \cos(\gamma_1)\tilde{W}_{1,\infty} - \cos(\gamma_2)W_{2,\infty} + \cos(\gamma_2)\tilde{W}_{2,\infty}.$$

Notice that $W_{1,\infty} = 0$ if $\cos(\gamma_1) < 0$ and $\tilde{W}_{1,\infty} = 0$ if $\cos(\gamma_1) > 0$, and that $W_{2,\infty} = 0$ if $\cos(\gamma_2) < 0$ and $\tilde{W}_{2,\infty} = 0$ if $\cos(\gamma_2) > 0$.

Using the arguments in [S, Section 3 up to the top of page 373 (including (3.5)′)], we see the following.

(i) For each $\rho > 0$, there is a sequence $\{\delta_k\}$ of positive reals that converges to zero such that

$$B^3(O_3,\rho)\cap \mathcal{M}_{j_k} \subset \{Y \in B^3(O_3,\rho) : \operatorname{dist}(Y, \operatorname{spt}(\|V_\infty\|)) < \delta_k\},$$

where $\mathcal{M}_{j_k} = \mu_{1/\epsilon_{j_k}}\mathcal{M}$ for each $k \in \mathbb{N}$ (that is, [S, (2.7)].)

(ii) $\mathcal{M}_\infty = \lim_{k\to\infty} \mathcal{M}_{j_k}$, taken in $\boldsymbol{\Omega}_\infty \times \mathbb{R}$ in the varifold sense, exists, and we have

$$V_\infty \llcorner (\boldsymbol{\Omega}_\infty \times \mathbb{R}) = \mathbf{v}(\mathcal{M}_\infty)$$

and

$$\mu_r(\mathcal{M}_\infty) = \mathcal{M}_\infty \quad \text{for } r > 0 \quad \text{(that is, } \mathcal{M}_\infty \text{ is a cone).}$$

(iii) $\mathcal{M}_\infty$ is empty or $\mathcal{M}_\infty = \bigcup_{j=1}^N \pi_j \cap (\boldsymbol{\Omega}_\infty \times \mathbb{R})$, where the π_j are planes through the origin and $\pi_i \cap \pi_j \cap (\boldsymbol{\Omega}_\infty \times \mathbb{R}) = \varnothing$ if $i \neq j$.

(iv) Either

Case 1. $N = 1$ and $\mathcal{M}_\infty = \pi_1 \cap (\boldsymbol{\Omega}_\infty \times \mathbb{R})$ for some plane π_1 whose intersection with $\{O\}\times\mathbb{R}$ is $\{O_3\}$; or

Case 2. $N < \infty$ and $\mathcal{M}_\infty = \bigcup_{j=1}^N \pi_j \cap (\boldsymbol{\Omega}_\infty \times \mathbb{R})$, where $\pi_1, \dots, \pi_N$ are planes with the line $\{O\}\times\mathbb{R}$ in common.

(v) The subgraphs U_{j_k} of u_{j_k} and $U_\infty = \lim_{k\to\infty} \mu_{1/\epsilon_{j_k}}(U)$ minimize appropriate functionals (for example, [S, (3.4)′]).

Using (19) and arguing as in the proof of [S, (3.7), pages 373–4], we see that

$$\mathcal{M}_\infty \neq \varnothing \quad \text{and} \quad V_\infty = \mathbf{v}(\mathcal{M}_\infty).$$

(We note that this argument, specifically in the paragraph after [S, (3.7)], implies $\Theta^{*1}(\|\delta V\|, O_3) < \infty$ and allows (15) to be used.) In particular, spt $(\|V_\infty\|) = \mathcal{M}_\infty$ and so (i) says that for each $\rho > 0$, there is a sequence $\{\delta_k\}$ of positive reals that converges to zero such that

$$B^3(O_3,\rho) \cap \mathcal{M}_{j_k} \subset \{Y \in B^3(O_3,\rho) : \operatorname{dist}(Y, \mathcal{M}_\infty) < \delta_k\}. \tag{24}$$

The conclusions in [S, Sections 2 and 3 up to, but not including, the paragraph containing (3.17)] hold and imply that $N = 1$, $\mathcal{M}_\infty = \pi_1 \cap (\Omega_\infty \times \mathbb{R})$ and either

- $\pi_1 \cap (\{O\} \times \mathbb{R}) = \{O_3\}$, or
- $\{O\} \times \mathbb{R} \subset \pi_1$.

(See also [Jeffres and Lancaster 2007].) We observe that the first is impossible when $(\gamma_1, \gamma_2) \in D_2^+$ (or $(\gamma_1, \gamma_2) \in D_2^-$) since no plane can meet $\partial^+\Omega_\infty \times \mathbb{R}$ in angle γ_1 and $\partial^-\Omega_\infty \times \mathbb{R}$ in angle γ_2 (as Concus and Finn [1996] observed and an easy calculation confirms). Therefore there exists $\xi_1, \xi_2 \in \mathbb{R}$ with $\xi_1^2 + \xi_2^2 = 1$ and $\xi = (\xi_1, \xi_2, 0) \in S^2$ such that

$$\pi_1 = \{X \in \mathbb{R}^3 : X \cdot \xi = 0\} \quad \text{and} \quad U_\infty = \{X \in \Omega_\infty \times \mathbb{R} : X \cdot \xi > 0\}.$$

Hence

$$\mathcal{M}_\infty = \{X \in \mathbb{R}^3 : X \cdot \xi = 0\} \cap (\Omega_\infty \times \mathbb{R}) \tag{25}$$

and we may write $U_\infty = U_\infty^{(1)} \times \mathbb{R}$, where $U_\infty^{(1)} = \{x \in \Omega_\infty : \xi \cdot (x, 0) > 0\}$. Using the arguments in [S, pages 374–5], which yield [S, (3.13), (3.15)–(3.16) and (3.18)–(3.18)′] and, for example, defining

$$\begin{aligned} E_\infty^{(1)}(W) = \mathcal{H}^1(\partial W \cap \Omega_\infty \cap B(O,1)) - \cos(\gamma_1)\mathcal{H}^1(\partial W \cap \partial^+\Omega_\infty \cap B(O,1)) \\ - \cos(\gamma_2)\mathcal{H}^1(\partial W \cap \partial^-\Omega_\infty \cap B(O,1)) \end{aligned}$$

for any open set $W \subset \Omega_\infty$ satisfying

$$\mathcal{H}^1(\partial W \cap B(O,1)) < \infty \quad \text{and} \quad (W \triangle U_\infty^{(1)}) \cap B(O,1) \subset\subset B(O,1),$$

we obtain

$$E_\infty^{(1)}(U_\infty^{(1)}) \le E_\infty^{(1)}(W)$$

for any set W as described above (compare with [S, (3.16)].)

Note that $f_j \in BV(\Omega_j)$ is a variational solution and hence a generalized solution of (1) and (2) with Ω, γ and $H(x,z)$ replaced by Ω_j, γ_j and $H_j^*(x) = \epsilon_j H^*(\epsilon_j x)$ (with H^* as in (46)) respectively. By Lemma 3.2, (f_{j_k}) has a subsequence, still denoted (f_{j_k}), that converges to a generalized solution $f_\infty : \Omega_\infty \to [-\infty, \infty]$ of

$$\begin{aligned} Nv &= 0 && \text{in } \Omega_\infty, \\ Tv \cdot \nu_\infty^+ &= \cos(\gamma_1) && \text{almost everywhere on } \partial^+\Omega_\infty, \\ Tv \cdot \nu_\infty^- &= \cos(\gamma_2) && \text{almost everywhere on } \partial^-\Omega_\infty, \end{aligned}$$

where $\nu_\infty^+ = (\cos(\alpha + \frac{1}{2}\pi), \sin(\alpha + \frac{1}{2}\pi))$ and $\nu_\infty^- = (\cos(-\alpha - \frac{1}{2}\pi), \sin(-\alpha - \frac{1}{2}\pi))$. Let us denote the subgraph of f_j by

$$U_j^* = \{(x, t) \in \Omega_j \times \mathbb{R} : t < f_j(x)\} \quad \text{for } j \in \mathbb{N}, \tag{26}$$

and denote by U_∞^* the subgraph of f_∞. Notice that (16) and $f_j(x_j/|x_j|) = 0$ for $j \in \mathbb{N}$ imply

(a) if K is open with $K \subset\subset \Omega_\infty$ and $(\cos(\beta), \sin(\beta)) \in K$, then there exists $m(K) \in \mathbb{N}$ such that $K \subset \Omega_j$ and $x_j/|x_j| \in K$ whenever $j \geq m(K)$;

(b) $(x_j/|x_j|, 0) \in \partial U_j^*$ for $j \in \mathbb{N}$; and

(c) $(x_j/|x_j|, 0) \to (\cos(\beta), \sin(\beta), 0)$ as $j \to \infty$.

Set $x_\beta = (\cos(\beta), \sin(\beta))$ and $X_\beta = (\cos(\beta), \sin(\beta), 0)$. From interior density bounds (for example, [Tam 1986b, Lemma 3.1]), we see that $X_\beta \in \partial U_\infty^* \cap (\Omega_\infty \times \mathbb{R})$. Notice that $\mathcal{M}^* = \partial U_\infty^* \cap (\Omega_\infty \times \mathbb{R})$ is a smooth surface whose ("downward") unit normal can be denoted by $\vec{\chi}(X) = (\chi_1(X), \chi_2(X), \chi_3(X))$ for $X \in \mathcal{M}^*$; then $\chi_3(X) \leq 0$ for all $X \in \mathcal{M}^*$. By Proposition 2.1, we see that

$$\vec{n}_{j_k}(y_k) \to \vec{\chi}(X) \quad \text{as } k \to \infty \quad \text{whenever } (y_k, f_{j_k}(y_k)) \to X \in \mathcal{M}^*; \tag{27}$$

in particular, (c) implies $\vec{n}_{j_k}(x_{j_k}/\epsilon_{j_k}) \to \vec{\chi}(X_\beta)$ as $k \to \infty$ (with the set Ω in the proposition being a neighborhood of X (or X_β) in $\mathbb{R}^3$.) We claim that either

($\aleph$) $\chi_3(X) < 0$ for all $X \in \mathcal{M}^*$ or

(ω) $\vec{\chi}$ is constant, $\chi_3(X) = 0$ for all $X \in \mathcal{M}^*$ and $\mathcal{M}^*$ is the intersection of $\Omega_\infty \times \mathbb{R}$ with the vertical plane π_2 containing X_β and normal to $\vec{\chi}$.

(To see this, we may represent the minimal surface $\mathcal{M}^*$ in isothermal coordinates as the (downward oriented) parametric surface $X : B(O, 1) \to \mathbb{R}^3$ (for example, [Courant 1977; Lancaster 1985; Elcrat and Lancaster 1986; Lancaster and Siegel 1996a]) and obtain the Weierstrass (f, g)-representation of $\mathcal{M}^*$, where

$$g(w) = S(\vec{\chi}(X(u, v))) \quad \text{for } w = u + iv \in \mathbb{C}, \ |w| < 1,$$

is the composition of the (north pole) stereographic projection S with the Gauss map $\vec{\chi} \circ X : B(O, 1) \to S_-^2$. Then g is a holomorphic map from the open unit ball in $\mathbb{C}$ into the closed unit ball in $\mathbb{C}$. If $\chi_3(X_p) = 0$ for some $X_p \in \mathcal{M}^*$, then

$X_p = X(u_p, v_p)$ for some $(u_p, v_p) \in B(O, 1)$ and $|g(w_p)| = 1$ for $w_p = u_p + iv_p$; the maximum modulus principle implies g is constant. The claim follows.)

Suppose ($\aleph$) holds and $\chi_3(X) < 0$ for all $X \in \mathcal{M}^*$. Then $f_\infty \in C^2(\Omega_\infty)$ and $\mathcal{M}^*$ is the graph of f_∞ over Ω_∞. Let $0 < R < \bar{R} < \operatorname{dist}(x_\beta, \partial\Omega_\infty)$; then there exists $L \geq 0$ such that $|\nabla f_\infty(x)| \leq L$ for all $x \in \overline{B(x_\beta, R)}$. Now (27) together with the uniform interior Hölder estimates for the unit normal (or Gauss) map of the graphs of solutions of (1) (for example, [Gilbarg and Trudinger 1983, Theorem 16.18] with $K = -1$ and $K' = J(\bar{R} - R)^2$ or [S, (3.1)]) imply there exists $K(R, \bar{R})$ such that if $k \in \mathbb{N}$ satisfies $k \geq K(R, \bar{R})$, then $\overline{B(x_\beta, \bar{R})} \subset \Omega_{j_k}$ and

$$|\nabla u_{j_k}(x)| \leq L + 1 \quad \text{for all } x \in \overline{B(x_\beta, R)}. \tag{28}$$

Notice that for k large enough, (28) contradicts (24) and (25) and so ($\aleph$) cannot hold. Therefore (ω) holds, $\mathcal{M}^*$ is the intersection of $\Omega_\infty \times \mathbb{R}$ with the plane π_2, we may write

$$\vec{\chi} = (\cos(\theta), \sin(\theta), 0) \quad \text{for some } \theta \in (-\pi, \pi]$$

and $U^*_\infty = \{X \in \Omega_\infty \times \mathbb{R} : (X - X_\beta) \cdot \vec{\chi} > 0\}$. (Notice then that (24) and (25) imply $\vec{\chi} = \vec{\xi}$ and $\pi_2 = \pi_1$.)

We will use the theory of generalized solutions (for example, [Giusti 1980]) to determine θ. We claim that

$$\theta = \begin{cases} -\alpha + \pi - \gamma_2 + \pi/2 & \text{if } \beta \in (-\alpha, -\alpha + \pi - \gamma_2], \\ \beta + \pi/2 & \text{if } \beta \in [-\alpha + \pi - \gamma_2, \alpha - \gamma_1], \\ \alpha - \gamma_1 + \pi/2 & \text{if } \beta \in [\alpha - \gamma_1, \alpha). \end{cases} \tag{29}$$

The sets

$$\mathcal{P} = \{x \in \Omega_\infty : f_\infty = \infty\} \quad \text{and} \quad \mathcal{N} = \{x \in \Omega_\infty : f_\infty = -\infty\} \tag{30}$$

each minimize an appropriate functional, and the arguments in [JL][2] show that $U_\infty = \mathcal{P} \times \mathbb{R}$, where $\mathcal{P}$ is given in one of [JL, (iv), (vi) or (viii) of Theorem 1] and $\mathcal{N} = \Omega_\infty \setminus \overline{\mathcal{P}}$.

Suppose $\beta \in [-\alpha + \pi - \gamma_2, \alpha - \gamma_1]$ holds. We see that [JL, Theorem 1, case (viii)] must hold. Since $\partial\mathcal{P}$ is a line going through O and $(\cos(\beta), \sin(\beta))$, we have $\vec{\chi} = (-\sin(\beta), \cos(\beta), 0)$ and $\theta = \beta + \pi/2$.

Suppose $\beta \in (-\alpha, -\alpha + \pi - \gamma_2]$ holds. Then [JL, Theorem 1, case (vi)] must hold, $\vec{\chi} = (-\sin(-\alpha + \pi - \gamma_2), \cos(-\alpha + \pi - \gamma_2), 0)$ and $\theta = -\alpha + \pi - \gamma_2 + \pi/2$.

Finally, suppose $\beta \in [\alpha - \gamma_1, \alpha)$. Then [JL, Theorem 1, case (iv)] must hold, $\vec{\chi} = (-\sin(\alpha - \gamma_1), \cos(\alpha - \gamma_1), 0)$ and $\theta = \alpha - \gamma_1 + \pi/2$. Our claim (29) is therefore proven.

[2] Here [JL] stands for [Jeffres and Lancaster 2007].

We have taken an arbitrary sequence (x_j) in Ω that satisfies (16) and shown that it has a subsequence (x_{j_k}) for which $(\vec{n}(x_{j_k}))$ converges to $\vec{\chi} = (\cos(\theta), \sin(\theta), 0)$ with θ given by (29). Therefore, if (y_j) is any sequence in Ω that satisfies (16) such that $\lim_{j\to\infty} \vec{n}(y_j) = \vec{\lambda}$ for some $S^2 \ni \vec{\lambda} \neq \vec{\chi}$, it must have a subsequence (y_{j_k}) for which $\vec{n}(y_{j_k})$ converges to both $\vec{\lambda}$ and $\vec{\chi}$, which is a contradiction. Thus, we see that the conclusion of Theorem 2.1 follows. □

Remark 2.2. Notice that $(\gamma_1, \gamma_2) \in D_2^-$ if and only if $(\gamma_2, \gamma_1) \in D_2^+$ and therefore we see that the conclusion of an appropriate version of Theorem 2.1 for the situation where $(\gamma_1, \gamma_2) \in D_2^-$ follows using a reflection in the x-axis and Theorem 2.1.

Remark 2.3. The proof of [Tam 1986c, Section 1] is essentially the same as that used in [Simon 1980] with the modification that [S, (1.12)] does not hold, the two-dimensional density $\Theta^2(\|Z\|, O_3) = 0$ and $\mathcal{M}_\infty = \varnothing$. Unfortunately the proofs of the claim in [Tam 1986c, Section 2] that (i) a subsequence of $\{f_j\}$ (called $\{u_j\}$ therein) converges "locally to a generalized solution" f_∞ (called therein u_∞) (that is, $\phi_{U_j^*} \to \phi_{U_\infty^*}$ in $L^1_{\text{loc}}(\Omega_\infty \times \mathbb{R})$ with U_j^* given by (26)) and (ii) the "graph" of this generalized solution (that is, ∂U_∞^*) is a vertical plane are absent; the "blow-up" in that section does not correspond to the process of blowing up with respect to a fixed point (that is, O_3) used in [Simon 1980]. (In spite of this, the ideas in [Tam 1986c, Section 2] are remarkable.) One difficulty is that even if a subsequence of $\{f_j\}$ should happen to converge in the sense of [Giusti 1980] to a generalized solution h_∞, the technique used here (for example, (24), (25), (28)) to show that ∂U_∞^* is a vertical plane cannot be used in [Tam 1986c] to show that h_∞ is the generalized solution u_∞ illustrated in [Tam 1986c, Figure 2 (see page 478)]. Even if Tam's proof can be correctly completed, the details would be sufficiently nontrivial that they should be provided to the reader. This new proof might be somewhat similar in outline to that of Theorem 2.1 above. (Of course, if $\alpha+\gamma < \pi/2$ in [Tam 1986c], no such proof could exist; the potential correction would need to be cognizant of this fact.)

Remark 2.4. In some uses of geometric measure theory in the literature (for example, [Allard 1972; Taylor 1977]), the authors leave important details to the reader or adopt a glib, hand waving, style. In this style, the proof of Theorem 2.1 can be shortened to the following:

Proof sketch. Suppose (x_j) is a sequence in Ω converging to O as $j \to \infty$ and satisfying (16). For each $j \in \mathbb{N}$, set $\epsilon_j = |x_j|$ and $\Omega_j = \{x \in \mathbb{R}^2 : \epsilon_j x \in \Omega\}$, and define $f_j, u_j \in C^2(\Omega_j) \cap C^1(\overline{\Omega_j} \setminus \{O\})$ by

$$f_j(x) = \frac{f(\epsilon_j x) - f(x_j)}{\epsilon_j} \quad \text{and} \quad u_j(x) = \frac{f(\epsilon_j x) - Rf(\beta)}{\epsilon_j}.$$

Using the techniques and results in [Simon 1980], we see that there is a vertical plane π_1 containing O_3 with unit normal $\vec{\chi}$ such that for each $\rho > 0$, there is a sequence $\{\delta_k\}$ of positive reals that converges to zero such that

$$B^3(O_3, \rho) \cap \mathcal{M}_{j_k} \subset \{Y \in B^3(O_3, \rho) : \operatorname{dist}(Y, \pi_1) < \delta_k\}, \tag{31}$$

where $\mathcal{M}_j = \mu_{1/\epsilon_j}\mathcal{M}$ for each $k \in \mathbb{N}$; recall that $\mu_r(X) = rX,\ \ X \in \mathbb{R}^3$, and $\mathcal{M} = \{(x, f(x) - Rf(\beta)) : x \in \overline{\Omega \cap B(O, 3\delta^*)} \setminus \{O\}\}$.

Now the sequence $\{f_j\}$ has a subsequence that converges (as in [Giusti 1980]) to a generalized solution f_∞ of (1) and (2) (this is Lemma 3.2). Since $f_j(x_j) = 0$, we have $(x_j, 0) \in \partial U_j^*$ for each $j \in \mathbb{N}$, where $U_j^* = \{(x, t) \in \Omega_j \times \mathbb{R} : t < f_j(x)\}$. Interior density bounds (for example, [Tam 1986b, Lemma 3.1]) imply

$$(\cos(\beta), \sin(\beta), 0) \in \partial U_\infty^* \cap (\Omega_\infty \times \mathbb{R}),$$

where U_∞^* is the subgraph of f_∞. If ∂U_∞^* is not a vertical plane, then [Massari and Pepe 1975, Theorem 3] (that is, Proposition 2.1) and [Gilbarg and Trudinger 1983, Theorem 16.18] imply a uniform bound on $|\nabla f_j|$ in a neighborhood of $(\cos(\beta), \sin(\beta))$ in Ω_∞, and this contradicts (31) since $\nabla u_j = \nabla f_j$ for each $j \in \mathbb{N}$. The conclusions of Theorem 2.1 now follow from [Jeffres and Lancaster 2007, Theorem 1].

3. Proof of Theorem 1.1

The proof of this theorem uses the conformal (or isothermal) representation of a prescribed mean curvature surface discussed in [Lancaster and Siegel 1996a] and properties of two-dimensional quasiconformal maps to obtain a contradiction to the assumption that the solution f is continuous at the origin. This proof uses Kenmotsu's theorem [1979], Theorem 2.1, Gehring's lemma [1973] and properties of solutions of Riemann–Hilbert problems to obtain a contradiction, illustrated in Figure 3, of the Phragmén–Lindelof theorem.

Proof. By Remark 2.2, we may assume $(\gamma_1, \gamma_2) \in D_2^+$. Assume f is continuous at O; then f is bounded in a neighborhood of O. Fix $\theta_1 \in (-\alpha, -\alpha + \pi - \gamma_2)$ and $\theta_2 \in (\alpha - \gamma_1, \alpha)$. By making $\delta_0 > 0$ smaller if necessary, we may assume

$$\Omega^* = \{(r\cos(\theta), r\sin(\theta)) : 0 < r < \delta_0,\ \theta_1 < \theta < \theta_2\}$$

is contained in Ω. Let $\partial^+\Omega^* = \{(r\cos(\theta_2), r\sin(\theta_2)) : 0 \le r \le \delta_0\}$ and define $\gamma_0^+ : \partial^+\Omega^* \to [0, \pi]$ so that $\cos(\gamma_0^+(x, y)) = Tf(x, y) \cdot (\cos(\theta_2 + \frac{1}{2}\pi), \sin(\theta_2 + \frac{1}{2}\pi))$ and notice that Theorem 2.1(iii) implies

$$\gamma_0^+(x, y) \to \gamma_1 + \theta_2 - \alpha \qquad \text{as } (x, y) \in \partial^+\Omega^* \text{ goes to } (0, 0). \tag{32}$$

Let $\partial^-\Omega^* = \{(r\cos(\theta_1), r\sin(\theta_1)) : 0 \le r \le \delta_0\}$ and define $\gamma_0^- : \partial^-\Omega^* \to [0,\pi]$ so that $\cos(\gamma_0^-(x,y)) = Tf(x,y)\cdot(\cos(\theta_1 - \frac{1}{2}\pi), \sin(\theta_1 - \frac{1}{2}\pi))$ and notice that Theorem 2.1(ii) implies

$$\gamma_0^-(x,y) \to \alpha + \gamma_2 + \theta_1 \quad \text{as } (x,y) \in \partial^-\Omega^* \text{ goes to } (0,0). \tag{33}$$

Set

$$\Pi = \{(\cos(\beta + \tfrac{1}{2}\pi), \sin(\beta + \tfrac{1}{2}\pi), 0) : \pi - \alpha - \gamma_2 \le \beta \le \alpha - \gamma_1\}$$

and, for $s \in (0,\delta_0]$, let $\Omega_s = \{(x,y) \in \Omega^* : x^2 + y^2 < s^2\}$; notice that Theorem 2.1 implies

$$\cap_{s>0}\, \overline{\vec{n}(\Omega_s)} = \Pi. \tag{34}$$

Since $\alpha \le \frac{1}{2}\pi$ and $\gamma_2 - \gamma_1 > \pi - 2\alpha$, we have $0 < \frac{3}{2}\pi - \alpha - \gamma_2 < \frac{1}{2}\pi + \alpha - \gamma_1 < \pi$.

We now wish to examine the stereographic projection of the Gauss map near $(0,0,f(0,0))$ and represent it as the sum of a holomorphic function and a continuous function (that is, (44)).

From (32), (33) and (34) and the fact that $\gamma_1, \gamma_2 \in (0,\pi)$, we see that there exists $\sigma \in (0,\delta_0]$ small enough that

$$\vec{n}(\Omega_\sigma) \subset \{\omega(\theta,\phi) : \tfrac{1}{4}(3\pi - 2\alpha - 2\gamma_2) < \theta < \tfrac{1}{4}(3\pi + 2\alpha - 2\gamma_1),\ \tfrac{1}{2}\pi < \phi < \tfrac{3}{4}\pi\}, \tag{35}$$

where $\omega(\theta,\phi) = (\sin(\phi)\cos(\theta), \sin(\phi)\sin(\theta), \cos(\phi))$, and there exists $\lambda > 0$ such that $\lambda < \gamma_0^\pm(x) < \pi - \lambda$ for $x \in \partial\Omega^* \setminus \{O\}$ with $|x| \le \sigma$. Notice that $f \in C^0(\overline{\Omega_\sigma}) \cap C^2(\overline{\Omega_\sigma} \setminus \{O\})$ and that f satisfies $Nf = H(x,f)$ on Ω_σ, $Tf \cdot \boldsymbol{\nu} = \cos(\gamma_0^+)$ on $\partial^+\Omega_\sigma = \Omega_\sigma \cap \partial^+\Omega^*$ and $Tf \cdot \boldsymbol{\nu} = \cos(\gamma_0^-)$ on $\partial^-\Omega_\sigma = \Omega_\sigma \cap \partial^-\Omega^*$. Define

$$S_0 = \{(x,y,f(x,y)) : (x,y) \in \Omega_\sigma\} \quad \text{and} \quad \Gamma_0 = \{(x,y,f(x,y)) : (x,y) \in \partial\Omega_\sigma\}.$$

If $\Gamma_0^\pm = \{(x,y,f(x,y)) : (x,y) \in \partial^\pm\Omega_\sigma,\ x^2 + y^2 < \sigma^2\}$ and $\Gamma_0^\sigma = \Gamma_0 \setminus (\Gamma_0^+ \cup \Gamma_0^-)$, then $\Gamma_0 = \Gamma_0^+ \cup \Gamma_0^- \cup \Gamma_0^\sigma$.

We will use the unit disk $E = \{(u,v) : u^2 + v^2 < 1\}$ as a parameter domain. From step 1 of the proof of [Lancaster and Siegel 1996a, Theorem 1] and from [Kenmotsu 1979] (also [Kenmotsu 2003]), we obtain the following facts.

There is a parametric description of the surface S_0

$$X(u,v) = (x(u,v), y(u,v), z(u,v)) \in C^2(E : \mathbb{R}^3) \cap W^{1,2}(E : \mathbb{R}^3)$$

with the following properties:

(i) X is a homeomorphism of E onto S_0.

(ii) X maps ∂E strictly monotonically onto Γ_0.

(iii) X is conformal on E, that is, $X_u \cdot X_v = 0$ and $|X_u| = |X_v|$ on E.

(iv) Let $\tilde{H}(u,v) = H(X(u,v))$ denote the prescribed mean curvature of S_0 at $X(u,v)$. Then $\triangle X := X_{uu} + X_{vv} = \tilde{H} X_u \times X_v$.

(v) $X \in C^0(\overline{E})$ and $X(1,0) = (0,0,z_0)$, where $z_0 = f(0,0)$.

(vi) Write $G(u,v) = (x(u,v), y(u,v))$. Then $G(\cos t, \sin t)$ moves clockwise about $\partial\Omega_\sigma$ as t increases in $0 \le t \le 2\pi$, and G is an orientation-reversing homeomorphism from $\overline{E}$ onto $\overline{\Omega_\sigma}$.

(vii) [Kenmotsu 1979, Lemma 1 and Corollary] Let $\pi_S : S^2 \to \mathbb{C}$ denote the stereographic projection from the north pole and define $g(u+iv) = \pi_S(\vec{n}(G(u,v)))$ for $(u,v) \in E$. Then

$$|g_{\bar\zeta}| = \tfrac{1}{2}|\tilde{H}|(1+|g|^2)|X_u|, \tag{36}$$

where $\zeta = u + iv$,

$$\frac{\partial}{\partial \zeta} = \frac{1}{2}\Big(\frac{\partial}{\partial u} - i\frac{\partial}{\partial v}\Big) \quad \text{and} \quad \frac{\partial}{\partial \bar\zeta} = \frac{1}{2}\Big(\frac{\partial}{\partial u} + i\frac{\partial}{\partial v}\Big).$$

For convenience with complex variables, set $E_1 = \{\zeta \in \mathbb{C} : |\zeta| < 1\}$.

Now Theorem 2.1 implies

$$g(1+) = \lim_{\theta \downarrow 0} g(e^{i\theta}) = \cos(\alpha - \gamma_1 + \tfrac{1}{2}\pi) + i\sin(\alpha - \gamma_1 + \tfrac{1}{2}\pi)$$

and

$$g(1-) = \lim_{\theta \uparrow 0} g(e^{i\theta}) = \cos(\tfrac{3}{2}\pi - \alpha - \gamma_2) + i\sin(\tfrac{3}{2}\pi - \alpha - \gamma_2).$$

Define $\tilde{n}(u,v) = \vec{n}(x(u,v), y(u,v))$ for $(u,v) \in E$. Notice that if

$$\tilde{n}(u,v) = (\tilde{n}_1(u,v), \tilde{n}_2(u,v), \tilde{n}_3(u,v)),$$

then, from the choice of σ,

$$-\cot\big(\tfrac{1}{4}\pi + \tfrac{1}{2}(\alpha - \gamma_1)\big) < \frac{\tilde{n}_1(u,v)}{\tilde{n}_2(u,v)} < \cot\big(\tfrac{3}{4}\pi - \tfrac{1}{2}(\alpha - \gamma_2)\big) \tag{37}$$

and

$$\min\{-\csc\big(\tfrac{1}{4}(3\pi - 2\alpha - 2\gamma_2)\big), -\csc\big(\tfrac{1}{4}(3\pi + 2\alpha - 2\gamma_1)\big)\} < \frac{\tilde{n}_3(u,v)}{\tilde{n}_2(u,v)} < 0. \tag{38}$$

Now

$$\tilde{n}(u,v) = \frac{X_u \times X_v}{|X_u \times X_v|} = \frac{1}{|X_u|^2}(y_u z_v - y_v z_u, x_v z_u - x_u z_v, x_u y_v - x_v y_u);$$

hence

$$\frac{|x_u y_v - x_v y_u|}{|x_v z_u - x_u z_v|} = \frac{|\tilde{n}_3|}{|\tilde{n}_2|} < A \quad \text{and} \quad \frac{|y_u z_v - y_v z_u|}{|x_v z_u - x_u z_v|} = \frac{|\tilde{n}_1|}{|\tilde{n}_2|} < B, \tag{39}$$

where

$$A = \max\left\{\csc\left(\tfrac{1}{4}(3\pi - 2\alpha - 2\gamma_2)\right), \csc\left(\tfrac{1}{4}(3\pi + 2\alpha - 2\gamma_1)\right)\right\},$$
$$B = \max\left\{\cot\left(\tfrac{1}{4}\pi + \tfrac{1}{2}(\alpha - \gamma_1)\right), \cot\left(\tfrac{3}{4}\pi - \tfrac{1}{2}(\alpha - \gamma_2)\right)\right\}.$$

Now $(\partial f/\partial y)(x(u,v), y(u,v)) = -\tilde{n}_2(u,v)/\tilde{n}_3(u,v)$ and so (38) implies

$$\frac{\partial f}{\partial y} \geq \min\left\{\sin\left(\tfrac{1}{4}(3\pi - 2\alpha - 2\gamma_2)\right), \sin\left(\tfrac{1}{4}(3\pi + 2\alpha - 2\gamma_1)\right)\right\} > 0 \tag{40}$$

and so $S_0 = \{(x, y, f(x, y)) : (x, y) \in \Omega_\sigma\} = X(E)$ is the graph $y = \phi(z, x)$ over the (z, x)-plane of a C^2 function over the projection U of $\mathcal{S}_0$ on the (z, x)-plane. Notice that $\phi \in C^0(\overline{U})$ and ∂U is the projection of Γ_0 on the (z, x)-plane.

If $\partial_0 U = \{(z, x) : (x, y, z) \in \Gamma_0^+ \cup \Gamma_0^-\}$ and $\partial_1 U = \{(z, x) : (x, y, z) \in \Gamma_0^\sigma\}$, then $\partial U = \partial_0 U \cup \partial_1 U$. Now Theorem 2.1 implies $|\nabla f(x, y)| \to \infty$ as $(x, y) \in \Omega^*$ goes to O. Also

$$Tf(r\cos(\theta_2), r\sin(\theta_2)) \cdot (\cos(\theta_2), \sin(\theta_2)) \to \cos(\alpha + \tfrac{1}{2}\pi - \gamma_1 - \theta_2) > 0, \tag{41}$$

since $\frac{1}{2}\pi - \gamma_1 < \alpha + \frac{1}{2}\pi - \gamma_1 - \theta_2 < \frac{1}{2}\pi$, and

$$Tf(r\cos(\theta_1), r\sin(\theta_1)) \cdot (\cos(\theta_1), \sin(\theta_1)) \to \cos(\tfrac{3}{2}\pi - \alpha - \gamma_2 - \theta_1) < 0,$$

since $\frac{1}{2}\pi < \frac{3}{2}\pi - \alpha - \gamma_2 - \theta_1 < \frac{3}{2}\pi - \gamma_2$. Thus the limits of the directional derivatives of f in the directions of $\partial^+\Omega_\sigma$ and $\partial^-\Omega_\sigma$ are

$$\lim_{r\downarrow 0} \nabla f(r\cos(\theta_2), r\sin(\theta_2)) \cdot (\cos(\theta_2), \sin(\theta_2)) = +\infty, \tag{42}$$
$$\lim_{r\downarrow 0} \nabla f(r\cos(\theta_1), r\sin(\theta_1)) \cdot (\cos(\theta_1), \sin(\theta_1)) = -\infty. \tag{43}$$

Hence Γ_0^+ is tangent to $\{(0, 0, z) : z \geq z_0\}$ and Γ_0^- is tangent to $\{(0, 0, z) : z \leq z_0\}$ at $(0, 0, z_0)$. In addition, $(\partial\phi/\partial z)(z_0, 0) = 0$. Thus $\Gamma_0^+ \cup \Gamma_0^-$ is a C^1 curve and ∂U is the union of the C^1 curve $\partial_0 U$ and the C^2 curve $\partial_1 U$. Since

$$|\nabla f(\sigma\cos(\theta_2), \sigma\sin(\theta_2))| < \infty,$$

$f_y(\sigma\cos(\theta_2), \sigma\sin(\theta_2)) > 0$ (by (40)) and the curves $y = \tan(\theta_2)x$ and $x^2 + y^2 = \sigma^2$ are orthogonal at $(\sigma\cos(\theta_2), \sigma\sin(\theta_2))$, we see that Γ_0^+ and Γ_0^σ do not meet tangentially at $(\sigma\cos(\theta_2), \sigma\sin(\theta_2), f(\sigma\cos(\theta_2), \sigma\sin(\theta_2)))$ and $\partial_0^+ U$ and $\partial_1 U$ do not meet tangentially at $(\sigma\cos(\theta_2), \sigma\sin(\theta_2))$. Similarly Γ_0^- and Γ_0^σ do not meet tangentially at $(\sigma\cos(\theta_1), \sigma\sin(\theta_1), f(\sigma\cos(\theta_1), \sigma\sin(\theta_1)))$ and $\partial_0^- U$ and $\partial_1 U$ do not meet tangentially at $(\sigma\cos(\theta_1), \sigma\sin(\theta_1))$. Therefore U is a simply connected Lipschitz domain and ∂U is a quasicircle (see [Gehring 2005, Theorem 6.3]).

Let us define $F \in C^2(E : \mathbb{R}^2) \cap W^{1,2}(E : \mathbb{R}^2)$ by $F(u, v) = (z(u, v), x(u, v))$. Note that F is a homeomorphism from $\overline{E}$ onto $\overline{U}$. Recall that $|DF|^2 = x_u^2 + x_v^2 + z_u^2 + z_v^2$

and the determinant of DF at (u, v) is

$$J((u,v),F) = x_v z_u - x_u z_v = |X_u|^2 \tilde{n}_2 > 0.$$

Since we are using conformal, or isothermal, coordinates, we obtain

$$\begin{aligned}|DF(u,v)|^2 \le 2|X_u|^2 &= 2|X_u \times X_v| \\ &= 2\sqrt{(y_u z_v - y_v z_u)^2 + (x_v z_u - x_u z_v)^2 + (x_u y_v - x_v y_u)^2} \\ &\le 2\sqrt{(B^2+1+A^2)(x_v z_u - x_u z_v)^2} = 2K\,J((u,v),F),\end{aligned}$$

where $K = \sqrt{B^2+1+A^2}$. Thus F is a K'-quasiconformal map from E to U, where $K' = (K - \sqrt{K^2-1})^{-1} = \sqrt{A^2+B^2+1} + \sqrt{A^2+B^2}$; see for example [Finn and Serrin 1958]. Then [Gehring 2005, Theorem 6.4] implies that there is a K'-quasiconformal extension $L : \mathbb{R}^2 \to \mathbb{R}^2$ of $F^{-1} : \bar{U} \to \bar{E}$ and hence there is a K'-quasiconformal extension $\tilde{F} : \mathbb{R}^2 \to \mathbb{R}^2$ of F, given by $\tilde{F} = L^{-1}$. Using Gehring's lemma [1973] or [Iwaniec and Martin 2001, Theorem 14.4.1], we see that $\tilde{F} \in W^{1,p}(B((1,0),\delta))$ for some $p > 2$. Since $\tilde{F} = F$ on $E \cap B((1,0),2\delta)$ and $F \in W^{1,\infty}(E \setminus B((1,0),\epsilon))$ for each $\epsilon > 0$, we see that $x_u, x_v, z_u, z_v \in L^p(E)$. Since $\tilde{n}$ is normal to $X(E)$, we have $X_u \cdot \tilde{n} = 0$ and $X_v \cdot \tilde{n} = 0$, which imply

$$y_u = \frac{\tilde{n}_1}{\tilde{n}_2} x_u + \frac{\tilde{n}_3}{\tilde{n}_2} z_u \quad \text{and} \quad y_v = \frac{\tilde{n}_1}{\tilde{n}_2} x_v + \frac{\tilde{n}_3}{\tilde{n}_2} z_v$$

and therefore $|y_u| \le B|x_u| + A|z_u|$ and $|y_v| \le B|x_v| + A|z_v|$. This implies X belongs to $W^{1,p}(E : \mathbb{R}^3)$.

The corollary on [Kenmotsu 1979, page 92] yields

$$|g_{\bar{\zeta}}| = \tfrac{1}{2}|\tilde{H}|(1+|g|^2)|X_u| \le |H|_\infty |X_u|$$

and so $g_{\bar{\zeta}} \in L^p(E_1 : \mathbb{R}^2)$. Let us set $\mu = (p-2)/p$. Then from [Monakhov 1983, Theorems 5 and 6, page 205], we see that

$$g(\zeta) = \psi(\zeta) + h(\zeta), \tag{44}$$

where ψ is a holomorphic function and $h \in L^\infty(E_1)$ is a uniformly Hölder continuous function on E_1 with Hölder exponent μ. Since g and h are bounded, so is ψ. Since h is continuous at $1 \in \partial E_1$ and $\psi(\zeta) = g(\zeta) - h(\zeta)$, the Phragmén–Lindelof theorem (for example, [Bear and Hile 1983]) and Theorem 2.1 imply

$$\lim_{r\to 0^+} \psi(1 + r\cos(\theta) + ir\sin(\theta)) = \psi(1-)(\theta/\pi - 1/2) + \psi(1+)(3/2 - \theta/\pi)$$

for $\pi/2 < \theta < 3\pi/2$, where $\psi(1+) = g(1+) - h(1)$ and $\psi(1-) = g(1-) - h(1)$, and so

$$\lim_{r\to 0^+} g(1 + r\cos(\theta) + ir\sin(\theta)) = g(1-)(\theta/\pi - 1/2) + g(1+)(3/2 - \theta/\pi) \tag{45}$$

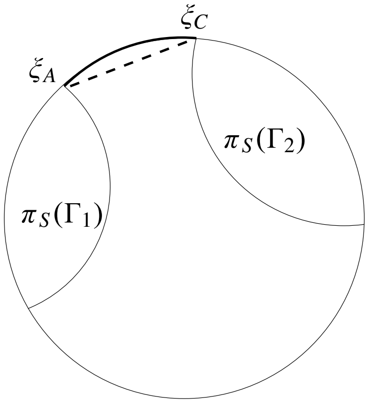

Figure 3. Differing limits of $g(u+iv)$ as $u+iv \to 1$.

for $\pi/2 < \theta < 3\pi/2$. Notice then that along any ray $\{1+r\cos(\theta)+ir\sin(\theta)\}$ with $\pi/2 < \theta < 3\pi/2$, in $|\zeta| < 1$, g converges to a point strictly inside the open unit disk as 1 is approached; see Figure 3. This contradicts Theorem 2.1, which implies $|g(u+iv)| \to 1$ as $u+iv \to 1$. Thus our assumption that f is continuous at O is invalid and the proof of Theorem 1.1 is complete. □

Remark 3.1. In [Lancaster and Siegel 1996a, Theorem 1], the hypotheses include, "*If* $\alpha < \pi/2$ *and there exist constants* $\underline{\gamma}^{\pm}, \overline{\gamma}^{\pm}, 0 < \underline{\gamma}^{\pm} \le \pi/2, \pi/2 \le \overline{\gamma}^{\pm} < \pi$, *satisfying*

$$\underline{\gamma}^{+} + \underline{\gamma}^{-} > \pi - 2\alpha \quad \textit{and} \quad \overline{\gamma}^{+} + \overline{\gamma}^{-} < 2\alpha + \pi$$

so that $\underline{\gamma}^{\pm} \le \gamma^{\pm}(s) \le \overline{\gamma}^{\pm}$ *for all* s *in* $0 < s < s_0$ *for some* s_0."

While the theorem is true as stated, the assumptions $\underline{\gamma}^{\pm} \le \pi/2$ and $\pi/2 \le \overline{\gamma}^{\pm}$ were added as an afterthought (by this author) and were unnecessary to the argument; [Lancaster and Siegel 1996a, Theorems 1 and 2] remain correct if one merely assumes $\underline{\gamma}^{\pm} \le \overline{\gamma}^{\pm}$. It is useful to note this fact because these assumptions artificially restrict the applicability of these theorems. (In fact, the remainder of that article correctly ignores this restriction.)

Appendix

We wish to discuss variational solutions of (1) and (2). We assume a solution $f \in C^2(\Omega) \cap C^{1,\rho^*}(\overline{\Omega} \setminus \{O\})$ is given and we define $H^* : \Omega \times \mathbb{R} \to \mathbb{R}$ by

$$H^*(x) = H(x, f(x)) \quad \text{for } x \in \Omega. \tag{46}$$

For the moment, we let Ω be any connected, open subset of $\mathbb{R}^2$ that has locally Lipschitz boundary and let $\gamma \in L^\infty(\partial\Omega)$ with $0 \le \gamma(x, y) \le \pi$ for $(x, y) \in \partial\Omega$; for convenience of notation, we assume Ω is bounded. The usual definition of a

$BV(\Omega)$ (variational) solution of (1) and (2) is a function $u \in BV(\Omega)$ that minimizes the functional

$$\mathscr{E}(u) = \int_\Omega \sqrt{1+ \mid Du \mid^2}\, dH_2 + \int_\Omega \int_0^u H^*(\,\cdot\,, t)\, dt\, dH_2 - \int_{\partial\Omega} \cos(\gamma) u\, dH_1$$

over $BV(\Omega)$. In some cases (for example, Ω is unbounded), individual terms in the functional may be infinite; Finn [1986, Definition 7.1] offers a more general definition of variational solution in his book. Another type of variational solution of (1) and (2) is that of a generalized solution, which we describe next.

We denote by $\mathscr{F}$ the (formal) functional given by

$$\mathscr{F}(U) = \int_{\Omega\times\mathbb{R}} |D\phi_U| + \int_{\Omega\times\mathbb{R}} H^*\phi_U\; dt\, dH_2 - \int_{\partial\Omega\times\mathbb{R}} \cos(\gamma)\phi_U\, dH_2.$$

For each $T \in (0, \infty)$ and $K \subset\subset \mathbb{R}^2$, we define the functional

$$\text{(47)}\quad \mathscr{F}_{T,K}(U) = \int\limits_{\Omega(T,K)} |D\phi_U| + \int\limits_{\Omega(T,K)} H^*\phi_U\, dt\, dH_2 - \int\limits_{\delta\Omega(T,K)} \cos(\gamma)\phi_U\, dH_2$$

when $U \subset \Omega \times \mathbb{R}$ is a Caccioppoli set (that is, a Borel set with locally finite perimeter), where $\Omega(T, K) = (\Omega \cap K) \times (-T, T)$ and $\delta\Omega(T, K) = (\partial\Omega \cap K) \times (-T, T)$.

Definition 3.1. A Caccioppoli set $U \subset \Omega \times \mathbb{R}$ is said to be a *local solution* for $\mathscr{F}$ if and only if for each $T > 0$ and $K \subset\subset \mathbb{R}^2$, we have $\mathscr{F}_{T,K}(U) \le \mathscr{F}_{T,K}(V)$ whenever $V \subset \Omega \times \mathbb{R}$ is a Caccioppoli set such that the support of $\phi_U - \phi_V$ is contained in $\Omega(T, K)$.

As noted in [Finn 1986, Section 7.3], a function $u \in BV(\Omega)$ minimizes $\mathscr{E}$ if and only if its subgraph $U = \{(x, y, t) \in \Omega \times \mathbb{R} : t < u(x, y)\}$ is a local solution for $\mathscr{F}$ [Miranda 1977].

Definition 3.2. A function $u : \Omega \to [-\infty, \infty]$ is called a *generalized solution* of (1) and (2) if and only if its subgraph U is a local solution for $\mathscr{F}$.

Definition 3.3. A sequence (u_j) in $BV(\Omega)$ is said to *converge locally in* Ω *to* u_∞ if and only if ϕ_{U_j} converges to ϕ_{U_∞} in $L^1_{\text{loc}}(\Omega \times \mathbb{R})$ as $j \to \infty$, where U_j and U_∞ are the subgraphs of u_j and u_∞, respectively.

Definition 3.4. For each $\Lambda \subset \mathbb{R}^2$ and $\epsilon > 0$, set $\Lambda^\epsilon = \Lambda \setminus B(O, \epsilon)$ and $\Sigma^\epsilon = \partial\Lambda \setminus B(O, \epsilon)$. We will say the triple (λ, Λ, O) is *admissible* if and only if Λ is an open set in $\mathbb{R}^2$, $O \in \partial\Lambda$, the map $\lambda : \partial\Lambda \setminus \{O\} \to (0, \pi)$ is in $C^{0,\rho^*}(\partial\Lambda \setminus \{O\})$ and, for some $\epsilon_0 > 0$ and all $\epsilon \in (0, \epsilon_0]$, there exist $a = a(\epsilon) \in (0, 1)$, $\tau = \tau(\epsilon) > 0$, $N = N(\epsilon)$, $N_1 = N_1(\epsilon) \le N(\epsilon)$, a finite open cover $\{\Lambda^\epsilon_j : j = 2, \dots, N\}$ of $\overline{\Omega^\epsilon}$ with $O \notin \bigcup_{j=2}^N \overline{\Lambda^\epsilon_j}$ and rigid motions $F_j : \mathbb{R}^2 \to \mathbb{R}^2$ for $2 \le j \le N_1$, such that $\Lambda^\epsilon_j \cap \partial\Lambda \ne \varnothing$ if $1 \le j \le N_1$ and $\Lambda^\epsilon_j \cap \partial\Lambda = \varnothing$ if $N_1 < j \le N$, the set $\Sigma^\epsilon_j = \partial\Lambda \cap \Lambda^\epsilon_j$ is open and connected in the relative topology of Σ^ϵ, $F_j(\Sigma^\epsilon_j)$ can be represented over some interval $a_j < x < b_j$ with $a_j < b_j$ by a Lipschitz function $y = \psi_j(x)$ with Lipschitz constant L_j, the set $T_j = \{(x, y+\psi_j(x)) : a_j < x < b_j,\ -\tau < y < 0\}$ lies in $F_j(\Omega)$ and $|\cos(\gamma)|\sqrt{1+L_j^2} \le a(\epsilon)$ on Σ_j^ϵ for $j = 2, \dots, N_1$. Compare this with [Finn 1986, Section 6.3].

Lemma 3.1. *Suppose $\gamma \in C^{0,\rho^*}(\partial\Omega \setminus \{O\})$ satisfies* (4), $|\gamma_1 - \gamma_2| > \pi - 2\alpha$ *(so that $(\gamma_1, \gamma_2) \in D_2^+ \cup D_2^-$) and (γ, Ω, O) is admissible. Then there exist $\zeta > 0$, $\mu = \mu(a(\zeta), \Omega)$ and $\Upsilon = \Upsilon(a(\zeta), \Omega)$ with $\mu \in [a(\zeta), 1)$ such that for each $T > 0$, $\lambda > 0$ and $f \in BV(\Omega \times (-T, T))$ with $f \ge 0$ almost everywhere on $\Omega \times (-T, T)$, we have*

$$\Bigl|\int_{\Sigma\times(-T,T)} \cos(\gamma) f^* dH_2\Bigr| \le \mu \int_{\mathcal{A}_\lambda\times(-T,T)} |Df| + \Upsilon \int_{\mathcal{A}_\lambda\times(-T,T)} f, \tag{48}$$

where $\mathcal{A}_\lambda \subset \Omega$ is the strip of width λ adjacent to $\Sigma = \partial\Omega$ and we denote by $f^* \in L^1(\partial\Omega \times (-T, T))$ the trace of f on $\partial\Omega \times (-T, T)$.

Proof. Fix $T > 0$ and $\lambda > 0$. Let $f \in BV(\Omega \times (-T, T))$ such that $f \ge 0$ almost everywhere in $\Omega \times (-T, T)$; then $f^* \ge 0$ almost everywhere on $\partial\Omega$. We see from [Giusti 1984, Remark 2.12] that there exists a sequence

$$\{f_k\} \subset C^\infty(\Omega \times (-T, T)) \cap BV(\Omega \times (-T, T))$$

such that

$$\lim_{k\to\infty} \int_{\Omega\times(-T,T)} |f_k - f|\, dx = 0, \tag{49}$$

$$\lim_{k\to\infty} \int_{\Omega\times(-T,T)} |Df_k|\, dx = \int_{\Omega\times(-T,T)} |Df| \tag{50}$$

and

$$f_k^* = f^* \quad \text{on } \partial(\Omega \times (-T, T)) \text{ for each } k \in \mathbb{N}, \tag{51}$$

where f_k^* and f^* denote the traces of f_k and f on $\partial(\Omega \times (-T, T))$, respectively. An examination of the construction of the f_k in [Giusti 1984, Theorem 1.17] shows that $f_k \ge 0$ on $\Omega \times (-T, T)$ for $k = 1, 2, 3, \dots$, since $f \ge 0$ almost everywhere on $\Omega \times (-T, T)$. (In fact, each f_k is actually a function f_ϵ for a suitably small $\epsilon > 0$ in the construction in the proof of that theorem.)

Since $\int |Df|$ is a Radon measure on $\Omega \times (-T, T)$,

$$\int_{\partial\mathcal{A}_\sigma\times(-T,T)} |Df| = 0 \quad \text{for almost all } \sigma \in (0, \lambda] \text{ and all } T > 0; \tag{52}$$

by replacing λ by a $\sigma \in (0, \lambda]$ that satisfies (52), we may assume

$$\int_{\partial\mathcal{A}_\lambda\times(-T,T)} |Df| = 0 \tag{53}$$

always holds.

We shall focus on functions $h \in C^1(\Omega\times(-T,T))\cap BV(\Omega\times(-T,T))$ with $h\geq 0$ in $\Omega\times(-T,T))$, obtain (48) for such functions, and then use the approximation above to establish (48) for f.

Case 1 ($(\gamma_1,\gamma_2)\in D_2^+$ and $\gamma_2\leq\pi/2$). This case is defined by $\gamma_2-\gamma_1>\pi-2\alpha$, and so $\gamma_1<\pi/2$ and $2\alpha>\pi/2$. Fix $\epsilon\in(0,\gamma_1)$. We wish to select $\sigma\in(0,\pi/2)$ such that $\sigma<\gamma_1-\epsilon$, $0<\pi-2\alpha-\sigma<\gamma_2-\epsilon$. Now these conditions require that $\sigma\in(0,\gamma_1-\epsilon)\cap(\pi-2\alpha-\gamma_2+\epsilon,\pi-2\alpha)$; this intersection is nonempty since $\gamma_1-\epsilon-(\pi-2\alpha-\gamma_2+\epsilon)>2\gamma_1-2\epsilon>0$ and so $\gamma_1-\epsilon>\pi-2\alpha-\gamma_2+\epsilon$.

Let $\zeta>0$ be small enough that

(a) $|\gamma(x)-\gamma_1|<\epsilon/2$ whenever $x\in\partial^+\Omega\setminus\{O\}$ with $|x|\leq 2\zeta$, and

(b) $|\gamma(x)-\gamma_2|<\epsilon/2$ whenever $x\in\partial^-\Omega\setminus\{O\}$ with $|x|\leq 2\zeta$.

Recall that $\tau^+(x)=\gamma(x)-\pi/2$ for $x\in\partial^+\Omega\cap B_{\delta^*}(O)$ and $\tau^-(x)=\gamma(x)+\pi/2$ for $x\in\partial^-\Omega\cap B_{\delta^*}(O)$; hence $|\tau^+(x)-\alpha|<\epsilon/2$ whenever $x\in\partial^+\Omega$ with $|x|\leq 2\zeta$ and $|\tau^-(x)+\alpha|<\epsilon/2$ whenever $x\in\partial^-\Omega$ with $|x|\leq 2\zeta$.

Let $\tau=\zeta$ and $R_1:\mathbb{R}^2\to\mathbb{R}^2$ be the rotation about the origin through the angle $-\alpha-\sigma$. Then $R_1(\partial^+\Omega)$ and $R_1(\partial^-\Omega)$ are the graphs $y=\psi_1^+(x)$ and $y=\psi_1^-(x)$ of Lipschitz functions with Lipschitz constants

$$L_1^+\leq\tan(\sigma+\epsilon/2)\quad\text{and}\quad L_1^-\leq\tan(\pi-2\alpha-\sigma+\epsilon/2),$$

respectively; notice that $\operatorname{dom}(\psi_1^+)=[0,x_0^+)$ and $\operatorname{dom}(\psi_1^-)=(x_0^-,0]$, where

$$|(x_0^+,\psi_1^+(x_0^+))|=2\zeta\quad\text{and}\quad|(x_0^-,\psi_1^-(x_0^-))|=2\zeta.$$

Set $L_1=\max\{L_1^+,L_1^-\}$ and let $\delta>0$ satisfy $\delta^2+(L_1\delta+\tau)^2=4\zeta^2$ (so that $\delta=\zeta((3L_1^2+4)^{1/2}-1)/(L_1^2+1)$).

For $0<x\leq\delta$, we have $\sigma+\epsilon/2<\gamma_1-\epsilon/2<\gamma(x)$ and so

$$\cos(\gamma(x))\sqrt{1+(L_1^+)^2}<\cos(\gamma_1-\epsilon/2)\sec(\sigma+\epsilon/2)<\frac{\cos(\gamma_1-\epsilon/2)}{\cos(\gamma_1-\epsilon/2)}=1.$$

For $-\delta\leq x<0$, we have $\pi-2\alpha-\sigma+\epsilon/2<\gamma_2-\epsilon/2<\gamma(x)$ and so

$$\cos(\gamma(x))\sqrt{1+(L_1^-)^2}<\cos(\gamma_2-\epsilon/2)\sec(\pi-2\alpha-\sigma+\epsilon/2)<\frac{\cos(\gamma_2-\epsilon/2)}{\cos(\gamma_2-\epsilon/2)}=1.$$

Set $S_1=(-\delta,\delta)\times(-L_1\delta-\tau,0)$,

$$\mu_1=\frac{\cos(\gamma_1-\epsilon/2)}{\cos(\sigma+\epsilon/2)}\quad\text{and}\quad\mu_2=\frac{\cos(\gamma_2-\epsilon/2)}{\cos(\pi-2\alpha-\sigma+\epsilon/2)}.$$

Then $\mu_1<1$, $\mu_2<1$ and

$$\sqrt{1+(L_1^+)^2}\cos(\gamma\circ R_1^{-1}(x))\leq\mu_1\quad\text{for } x\in R_1(\partial^+\Omega)\cap S_1,\tag{54}$$

and

$$(55)\qquad \sqrt{1+(L_1^-)^2}\cos(\gamma\circ R_1^{-1}(x)) \le \mu_2 \quad \text{for } x\in R_1(\partial^-\Omega)\cap S_1,$$

We will now establish

$$(56)\qquad \int_{\Sigma\times(-T,T)}\cos(\gamma)h^* dH_2 \le \mu\int_{\mathscr{A}_\lambda\times(-T,T)}|Dh| + \Upsilon\int_{\mathscr{A}_\lambda\times(-T,T)} h,$$

when $h\in C^1(\Omega\times(-T,T))\cap BV(\Omega\times(-T,T))$ with $h\ge 0$ in $\Omega\times(-T,T))$. To a great extent, we will follow the proof of [Finn 1986, Lemma 6.1]. In Definition 3.4, set ϵ equal to δ, $N=N(\delta)$, $N_1=N_1(\delta)$, $\tau=\tau(\delta)$ and obtain a finite, open cover $\{\Lambda_j^\delta : j=2,\dots,N\}$ of Ω^δ in $\mathbb{R}^2$ with the properties described in the definition. Set $\Omega_j^\delta=\Lambda_j^\delta\cap\overline{\Omega}$ for $j=2,\dots,N$ and set $\Omega_1^\delta=R_1^{-1}(S_1)\cap\overline{\Omega}$. Notice that $\{\Omega_j^\delta : j=1,\dots,N\}$ is an open (in the relative topology of $\overline{\Omega}$) cover of $\overline{\Omega}$. Let $\{\varphi_j : j=1,\dots,N\}$ be a partition of unity of $\overline{\Omega}$ subordinate to $\{\Omega_j^\delta : j=1,\dots,N\}$. Notice since $O\notin\bigcup_{j=2}^N\overline{\Lambda_j^\delta}$ that $\varphi_1\equiv 1$ in some neighborhood of O.

Using the techniques in the proof of [Finn 1986, Lemma 6.1], one sees that

$$\Big|\int_{\Sigma\times(-T,T)}\varphi_j\cos(\gamma)h^* dH_2\Big| \le a(\delta)\int_{\mathscr{A}_\lambda\times(-T,T)}\varphi_j|Dh| + \Upsilon\int_{\mathscr{A}_\lambda^j\times(-T,T)} h,$$

where $\mathscr{A}_\lambda^j=\mathscr{A}_\lambda\cap F_j^{-1}(T_j)$ for each $j=2,\dots,N_1$ and $k\in\mathbb{N}$. Notice also that these techniques yield

$$\Big|\int_{\partial^+\Omega\times(-T,T)}\varphi_1 h^* dH_2\Big| \le \sqrt{1+(L_1^+)^2}\int_{\mathscr{A}_\lambda^+\times(-T,T)}\varphi_1|Dh| + \Upsilon\int_{\mathscr{A}_\lambda^+\times(-T,T)} h$$

and

$$\Big|\int_{\partial^-\Omega\times(-T,T)}\varphi_1 h^* dH_2\Big| \le \sqrt{1+(L_1^-)^2}\int_{\mathscr{A}_\lambda^-\times(-T,T)}\varphi_1|Dh| + \Upsilon\int_{\mathscr{A}_\lambda^-\times(-T,T)} h,$$

where

$$\mathscr{A}_\lambda^+ = \Omega\cap R_1^{-1}(\{(x,y): 0<x<\delta,\ \psi_1^+(x)-\lambda<y<\psi_1^+(x)\}),$$
$$\mathscr{A}_\lambda^- = \Omega\cap R_1^{-1}(\{(x,y): -\delta<x<0,\ \psi_1^-(x)-\lambda<y<\psi_1^-(x)\}).$$

Then

$$\int_{\Sigma\times(-T,T)}\varphi_1\cos(\gamma)h^* dH_2 \le \mu_1\int_{\mathscr{A}_\lambda^+\times(-T,T)}\varphi_1|Dh| + \Upsilon\int_{\mathscr{A}_\lambda^+\times(-T,T)} h$$
$$+\mu_2\int_{\mathscr{A}_\lambda^-\times(-T,T)}\varphi_1|Dh| + \Upsilon\int_{\mathscr{A}_\lambda^-\times(-T,T)} h,$$

and so, if we set $\mu_0 = \max\{\mu_1, \mu_2\} < 1$,

$$\int_{\Sigma\times(-T,T)} \varphi_1 \cos(\gamma)h^* dH_2 \leq \mu_0 \int_{\mathcal{A}_\lambda\times(-T,T)} \varphi_1|Dh| + \Upsilon \int_{\mathcal{A}^1_\lambda\times(-T,T)} h, \tag{57}$$

where $\mathcal{A}^1_\lambda = \mathcal{A}_\lambda \cap B_{2\zeta}(O)$. Therefore, if we set $\mu = \max\{a(\delta), \mu_0\} < 1$, we obtain

$$\begin{aligned}\int_{\Sigma\times(-T,T)} \cos(\gamma)h^* dH_2 &= \sum_{j=1}^{N_1} \int_{\Sigma\times(-T,T)} \varphi_j \cos(\gamma)h^* dH_2 \\ &\leq \sum_{j=1}^{N_1}\biggl(\mu \int_{\mathcal{A}_\lambda\times(-T,T)} \varphi_j|Dh| + \Upsilon \int_{\mathcal{A}^j_\lambda\times(-T,T)} h\biggr) \\ &\leq \mu \int_{\mathcal{A}_\lambda\times(-T,T)} |Dh| + \Upsilon_1 \int_{\mathcal{A}_\lambda\times(-T,T)} h\end{aligned}$$

and thus we obtain (56).

Now set $h = f_k$ and obtain $\int_\Sigma \cos(\gamma) f_k^* \, ds \leq \mu \int_{\mathcal{A}_\delta} |Df_k| + \Upsilon_1 \int_{\mathcal{A}_\delta} |f_k|$ for each $k \in \mathbb{N}$. From (50), (53) and [Giusti 1984, Proposition 1.13], we see that

$$\lim_{k\to\infty} \int_{\mathcal{A}_\lambda\times(-T,T)} |Df_k| \, dx = \int_{\mathcal{A}_\lambda\times(-T,T)} |Df|$$

and therefore using this together with (49) and (51) yields

$$\begin{aligned}\int_\Sigma \cos(\gamma) f^* \, dH^2 = \int_\Sigma \cos(\gamma) f_k^* \, dH^2 &\leq \mu \int_{\mathcal{A}_\delta} |Df_k| + \Upsilon_1 \int_{\mathcal{A}_\delta} |f_k| \\ &\leq \mu \int_{\mathcal{A}_\delta} |Df_k| + \Upsilon_1 \int_{\mathcal{A}_\delta} |f| + \Upsilon_1 \int_{\mathcal{A}_\delta} |f - f_k|.\end{aligned}$$

If we take the limit as $k \to \infty$, we obtain

$$\int_\Sigma \cos(\gamma) f^* \, dH^2 \leq \mu \int_{\mathcal{A}_\delta} |Df| + \Upsilon_1 \int_{\mathcal{A}_\delta} |f|. \tag{58}$$

We wish to prove (58) with f replaced by $-f$. Fix $\epsilon \in (0, \min\{\gamma_1, \pi/2 - \gamma_1\})$. Let $\zeta \in (0, \delta^*/2)$ be small enough that

(a) $|\gamma(x) - \gamma_1| < \epsilon/2$ whenever $x \in \partial^+\Omega \setminus \{O\}$ with $|x| \leq 2\zeta$, and

(b) $|\gamma(x) - \gamma_2| < \epsilon/2$ whenever $x \in \partial^-\Omega \setminus \{O\}$ with $|x| \leq 2\zeta$.

Notice that if $x \in \partial^+\Omega \setminus \{O\} \cap B_{2\zeta}(O)$, then $\pi - \gamma(x) > \pi - \gamma_1 - \epsilon > \pi/2$ and so

$$\cos(\pi - \gamma) < 0 \quad \text{on } (\partial^+\Omega \setminus \{O\}) \cap B_{2\zeta}(O). \tag{59}$$

Also, if $x \in \partial^-\Omega \setminus \{O\} \cap B_{2\zeta}(O)$, then

$$|\tau^-(x) + \alpha| < \epsilon/2 \quad \text{and} \quad \pi - \gamma(x) > \pi - \gamma_2 - \epsilon/2 > \pi/2 - \epsilon/2.$$

Let $\tau=\zeta$ and let $R_2:\mathbb{R}^2\to\mathbb{R}^2$ be the rotation about the origin through the angle $\alpha-\pi$. Then $R_2(\partial^-\Omega)\cap B_{2\zeta}(O)$ is the graph $y=\psi_2^-(x)$ of a Lipschitz function with Lipschitz constant $L_2^-\le\tan(\epsilon/2)$; notice that $\operatorname{dom}(\psi_1^-)=(x_0^-,0]$, where $|(x_0^-,\psi_1^-(x_0^-))|=2\zeta$. Set $L_2=L_2^-$ and let $\delta>0$ satisfy $\delta^2+(L_2\delta+\tau)^2=4\zeta^2$ (so that $\delta=\zeta((3L_2^2+4)^{1/2}-1)/(L_2^2+1)$). For $-\delta\le x<0$, we have $\pi-\gamma(x)\ge\pi/2-\epsilon/2$ and so

$$\cos(\pi-\gamma(x))\sqrt{1+(L_2^-)^2}<\cos(\pi/2-\epsilon/2)\sec(\epsilon/2)<\frac{\cos(\pi/4+\gamma_1/2)}{\cos(\pi/4-\gamma_1/2)}<1.$$

We will now establish

$$-\int_{\Sigma\times(-T,T)}\cos(\gamma)h^*\,dH_2\le\mu\int_{\mathcal{A}_\lambda\times(-T,T)}|Dh|+\Upsilon_1\int_{\mathcal{A}_\lambda\times(-T,T)}h \tag{60}$$

when $h\in C^1(\Omega\times(-T,T))\cap BV(\Omega\times(-T,T))$ with $h\ge0$ in $\Omega\times(-T,T))$, where $\mu=\max\{a(\delta),\mu_3\}<1$ and $\mu_3=\cos(\pi/2-\epsilon/2)/\cos(\epsilon/2)$. Let us write (60) as

$$\int_{\Sigma\times(-T,T)}\cos(\pi-\gamma)h^*dH_2\le\mu\int_{\mathcal{A}_\lambda\times(-T,T)}|Dh|+\Upsilon_1\int_{\mathcal{A}_\lambda\times(-T,T)}h.$$

Using the techniques in the proof of [Finn 1986, Lemma 6.1], one sees that

$$\left|\int_{\Sigma\times(-T,T)}\varphi_j\cos(\pi-\gamma)h^*dH_2\right|\le a(\delta)\int_{\mathcal{A}_\lambda\times(-T,T)}\varphi_j|Dh|+\Upsilon\int_{\mathcal{A}_\lambda^j\times(-T,T)}h,$$

where $\mathcal{A}_\lambda^j=\mathcal{A}_\lambda\cap F_j^{-1}(T_j)$ for each $j=2,\dots,N_1$ and $k\in\mathbb{N}$. Notice also that these techniques yield

$$\left|\int_{\partial^-\Omega\times(-T,T)}\varphi_1h^*dH_2\right|\le\sqrt{1+(L_1^-)^2}\int_{\mathcal{A}_\lambda^-\times(-T,T)}\varphi_1|Dh|+\Upsilon\int_{\mathcal{A}_\lambda^-\times(-T,T)}h,$$

where $\mathcal{A}_\lambda^-=\Omega\cap R_2^{-1}\big(\{(x,y):-\delta<x<0,\ \psi_2^-(x)-\lambda<y<\psi_2^-(x)\}\big)$. Then

$$\int_{\Sigma\times(-T,T)}\varphi_1\cos(\pi-\gamma)h^*dH_2\le\mu_3\int_{\mathcal{A}_\lambda^-\times(-T,T)}\varphi_1|Dh|+\Upsilon\int_{\mathcal{A}_\lambda^-\times(-T,T)}h. \tag{61}$$

Therefore, if we set $\mu=\max\{a(\delta),\mu_0\}<1$, we obtain

$$\begin{aligned}\int_{\Sigma\times(-T,T)}\cos(\pi-\gamma)h^*dH_2&=\sum_{j=1}^{N_1}\int_{\Sigma\times(-T,T)}\varphi_j\cos(\gamma)h^*dH_2\\&\le\sum_{j=1}^{N_1}\Big[\ \mu\int_{\mathcal{A}_\lambda\times(-T,T)}\varphi_j|Dh|+\Upsilon\int_{\mathcal{A}_\lambda^j\times(-T,T)}h\ \Big]\\&\le\mu\int_{\mathcal{A}_\lambda\times(-T,T)}|Dh|+\Upsilon_1\int_{\mathcal{A}_\lambda\times(-T,T)}h,\end{aligned}$$

and thus we obtain (60). If we reason as before when we used (56) to obtain (58), we see that (60) and the approximation of f by the (f_k) implies

$$-\int_{\Sigma} \cos(\gamma) f^* \, dH^2 \le \mu \int_{\mathcal{A}_\delta} |Df| + \Upsilon_1 \int_{\mathcal{A}_\delta} |f|. \tag{62}$$

Since (58) and (62) are together equivalent to (48), we see that the lemma is proven when $(\gamma_1, \gamma_2) \in D_2^+$ and $\gamma_2 \le \pi/2$.

Case 2 ($(\gamma_1, \gamma_2) \in D_2^+$ and $\gamma_1 \ge \pi/2$). In this case, $\gamma_2 > \pi/2$. Let us set $\tilde{\gamma} = \pi - \gamma$, $\tilde{\gamma}_1 = \pi - \gamma_1$ and $\tilde{\gamma}_2 = \pi - \gamma_2$. Notice that $\tilde{\gamma}_1 - \tilde{\gamma}_2 = \gamma_2 - \gamma_1 > \pi - 2\alpha$ and so $(\tilde{\gamma}_1, \tilde{\gamma}_2) \in D_2^-$. Then $(\tilde{\gamma}_2, \tilde{\gamma}_1) \in D_2^+$ with $\tilde{\gamma}_1 \le \pi/2$. By reflecting Ω and γ about the x-axis, we see from our previous argument with $\gamma_2 \le \pi/2$ that (48) holds.

Case 3 ($(\gamma_1, \gamma_2) \in D_2^+$, $\gamma_1 < \pi/2$ and $\gamma_2 > \pi/2$). We use the same argument used to establish (48) when $\gamma_2 \le \pi/2$—that is, only one of the sides, $\partial^+\Omega$ or $\partial^-\Omega$, contributes to each integral since an inequality like (59) holds on the other side, and, by rotating through a suitable angle, we can make the intersection of the contributing side with a sufficiently small ball centered at O the graph of a function over the x-axis with arbitrarily small Lipschitz constant. Then we see that (48) holds in this case.

Case 4 ($(\gamma_1, \gamma_2) \in D_2^-$). In this case $\gamma_1 - \gamma_2 > \pi - 2\alpha$. Then $(\gamma_2, \gamma_1) \in D_2^+$ and, by reflection about the x-axis, we see our previous arguments show that (48) holds. □

Remark 3.2. Suppose $\Omega_j \to \Omega_\infty$ in that $\Omega_j = \{x \in \mathbb{R}^2 : \epsilon_j x \in \Omega\}$ when $\epsilon_j \to 0$ as $j \to \infty$ and $\Omega_\infty = \{(r\cos(\theta), r\sin(\theta)) : r > 0,\ -\alpha < \theta < \alpha\}$. Assuming we define other quantities appropriately (for example, $\gamma_j \in C^0(\partial\Omega_j)$ defined by $\gamma_j(x) = \gamma(\epsilon_j x)$ for $x \in \partial\Omega_j$), then an examination of the proofs of [Finn 1986, Lemmas 6.1 and 7.6] shows that the constants ζ, $a(\zeta)$, Υ and μ can be assumed to be independent of j in Lemma 3.1.

Remark 3.3. Notice, in particular, that if U is a Caccioppoli set in $\Omega \times \mathbb{R}$, then, with $f = \phi_U$ and $f = \phi_{U'}$, (48) implies

$$\left|\int_{\Sigma\times(-T,T)} \cos(\gamma)\phi_U^* dH_2\right| \le \mu \int_{\Sigma_\lambda\times(-T,T)} |D\phi_U| + \Upsilon \int_{\Sigma_\lambda\times(-T,T)} \phi_U \tag{63}$$

and

$$\left|\int_{\Sigma\times(-T,T)} \cos(\pi-\gamma)\phi_{U'}^* dH_2\right| \le \mu \int_{\Sigma_\lambda\times(-T,T)} |D\phi_{U'}| + \Upsilon \int_{\Sigma_\lambda\times(-T,T)} \phi_{U'}, \tag{64}$$

where $U' = (\Omega \times \mathbb{R}) \setminus U$ for $T > 0$.

Emmer's lemma (for example, [Emmer 1973]), in this case Lemma 3.1, is the key ingredient needed to obtain lower semicontinuity of the functional in question. Slight modifications of arguments in [Finn 1986, Section 7.4] and [Tam 1986b, Lemma 1.2] show that $\mathscr{E}$ and $\mathscr{F}_{T,K}$ for $T > 0$ and $K \subset\subset \mathbb{R}^2$ are lower semicontinuous. The proof of [Tam 1986a, Lemma 2.3] (also [Tam 1984, Lemma 2.3]), adapted to the situation here, yields this:

Lemma 3.2. *Let Ω and γ be as in Theorem 1.1, and note that (γ, Ω, O) is admissible. Let (ϵ_j) be a sequence of positive reals such that $\lim_{j\to\infty} \epsilon_j = 0$. For each $j \in \mathbb{N}$, set $H_j^*(x) = \epsilon_j H^*(\epsilon_j x)$ for $x \in \Omega_j$ and $\gamma_j(x) = \epsilon_j \gamma(\epsilon_j x)$ for $x \in \partial\Omega_j$. For each $j \in \mathbb{N}$, suppose f_j is a generalized solution for*

$$\mathscr{E}_j(u) = \int_{\Omega_j} \sqrt{1+ \mid Du \mid^2}\, dH_2 + \int_{\Omega_j} H_j^* u \, dH_2 - \int_{\partial\Omega_j} \cos(\gamma) u \, dH_1.$$

Then (f_j) has a subsequence (f_{j_i}) that converges locally to a generalized solution f_∞ for

$$\mathscr{E}_\infty(u) = \int_{\Omega_\infty} \sqrt{1+ \mid Du \mid^2}\, dH_2 - \int_{\partial^+\Omega_\infty} \cos(\gamma_1) u \, dH_1 - \int_{\partial^-\Omega_\infty} \cos(\gamma_2) u \, dH_1.$$

Acknowledgments

I wish to thank Harold Parks for his comments and Leon Simon for his comments and his sketch of an alternative proof of [Simon 1980, Section 4] using a scaling argument (for example, [Simon 1976]). I thank the referee for efforts and suggestions. I especially thank Robert Finn for his support and encouragement over the years, his discovery and investigation (with Paul Concus and others) of this conjecture and his editorial efforts. I am indebted to Eberhard Zeidler and the Max-Planck-Institut für Mathematik in den Naturwissenschaften, to Robert Finn, Laurence Nédélec, Rafe Mazzeo, Leon Simon and the Department of Mathematics at Stanford University and to Ben Andrews, Maria Athanassenas and the Centre for Mathematics and its Applications of the Australian National University for their hospitality during some of the course of this study.

References

[Allard 1972] W. K. Allard, "On the first variation of a varifold", *Ann. of Math.* (2) **95** (1972), 417–491. MR 46 #6136 Zbl 0252.49028

[Athanassenas and Lancaster 2008] M. Athanassenas and K. Lancaster, "CMC capillary surfaces at reentrant corners", *Pacific J. Math.* **234**:2 (2008), 201–228. MR 2009b:53007 Zbl 1148.76010

[Bear and Hile 1983] H. S. Bear and G. N. Hile, "Behavior of solutions of elliptic differential inequalities near a point of discontinuous boundary data", *Comm. Partial Differential Equations* **8**:11 (1983), 1175–1197. MR 85g:35054 Zbl 0533.35040

[Concus and Finn 1996] P. Concus and R. Finn, "Capillary wedges revisited", *SIAM J. Math. Anal.* **27**:1 (1996), 56–69. MR 96m:76006 Zbl 0843.76012

[Concus et al. 1992] P. Concus, R. Finn, and F. Zabihi, "On canonical cylinder sections for accurate determination of contact angle in microgravity", pp. 125–131 in *Fluid Mechanics Phenomena in*

Microgravity, Applied Mechanics Division **154**, American Society of Mechanical Engineers, New York, 1992.

[Courant 1977] R. Courant, *Dirichlet's principle, conformal mapping, and minimal surfaces*, Springer, New York, 1977. MR 56 #13103 Zbl 0354.30012

[Elcrat and Lancaster 1986] A. R. Elcrat and K. E. Lancaster, "Boundary behavior of a nonparametric surface of prescribed mean curvature near a reentrant corner", *Trans. Amer. Math. Soc.* **297**:2 (1986), 645–650. MR 87h:35098 Zbl 0602.35042

[Emmer 1973] M. Emmer, "Esistenza, unicità e regolarità nelle superfici de equilibrio nei capillari", *Ann. Univ. Ferrara Sez. VII* (*N.S.*) **18** (1973), 79–94. MR 49 #1281 Zbl 0275.49005

[Federer 1969] H. Federer, *Geometric measure theory*, Die Grundlehren der mathematischen Wissenschaften **153**, Springer, New York, 1969. MR 41 #1976 Zbl 0176.00801

[Finn 1986] R. Finn, *Equilibrium capillary surfaces*, Die Grundlehren der Mathematischen Wissenschaften **284**, Springer, New York, 1986. MR 88f:49001 Zbl 0583.35002

[Finn 1988a] R. Finn, "Comparison principles in capillarity", pp. 156–197 in *Partial differential equations and calculus of variations*, edited by S. Hildebrandt and R. Leis, Lecture Notes in Math. **1357**, Springer, Berlin, 1988. MR 90d:53010 Zbl 0692.35006

[Finn 1988b] R. Finn, "Moon surfaces, and boundary behaviour of capillary surfaces for perfect wetting and nonwetting", *Proc. London Math. Soc.* (3) **57**:3 (1988), 542–576. MR 89m:49076 Zbl 0668.76019

[Finn 1996] R. Finn, "Local and global existence criteria for capillary surfaces in wedges", *Calc. Var. Partial Differential Equations* **4**:4 (1996), 305–322. MR 97f:53005 Zbl 0872.76017

[Finn 1999] R. Finn, "Capillary surface interfaces", *Notices Amer. Math. Soc.* **46**:7 (1999), 770–781. MR 2000g:76033

[Finn 2002a] R. Finn, "Eight remarkable properties of capillary surfaces", *Math. Intelligencer* **24**:3 (2002), 21–33. MR 2003f:76041

[Finn 2002b] R. Finn, "Some properties of capillary surfaces", *Milan J. Math.* **70** (2002), 1–23. MR 2003m:53012 Zbl 1053.76009

[Finn and Serrin 1958] R. Finn and J. Serrin, "On the Hölder continuity of quasi-conformal and elliptic mappings", *Trans. Amer. Math. Soc.* **89** (1958), 1–15. MR 20 #4094 Zbl 0082.29401

[Gehring 1973] F. W. Gehring, "The L^p-integrability of the partial derivatives of a quasiconformal mapping", *Acta Math.* **130** (1973), 265–277. MR 53 #5861 Zbl 0258.30021

[Gehring 2005] F. W. Gehring, "Quasiconformal mappings in Euclidean spaces", pp. 1–29 in *Handbook of complex analysis: Geometric function theory*, vol. 2, edited by R. Kühnau, Elsevier, Amsterdam, 2005. MR 2005k:30044 Zbl 1078.30014

[Gilbarg and Trudinger 1983] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, 2nd ed., Grundlehren der Mathematischen Wissenschaften **224**, Springer, Berlin, 1983. MR 86c:35035 Zbl 0562.35001

[Giusti 1980] E. Giusti, "Generalized solutions for the mean curvature equation", *Pacific J. Math.* **88**:2 (1980), 297–321. MR 83a:35030 Zbl 0461.49024

[Giusti 1984] E. Giusti, *Minimal surfaces and functions of bounded variation*, Monographs in Mathematics **80**, Birkhäuser, Basel, 1984. MR 87a:58041 Zbl 0545.49018

[Iwaniec and Martin 2001] T. Iwaniec and G. Martin, *Geometric function theory and non-linear analysis*, Oxford University Press, New York, 2001. MR 2003c:30001 Zbl 1045.30011

[Jeffres and Lancaster 2007] T. Jeffres and K. Lancaster, "Vertical blow ups of capillary surfaces in $\mathbb{R}^3$, I: Convex corners", *Electron. J. Differential Equations* (2007), no. 152. MR 2009b:49103

[Kenmotsu 1979] K. Kenmotsu, "Weierstrass formula for surfaces of prescribed mean curvature", *Math. Ann.* **245**:2 (1979), 89–99. MR 81c:53005b Zbl 0402.53002

[Kenmotsu 2003] K. Kenmotsu, *Surfaces with constant mean curvature*, Translations of Mathematical Monographs **221**, American Mathematical Society, Providence, RI, 2003. MR 2004m:53014 Zbl 1042.53001

[Lancaster 1985] K. E. Lancaster, "Boundary behavior of a nonparametric minimal surface in $\mathbf{R}^3$ at a nonconvex point", *Analysis* **5**:1-2 (1985), 61–69. Corrected in **6** (1986), 413. MR 86m:49053 Zbl 0601.35035

[Lancaster and Siegel 1996a] K. E. Lancaster and D. Siegel, "Existence and behavior of the radial limits of a bounded capillary surface at a corner", *Pacific J. Math.* **176**:1 (1996), 165–194. Figures corrected in **179**:2 (1997), 397–402. MR 98g:58030a Zbl 0866.76018

[Lancaster and Siegel 1996b] K. E. Lancaster and D. Siegel, "Behavior of a bounded non-parametric H-surface near a reentrant corner", *Z. Anal. Anwendungen* **15**:4 (1996), 819–850. MR 97m:53011 Zbl 0866.35046

[Massari and Pepe 1975] U. Massari and L. Pepe, "Successioni convergenti di ipersuperfici di curvatura media assegnata", *Rend. Sem Mat. Univ. Padova* **53** (1975), 53–68. MR 54 #8415 Zbl 0358.49020

[Miranda 1977] M. Miranda, "Superficie minime illimitate", *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4) **4**:2 (1977), 313–322. MR 58 #18063 Zbl 0352.49020

[Monakhov 1983] V. N. Monakhov, *Boundary value problems with free boundaries for elliptic systems of equations*, Translations of Mathematical Monographs **57**, American Mathematical Society, Providence, RI, 1983. MR 85a:35029 Zbl 0532.35001

[Shi and Finn 2004] D. Shi and R. Finn, "On a theorem of Lancaster and Siegel", *Pacific J. Math.* **213**:1 (2004), 111–119. MR 2004m:76038 Zbl 1156.76361

[Simon 1976] L. Simon, "Remarks on curvature estimates for minimal hypersurfaces", *Duke Math. J.* **43**:3 (1976), 545–553. MR 54 #6040 Zbl 0348.53003

[Simon 1980] L. Simon, "Regularity of capillary surfaces over domains with corners", *Pacific J. Math.* **88**:2 (1980), 363–377. MR 82d:49043 Zbl 0467.35022

[Tam 1984] L. F. Tam, *The Behavior of Capillary Surfaces as Gravity Tends to Zero*, thesis, Stanford University, 1984.

[Tam 1986a] L.-F. Tam, "The behavior of capillary surfaces as gravity tends to zero", *Comm. Partial Differential Equations* **11**:8 (1986), 851–901. MR 89c:76040 Zbl 0607.35041

[Tam 1986b] L.-F. Tam, "On existence criteria for capillary free surfaces without gravity", *Pacific J. Math.* **125**:2 (1986), 469–485. MR 88a:49023 Zbl 0604.49030

[Tam 1986c] L.-F. Tam, "Regularity of capillary surfaces over domains with corners: Borderline case", *Pacific J. Math.* **124**:2 (1986), 469–482. MR 87k:49049 Zbl 0604.49029

[Taylor 1976] J. E. Taylor, "The structure of singularities in soap-bubble-like and soap-film-like minimal surfaces", *Ann. of Math.* (2) **103**:3 (1976), 489–539. MR 55 #1208a Zbl 0335.49032

[Taylor 1977] J. E. Taylor, "Boundary regularity for solutions to various capillarity and free boundary problems", *Comm. Partial Differential Equations* **2**:4 (1977), 323–357. MR 58 #7336 Zbl 0357.35010

Received December 28, 2007. Revised March 17, 2010.

Kirk E. Lancaster
Department of Mathematics and Statistics
Wichita State University
Wichita, KS 67260-0033
United States
lancaster@math.wichita.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

THE EXISTENCE AND MONOTONICITY OF A THREE-DIMENSIONAL TRANSONIC SHOCK IN A FINITE NOZZLE WITH AXISYMMETRIC EXIT PRESSURE

JUN LI, ZHOUPING XIN AND HUICHENG YIN

We establish the existence of a multidimensional transonic shock solution in a class of slowly varying nozzles for the three dimensional steady full Euler system with axially symmetric exit pressure in the diverging part lying in an appropriate scope. We also show that the shock position depends monotonically on the exit pressure.

1. Introduction and the main results

The transonic shock problem in a de Laval nozzle is a fundamental one in fluid dynamics and has been extensively studied by many authors under the assumption that the transonic flow is quasi-one-dimensional or the transonic shock goes through some fixed point in advance [Chen et al. 2006; Chen et al. 2007; Chen and Feldman 2003; Chen 2008; Courant and Friedrichs 1948; Embid et al. 1984; Glaz and Liu 1984; Kuz'min 2002; Liu 1982a; 1982b; Xin et al. 2009; Xin and Yin 2005; 2008a; 2008b; Yuan 2006]. Courant and Friedrichs [1948, page 386] proposed a physically more interesting transonic shock wave pattern in a de Laval nozzle as follows: Given an appropriately large end pressure $p_e(x)$, if the upstream flow is still supersonic behind the throat of the nozzle, then at a certain place in the diverging part of the nozzle a shock front intervenes and the gas is compressed and slowed down to subsonic speed. The position and the strength of the shock front are automatically adjusted so that the end pressure at the exit becomes $p_e(x)$. This means that the position of the transonic shock should be completely free. Indeed, the assumption that the shock goes through some fixed point at the wall of the nozzle in advance may lead to overdetermined boundary conditions for the

MSC2000: primary 35L65, 35L67, 35L70; secondary 76N15.

Keywords: steady Euler system, multidimensional transonic shock, axisymmetric, first order elliptic system.

Li Jun and Yin Huicheng are supported by the National Natural Science Foundation of China grants 10871096 and 10931007 and the National Basic Research Program of China grant 2006CB805902. Xin Zhouping is supported in part by Hong Kong RGC Earmarked Research grants CUHK4042/08P and CUHK4028/04P, and a Focus Area Grant from The Chinese University of Hong Kong. Huicheng Yin is the corresponding author.

transonic shock problem for the full Euler system with the given exit pressure; see [Xin et al. 2009; Xin and Yin 2008a] for details. Here, we focus on the existence of a solution to this transonic shock problem for the three-dimensional full Euler system when the exit pressure $p_e(x)$ is axisymmetric and lies in an appropriate scope without other artificial constraints. In particular, we show the shock position depends monotonically on the exit pressure.

The steady and nonisentropic Euler system in three-dimensional space is

$$\begin{cases} \operatorname{div}(\rho u) = 0, \\ \operatorname{div}(\rho u \otimes u) + \nabla P = 0, \\ \operatorname{div}\big((\rho(e + \frac{1}{2}|u|^2) + P)u\big) = 0, \end{cases} \tag{1-1}$$

where $u = (u_1, u_2, u_3)$, ρ, P, e and S stand for the velocity, density, pressure, internal energy and specific entropy, respectively. The pressure function $P = P(\rho, S)$ and the internal energy function $e = e(\rho, S)$ are smooth in their arguments. It is assumed that $\partial_\rho P(\rho, S) > 0$ and $\partial_S e(\rho, S) > 0$ for $\rho > 0$.

For the ideal polytropic gases, the equations of state are given by

$$P = A\rho^\gamma e^{S/c_v} \quad \text{and} \quad e = \frac{P}{(\gamma-1)\rho},$$

where A, c_v and γ are positive constants, and $1 < \gamma < 3$ (in air, $\gamma \approx 1.4$).

We now describe the class of de Laval nozzle that will be studied later on; see also [Li et al. 2010a; 2010b]. The wall Γ of the nozzle is assumed to be $C^{3,\alpha}$-regular for $X_0 - 1 \le r \equiv (x_1^2 + x_2^2 + x_3^2)^{1/2} \le X_0 + 1$, where $X_0 > 0$ is a fixed large constant, and $\alpha \in (0, 1)$ and Γ consists of two curved surfaces Π_1 and Π_2; here Π_1 includes the converging part of the nozzle, and Π_2 constructs a symmetric curved diverging part of it. See Figure 1. More precisely, Π_2 is represented by the equation $x_2^2 + x_3^2 = x_1^2 \tan^2\theta_0$ with $x_1 > 0$ and $X_0 < r < X_0 + 1$, where θ satisfies $0 < \theta_0 < \pi/2$ and is sufficiently small. For simplicity, we assume that the $C^{3,\alpha}$-smooth supersonic incoming flow $(S_0^-, P_0^-(x), u_0^-(x))$ is spherically symmetric near $r = X_0$; here $S_0^-(x) = S_0^-$ is a constant, $P_0^-(x) = P_0^-(r)$, and $u_0^-(x) = U_0^-(r)x/r$. This assumption is easily satisfied because of the hyperbolicity of the supersonic incoming flow and the symmetry of Π_2.

Let shock Σ in the nozzle be given by $x_1 = \eta(x')$ with $x' = (x_2, x_3)$, and denote the flow field behind the shock by $(S^+(x), P^+(x), u^+(x))$. The Rankine–Hugoniot conditions on Σ imply

$$\begin{cases} [(1, -\nabla_{x'}\eta(x')) \cdot \rho u] = 0, \\ [((1, -\nabla_{x'}\eta(x')) \cdot \rho u)u] + (1, -\nabla_{x'}\eta(x'))^t [P] = 0, \\ [(1, -\nabla_{x'}\eta(x')) \cdot (\rho(e + \frac{1}{2}|u|^2) + P)u] = 0. \end{cases} \tag{1-2}$$

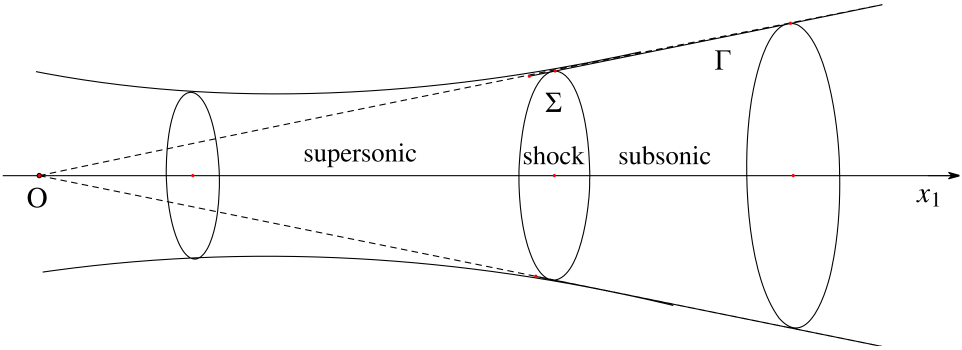

Figure 1

Here the brackets around function denotes the jump of that function across Σ.

In addition, $P^+(x)$ should satisfy the physical entropy condition

$$P^+(x) > P^-(x) \quad \text{on } x_1 = \eta(x_2, x_3); \tag{1-3}$$

see [Courant and Friedrichs 1948].

On the exit of the nozzle, the pressure is prescribed and axisymmetric:

$$P^+(x) = P_e + \varepsilon \tilde{P}(\theta) \quad \text{on } r = X_0 + 1. \tag{1-4}$$

Here P_e is a positive constant, $\varepsilon > 0$ is sufficiently small, $\theta = r^{-1} \arcsin(x_2^2 + x_3^2)^{1/2}$, and $\tilde{P}(\theta) \in C^{2,\alpha}[0, \theta_0]$ with $\tilde{P}'(0) = \tilde{P}'(\theta_0) = 0$. We require that for given exit pressure P_e, the Euler system (1-1) has a radial symmetric transonic shock lying at $r = r_0 \in (X_0, X_0 + 1)$ with supersonic incoming flow $(S_0^-, P_0^-(r), (U_0^-(r)/r)x)$ for $r \in (X_0, r_0)$. For the range of P_e and detailed information on the corresponding transonic shock solution $(S_0^\pm, P_0^\pm(r), (U_0^\pm(r)/r)x)$, see Theorem A.1.

The wall of the nozzle is assumed to be solid; thus

$$x_1 u_1^+ \tan^2 \theta_0 - x_2 u_2^+ - x_3 u_3^+ = 0 \quad \text{on } \Pi_2. \tag{1-5}$$

Finally, we assume X_0 and θ_0 to be suitably large and small respectively so that

$$X_0 \theta_0 = 1 \quad \text{and} \quad \tfrac{1}{2}\eta_0 < \theta_0 < \eta_0. \tag{1-6}$$

Here $\eta_0 > 0$ is a constant.

Note that (1-6) means that the nozzle wall $\Pi_2 : x_2^2 + x_3^2 = x_1^2 \tan^2 \theta_0$ is close to the cylindrical surface $x_2^2 + x_3^2 = 1$ for $X_0 \le r \le X_0 + 1$.

The main results in this paper can be stated as follows:

Theorem 1.1 (existence and monotonicity). *Under the assumptions above*, *with*

$$M_0^-(X_0) \equiv \frac{U_0^-(X_0)}{c(\rho_0^-(X_0))} > \sqrt{\frac{\gamma+3}{2}}$$

and $\varepsilon < 1/X_0^3$, *the problem* (1-1) *with the conditions* (1-2)–(1-5) *has a solution* $(S^+(x), P^+(x), u^+(x); \eta(x_2, x_3))$ *that admits the following estimates*:

(i) $\eta(x_2, x_3) \in C^{3,\alpha}(\bar{S}_e)$, with $S_e = \{(x_2, x_3) : (\eta(x_2, x_3), x_2, x_3) \in \Sigma\}$ being the projection of the shock surface Σ onto the (x_2, x_3)-plane. Moreover, there exists a constant $C_0 > 0$ (depending only on α and the supersonic incoming flow) such that

$$\|\eta(x_2, x_3) - (r_0^2 - x_2^2 - x_3^2)^{1/2}\|_{L^\infty(S_e)} \le C_0 X_0 \varepsilon,$$
$$\|\nabla_{x_2,x_3}(\eta(x_2, x_3) - (r_0^2 - x_2^2 - x_3^2)^{1/2})\|_{C^{2,\alpha}(\bar{S}_e)} \le C_0 \varepsilon.$$

(ii) Denote by

$$\Omega_+ = \{(x_1, x_2, x_3) : \eta(x_2, x_3) < x_1 < ((X_0+1)^2 - x_2^2 - x_3^2)^{1/2},\ x_2^2 + x_3^2 \le x_1^2 \tan^2 \theta_0\}$$

the subsonic region. Then $(S^+(x), P^+(x), u^+(x)) \in C^{2,\alpha}(\overline{\Omega}_+)$ satisfies

$$\|(S^+(x), P^+(x), u^+(x)) - (S_0^+, \hat{P}_0^+(r), \hat{u}_0^+(x))\|_{C^{2,\alpha}(\bar{\Omega}_+)} \le C_0 \varepsilon,$$

where $\hat{u}_0^+(x) = \hat{U}_0^+(r)x/r$, and $(S_0^+, \hat{P}_0^+(r), \hat{u}_0^+(r))$ stands for the extension of the background solution $(S_0^+, P_0^+(r), U_0^+(r)x/r)$ in Ω_+ described in more detail in Theorem A.1 and Remark A.3.

(iii) The position of the shock surface depends on the given exit pressure monotonically and continuously.

Remark 1.2. Showing that the shock position depends monotonically on the exit pressure is one of the keys to the existence result described by Theorem 1.1. When the exit pressure changes at order $O(\varepsilon)$, the shock position will change at order $X_0 O(\varepsilon)$ instead of $O(1)\varepsilon$; this will be crucial in our analysis.

Remark 1.3. The condition

$$M_0^-(X_0) \equiv \frac{U_0^-(X_0)}{c(\rho_0^-(X_0))} > \sqrt{\frac{\gamma+3}{2}}$$

on the supersonic Mach number is there to ensure that the shock position along the nozzle wall is monotonic in the subsonic pressure across the shock; this is the initial step toward showing the monotonic dependence of the shock position on the exit pressure. See (4-34), (4-36), (4-38), and (4-39) for more details.

Remark 1.4. Although in [Li et al. 2010a] we established by a completely different method (see [Li et al. 2009a] also) the existence of a three-dimensional transonic shock for a variety of conic nozzles with axisymmetric exit pressures, we did not show monotonic dependence of the shock position on the exit pressure.

There has already been much work on the steady transonic problem; see [Bers 1950; 1951; Čanić et al. 2000; Chen et al. 2006; Chen et al. 2007; Chen and Feldman 2003; Chen 2008; Courant and Friedrichs 1948; Embid et al. 1984; Glaz and Liu 1984; Kuz'min 2002; Li et al. 2009a; 2009b; 2010a; 2010b; Liu 1982a; 1982b; Morawetz 1994; Xin et al. 2009; Xin and Yin 2005; 2008b; 2008a; Yuan 2006; Zheng 2003; 2006] and the references therein. In particular, for a three-dimensional nozzle with a symmetric diverging part and a symmetric supersonic incoming flow near the diverging part of the nozzle, Xin and Yin [2008b] and Courant and Friedrichs [1948] have shown that there exist two constant pressures P_1 and P_2 with $P_1 < P_2$ such that if the exit pressure P_e is in the interval (P_1, P_2), then the transonic shock exists uniquely in the diverging part of the nozzle, and the position and the strength of the shock are completely determined by P_e and the resulting ordinary differential equations. Xin and Yin [2008b] also established global existence, stability and long time asymptotic behavior of an unsteady symmetric transonic shock under the exit pressure P_e when the initial unsteady shock lies in the symmetric diverging part of the three-dimensional nozzle; on the other hand a steady symmetric transonic shock is dynamically unstable if it lies in the symmetric converging part of the nozzle. In [Li et al. 2009b], we established for the two-dimensional steady Euler system, by a monotonicity argument on the shock position and the exit pressure, uniqueness and existence of a completely free two-dimensional transonic shock in a nozzle with variable end pressures at the exit. For the three-dimensional steady Euler system, we have shown in [Li et al. 2010b] the uniqueness of a completely free three-dimensional transonic shock solution of class $C^{3,\alpha}$ in a nozzle with general exit pressure; this regularity is higher than the $C^{2,\alpha}$ regularity of solutions in Theorem 1.1. In this paper, we will focus on the existence and monotonicity property of a completely free three-dimensional transonic shock for a certain class of the exit pressures.

Next we comment on the proofs of the main results in this paper. In almost all previous results dealing with transonic shocks in a nozzle with given exit pressure except, except for those in [Li et al. 2009b; 2010a; 2010b; Xin et al. 2009], the authors assume that the shock goes through a fixed point in advance; this plays the crucial role in the analysis, in particular, in the process of determining the shock position. However, for de Laval nozzles, this assumption is not physical since the shock position should be determined by the supersonic incoming flow, the geometry of the nozzle and the exit pressure, as pointed out by Courant and Friedrichs. Moreover, this constraint may lead in general to an over-determined problem. In [Li et al. 2009b; 2010a; 2010b], we have successfully removed this condition, and further determined the shock position and transonic flow in the nozzle. This leads to the well-posedness of the transonic shock problem in the two-dimensional case and the uniqueness of solutions to it in the three-dimensional case, as well as some new observations and techniques.

A key step in [Li et al. 2009b; 2010b] is to derive a priori gradient estimates instead of the solution itself, in order to establish that the shock position along the walls of the nozzle varies monotonically with exit pressure. This leads to the determination of a unique shock position and the desired stability estimates. However, it seems difficult to apply these methods directly to obtain the existence of the transonic shock in a three-dimensional nozzle. The main reasons are as follows: $C^{3,\alpha}$ regularity of the solution in the subsonic region plays a fundamental role in the theorems, but this higher order regularity is a source of great difficulties for nozzles with variable exit pressure. Compared with two-dimensional case, it seems much more difficult to find higher order compatibility conditions near the intersection curve of the shock surface with the wall of the nozzle, which is necessary to ensure $C^{3,\alpha}$ regularity of the solution nearby. In the two-dimensional case, higher order compatibility at the intersection points of the shock curve with the walls of the nozzle can be found directly from the Euler system together with the no-flow boundary condition of the walls of the nozzle, and Rankine–Hugoniot conditions on the shock curve. This yields naturally $C^{3,\alpha}$ regularity of the solution in [Li et al. 2009b]; similar approaches cannot be applied in the three-dimensional case; see [Xin and Yin 2008b, Lemma 6.1]. In addition, for the axially symmetric exit pressure in this paper, it is natural to introduce spherical coordinates in the space variables, which brings new technical difficulties in finding compatibility conditions on the symmetry axis and handling singularities and source terms in the transformed equations near the symmetry axis. Due to the singularity near the symmetry axis and the source terms for the Euler system in spherical coordinates, the key gradient estimate method in [Li et al. 2009b] cannot be applied here; see (2-8) and Remark 3.3.

To overcome these difficulties, our strategy is as follows: First, we will give some rather delicate computations and analysis of the three-dimensional Euler system and the related axisymmetric functions near the x_1-axis and the nozzle wall; this is to establish $C^{2,\alpha}$ regularity of the solutions; see Lemmas B.1–B.7 and Section 3. Second, to derive that the shock position is monotonic in the end pressure, we will focus directly on the first order elliptic system and how the two pressures and two shock positions (see (4-17)) differ from those in the gradient estimates of [Li et al. 2009b; 2010b]. The key step is to establish an ordinary differential-integral inequality in the difference of pressures (see (4-45)). Based on this result and the continuous dependence of the shock position on the exit pressure, we can finally complete the proof of Theorem 1.1.

The rest of the paper is organized as follows. In Section 2, we will reformulate the three-dimensional problem (1-1) with the boundary conditions (1-2)–(1-5). First we transform the nozzle wall into a cube surface, and decompose the velocity u^+ as the radial speed U_1^+ and two angular speeds U_2^+ and U_3^+. In the Euler system on $(S^+, P^+, U_1^+, U_2^+, U_3^+)$, with the exit boundary condition (1-4), it is natural to search for a solution with $U_3^+ \equiv 0$. Furthermore, we decompose the Euler system

(1-1) as a 2×2 first order elliptic system for ρ^+ and U_2^+/U_1^+, and two algebraic equations in U_1^+ and specific entropy S^+ respectively. In Section 3, we use the decomposition in Section 2 to linearize the compressible Euler system, establish an existence result under the assumption that the shock goes through some fixed point at the nozzle wall in advance, and obtain some key estimates based on the background solution. We note that this solution does not satisfy the boundary condition (1-4) unless the exit pressure is adjusted by an appropriate constant. In Section 4, we establish that the shock position is monotonic in the end pressure. In Section 5, we use the continuous dependence of the solution on the shock position to the existence result in Theorem 1.1. In Appendix A, we list some properties of the background solution. We give some useful inequalities and estimates in Appendix B. Finally, in Appendix C we give a detailed discussion of the regularity of $C^{3,\alpha}$ solutions to problem (1-1) with (1-2)–(1-5).

We will use the following conventions:

$O(\varepsilon)$ means that there exists a generic constant $C_1 > 0$ independent of X_0 and ε such that $\|O(\varepsilon)\|_{C^{1,\alpha}} \le C_1\varepsilon$.

$O(1/X_0^m)$ for $m > 0$ means that there exists a generic constant $C_2 > 0$ independent of X_0 and ε such that $\|O(1/X_0^m)\|_{C^{1,\alpha}} \le C_2/X_0^m$.

2. Reformulation of the problem

In this section, we will reformulate the nonlinear problem (1-1) with (1-2)–(1-5) to obtain a coupled first order elliptic system in the angular velocity exponent U_2^+ and the density ρ^+, and two first order equations, one in the radial velocity U_1^+ and the other in the specific entropy S^+. As in [Xin and Yin 2008b], we will need to derive relations between (ρ^+, U_1^+) and (U_2^+, U_3^+) in the shock Σ. Due to the symmetry of the nozzle wall Π_2 and the supersonic incoming flow in the diverging part, it will be more convenient to use the spherical coordinates

$$x_1 = r\cos\theta, \quad x_2 = r\sin\theta\cos\varphi, \quad x_3 = r\sin\theta\sin\varphi \tag{2-1}$$

and velocity decomposition

$$\begin{aligned} U_1^+ &= u_1^+\cos\theta + u_2^+\sin\theta\cos\varphi + u_3^+\sin\theta\sin\varphi,\\ U_2^+ &= u_1^+\sin\theta - u_2^+\cos\theta\cos\varphi - u_3^+\cos\theta\sin\varphi,\\ U_3^+ &= -u_2^+\sin\varphi + u_3^+\cos\varphi, \end{aligned} \tag{2-2}$$

where $\theta \in [0, \theta_0]$, $\varphi \in [0, 2\pi]$, and $r = (x_1^2 + x_2^2 + x_3^2)^{1/2}$.

In the spherical coordinates (2-1), set

$$\tilde{\nabla} := \Big(\partial_r, -\frac{1}{r}\partial_\theta, \frac{1}{r\sin\theta}\partial_\varphi\Big) \quad \text{and} \quad \tilde{U} = (U_1, U_2, U_3).$$

Then (1-1) and (1-2) are transformed respectively into

$$\begin{cases} \tilde{\nabla}\cdot(\rho^+\tilde{U}^+)+\Big(\frac{2}{r},-\frac{1}{r}\cot\theta\Big)\rho^+\cdot(U_1^+,U_2^+)=0,\\ (\tilde{U}^+\cdot\tilde{\nabla})\tilde{U}^+ +\frac{\tilde{\nabla}P^+}{\rho^+}+\frac{1}{r}\begin{pmatrix} -((U_2^+)^2+(U_3^+)^2)\\ U_1^+U_2^+ +(U_3^+)^2\cot\theta\\ U_1^+U_2^+ - U_2^+U_3^+\cot\theta\end{pmatrix}=0,\\ (\tilde{U}^+\cdot\tilde{\nabla})S^+=0,\end{cases} \tag{2-3}$$

and

$$\begin{cases} [\rho\tilde{U}]\cdot\Big(1,\frac{1}{\tilde{r}}\partial_\theta\tilde{r},-\frac{\partial_\varphi\tilde{r}}{\tilde{r}\sin\theta}\Big)=0,\\ [\rho\tilde{U}\otimes\tilde{U}+PI]\cdot\Big(1,\frac{1}{\tilde{r}}\partial_\theta\tilde{r},-\frac{\partial_\varphi\tilde{r}}{\tilde{r}\sin\varphi}\Big)=0\\ \big[(\rho(e+\frac{1}{2}|\tilde{U}|^2)+P)\tilde{U}\big]\cdot\Big(1,\frac{1}{\tilde{r}}\partial_\theta\tilde{r},-\frac{\partial_\varphi\tilde{r}}{\tilde{r}\sin\varphi}\Big)=0,\end{cases} \tag{2-4}$$

where $r=\tilde{r}(\theta,\varphi)$ is the equation of the shock surface Σ in the spherical coordinates (r,θ,φ).

Meanwhile, (1-4) and (1-5) are correspondingly changed into

$$P^+(r,\theta,\varphi)=P_e+\varepsilon\tilde{P}(\theta)\quad\text{on } r=X_0+1 \tag{2-5}$$

and

$$U_2^+=0\quad\text{on } \theta=\theta_0. \tag{2-6}$$

For the axisymmetric exit pressure (1-4), we will search for solutions of (2-3)–(2-6) in the form

$$(S^+,P^+,\tilde{U}^+;\tilde{r})=(S^+(r,\theta),P^+(r,\theta),U_1^+(r,\theta),U_2^+(r,\theta),0;\tilde{r}(\theta)), \tag{2-7}$$

that is, we look for a solution and shock surface independent of the variable φ.

In this case, using the notation

$$U\equiv(U_1,U_2),\qquad U^\perp\equiv(-U_2,U_1),\qquad \nabla\equiv(\partial_r,-(1/r)\partial_\theta),$$

we can simplify (2-3) and (2-4) to

$$\begin{cases} \nabla\cdot(\rho^+U^+)+\frac{1}{r}\rho^+(2,-\cot\theta)\cdot U^+=0,\\ (U^+\cdot\nabla)U^+ +\frac{1}{\rho^+}\nabla P^+ +\frac{U_2^+}{r}(U^+)^\perp=0,\\ (U\cdot\nabla)S^+=0,\end{cases} \tag{2-8}$$

and

$$\begin{cases} [\rho U]\cdot\Big(1, \dfrac{\tilde r'(\theta)}{\tilde r(\theta)}\Big)=0, \\ [\rho U\otimes U+PI]\cdot\Big(1, \dfrac{\tilde r'(\theta)}{\tilde r(\theta)}\Big)=0, \\ \big[(\rho(e+\tfrac{1}{2}|U|^2)+P)U\big]\cdot\Big(1, \dfrac{\tilde r'(\theta)}{\tilde r(\theta)}\Big)=0. \end{cases} \tag{2-9}$$

For convenience, we use the transformation

$$y_1=r \quad\text{and}\quad y_2=X_0\theta, \tag{2-10}$$

to change the fixed wall Π_2 into $y_2=1$.

In the following, we will drop the $+$ superscripts for simplicity in presentation.

In this case, (2-8) and (2-9) can be rewritten respectively as

$$\begin{cases} \nabla_y\cdot(\rho U)+\dfrac{\rho}{y_1}U\cdot\Big(2, -\cot\Big(\dfrac{y_2}{X_0}\Big)\Big)=0, \\ (U\cdot\nabla_y)U+\dfrac{1}{\rho}\nabla_y P+\dfrac{U_2}{y_1}U^\perp=0, \\ (U\cdot\nabla_y)S=0, \end{cases} \tag{2-11}$$

and

$$\begin{pmatrix} [\rho U] \\ [\rho U\otimes U+PI] \\ [(\rho(e+\frac{1}{2}|U|^2)+P)U] \end{pmatrix}\cdot\begin{pmatrix} 1 \\ \dfrac{X_0\xi'(y_2)}{\xi(y_2)} \end{pmatrix}=0, \tag{2-12}$$

where $\nabla_y\equiv(\partial_{y_1}, -(X_0/y_1)\partial_{y_2})$ and $\xi(y_2)=\tilde r(y_2/X_0)$, and (2-5) and (2-6) become respectively

$$P(y)=P_e+\varepsilon\tilde P(y_2/X_0) \quad \text{on } y_1=X_0+1 \tag{2-13}$$

and

$$U_2=0 \quad \text{on } y_2=1. \tag{2-14}$$

Next, we derive boundary conditions of (P, S, U_1) on the shock surface.

It follows from (2-12) that

$$\xi'(y_2)=-\frac{\xi(y_2)}{X_0}\frac{[\rho U_1U_2]}{[\rho U_2^2+P]}. \tag{2-15}$$

This, together with (2-12), yields on Σ that

$$
\begin{aligned}
&G_1(\rho, U, S) \equiv [\rho U_1][\rho U_2^2 + P] - \rho^2 U_1 U_2^2 = 0,\\
&G_2(\rho, U, S) \equiv [\rho U_1^2 + P][\rho U_2^2 + P] - (\rho U_1 U_2)^2 = 0,\\
&G_3(\rho, U, S) \equiv [(\rho e + \tfrac{1}{2}\rho|U|^2 + P)U_1][\rho U_2^2 + P]\\
&\qquad\qquad\qquad - \rho U_1(\rho e + \tfrac{1}{2}\rho|U|^2 + P)U_2^2 = 0.
\end{aligned}
\tag{2-16}
$$

It follows from a direct computation and the implicit function theorem that at the shock position Σ

$$
\begin{aligned}
&(S - S_0^+, P - P_0^+, U_1 - \hat{U}_0^+)(r_0)\\
&\qquad = (\tilde{g}_1, \tilde{g}_2, \tilde{g}_3)(U_2^2, P_0^- - P_0^-(r_0), U_0^- - U_0^-(r_0)),
\end{aligned}
\tag{2-17}
$$

where $\tilde{g}_j$ is smooth in its arguments and satisfies $\tilde{g}_j(0,0,0) = 0$ for $j = 1, 2, 3$. Moreover, by (1-6), the expected estimates in Theorem 1.1, and Remarks A.2 and A.3, it can be verified that

$$\tilde{g}_i = (O(\varepsilon) + O(1/X_0))(O(U_2) + O(\xi(y_2) - r_0)) \quad \text{for } i = 1, 2, 3.$$

This implies that on the shock surface, the influence of U_2 and $\xi(y_2) - r_0$ on $S - S_0^+$, $U_1 - \hat{U}_0^+$ and $P - \hat{P}_0^+$ can be almost neglected.

On the other hand, due to (2-1) and (2-10), the extension $(S_0^\pm, \hat{P}_0^\pm(r), \hat{U}_0^\pm(r))$ of the background solution in Appendix A will be changed into

$$(S_0^\pm, \hat{P}_0^\pm(y), \hat{U}_0^\pm(y)), \tag{2-18}$$

which satisfies for large X_0

$$\left|\frac{d^k \hat{P}_0^\pm(y_1)}{dy_1^k}\right| + \left|\frac{d^k \hat{U}_0^\pm(y_1)}{dy_1^k}\right| \le \frac{C}{X_0^k} \quad \text{for } k = 1, 2, 3, \tag{2-19}$$

where the constant $C > 0$ is independent of X_0 (see Remark A.2).

To treat the system (2-11) with (2-12)–(2-14), we introduce new coordinates

$$z_1 = \frac{y_1 - \xi(y_2)}{X_0 + 1 - \xi(y_2)} \quad \text{and} \quad z_2 = y_2, \tag{2-20}$$

which changes the free domain

$$R_+ = \{(y_1, y_2) : \xi(y_2) < y_1 < X_0 + 1,\ 0 < y_2 < 1\} \tag{2-21}$$

into a fixed square

$$E_+ = \{(z_1, z_2) : 0 < z_1 < 1,\ 0 < z_2 < 1\}. \tag{2-22}$$

There coordinates will decouple the system (2-11) with (2-12)–(2-14).

With some abuse of notation, we set

$$(S,P,U_1,U_2)(z)=(S,P,U_1,U_2)(\xi(z_2)+z_1(X_0+1-\xi(z_2)),z_2), \tag{2-23}$$

$$(\hat{P}_0^+,\hat{U}_0^+)(z_1)=(\hat{P}_0^+,\hat{U}_0^+)(r_0+z_1(X_0+1-r_0)). \tag{2-24}$$

Define

$$w=U_2/U_1. \tag{2-25}$$

We now derive a first order elliptic system in w and P.

In fact,

$$\frac{1}{\rho U_1^2}\times\big((\text{the third equation in (2-11)})-U_2\times(\text{the first equation in (2-11)})\big),$$

together with the fourth equation in (2-11), yields

$$\partial_{y_1}w-\frac{X_0}{y_1}\Big(\frac{1}{\rho U_1^2}-\frac{w^2}{\gamma P}\Big)\partial_{y_2}P-\frac{w}{\gamma P}\partial_{y_1}P-\frac{w}{y_1}+\frac{w^2}{y_1}\cot\frac{y_2}{X_0}=0.$$

While

$$\frac{y_1}{X_0\rho U_1^2}\times\big((\text{the second equation in (2-11)})-U_1\times(\text{the first equation in (2-11)})\big)$$

yields

$$\partial_{y_2}w+\frac{w}{X_0}\cot\frac{y_2}{X_0}+\frac{y_1}{X_0}\Big(\frac{1}{\rho U_1^2}-\frac{1}{\gamma P}\Big)\partial_{y_1}P+\frac{w}{\gamma P}\partial_{y_2}P-\frac{w^2+2}{X_0}=0.$$

In the (z_1,z_2) coordinates, we then have in E_+

$$\begin{aligned}\partial_{z_1}w-a_1\partial_{z_2}P&=F_1(S,P,U_1,U_2;\xi),\\ \partial_{z_2}w+\frac{1}{X_0}\cot\frac{z_2}{X_0}w+a_2\partial_{z_1}P&=F_2(S,P,U_1,U_2;\xi),\end{aligned} \tag{2-26}$$

where

$$\begin{aligned}a_1&=\frac{X_0(X_0+1-r_0)}{r_0}\frac{1}{\hat{\rho}_0^+(0)(\hat{U}_0^+(0))^2},\\ a_2&=\frac{r_0}{X_0(X_0+1-r_0)}\Big(\frac{1}{\hat{\rho}_0^+(0)(\hat{U}_0^+(0))^2}-\frac{1}{\gamma\hat{P}_0^+(0)}\Big),\end{aligned}$$

and

$$
\begin{aligned}
&F_1(S,P,U_1,U_2;\xi)\\
&\quad=\frac{X_0}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}\Big(\frac{1}{\rho U_1^2}-\frac{w^2}{\gamma P}\Big)\big((z_1-1)\xi'(z_2)\partial_{z_1}\\
&\qquad\qquad+(X_0+1-\xi(z_2))\partial_{z_2}\big)P+\frac{w}{\gamma P}\partial_{z_1}P-a_1\partial_{z_2}P\\
&\qquad\qquad+\frac{w(X_0+1-\xi(z_2))}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}-\frac{w^2(X_0+1-\xi(z_2))}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}\cot\frac{z_2}{X_0},\\
&F_2(S,P,U_1,U_2;\xi)\\
&\quad=a_2\partial_{z_1}P-\frac{\xi(z_2)+z_1(X_0+1-\xi(z_2))}{X_0(X_0+1-\xi(z_2))}\Big(\frac{1}{\rho U_1^2}-\frac{1}{\gamma P}\Big)\partial_{z_1}P\\
&\qquad\qquad-\frac{w}{\gamma P}\Big(\frac{(z_1-1)\xi'(z_2)}{X_0+1-\xi(z_2)}\partial_{z_1}+\partial_{z_2}\Big)P+\frac{(1-z_1)\xi'(z_2)}{X_0+1-\xi(z_2)}\partial_{z_1}w+\frac{w^2+2}{X_0}.
\end{aligned}
$$

It should be noted that in (2-26),

$$
\frac{w^2(X_0+1-\xi(z_2))}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}\cot\frac{z_2}{X_0}\quad\text{and}\quad\frac{1}{X_0}\cot\frac{z_2}{X_0}w
$$

are singular at $z_2=0$, and thus special care is required in our analysis.

In addition, it follows from the first equality and the fourth equality in (2-9) that

$$
\Big[\tfrac{1}{2}|U|^2+\frac{\gamma}{\gamma-1}\frac{P}{\rho}\Big]=0.
$$

This, together with the first and the fifth equation in (1-1) yields the Bernoulli's law

$$
\tfrac{1}{2}U_1^2(1+w^2)+\frac{\gamma}{\gamma-1}\frac{P}{\rho}=\tfrac{1}{2}(U_0^-(X_0))^2+\frac{\gamma}{\gamma-1}\frac{P_0^-(X_0)}{\rho_0^-(X_0)}. \tag{2-27}
$$

In terms of the fourth equation in (2-11), the equation for the entropy becomes

$$
\begin{aligned}
\Big(\Big(1+\frac{X_0w(1-z_1)\xi'(z_2)}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}\Big)\partial_{z_1}&\\
-\frac{X_0(X_0+1-\xi(z_2))w}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}\partial_{z_2}\Big)S&=0.
\end{aligned} \tag{2-28}
$$

The related boundary conditions of (S^+,P,U_1,U_2) are

$$
\begin{aligned}
&(S,P,U_1)(0,z_2)-(S_0^+,\hat{P}_0^+,U_0^+)(0)\\
&\qquad=(\tilde{g}_1,\tilde{g}_2,\tilde{g}_3)(U_2^2(0,z_2),P_0^-(\xi(z_2))-P_0^-(r_0),U_0^-(\xi(z_2))-U_0^-(r_0)).
\end{aligned} \tag{2-29}
$$

and

$$P(1, z_2) = P_e + \varepsilon \tilde{P}(z_2/X_0). \tag{2-30}$$

$$U_2(z_1, 1) = 0, \tag{2-31}$$

where the shock $\xi(z_2)$ is determined by

$$\xi'(z_2) = -\frac{\xi(z_2)}{X_0} \frac{(\rho U_1 U_2)(0, z_2)}{\rho(0, z_2) U_2^2(0, z_2) + P(0, z_2) - P_0^-(\xi(z_2))}. \tag{2-32}$$

Consequently, in order to show Theorem 1.1, we only need to solve the problem (2-26)–(2-28) with conditions (2-29)–(2-32).

3. The existence of a three-dimensional transonic shock for undetermined exit pressure

We will now establish the existence of a three-dimensional transonic shock in a nozzle when the transonic shock is assumed to go through some fixed point on the wall and when the end pressure $P_e + \varepsilon P_0(\theta)$ in (1-4) is adjusted by an appropriate constant. It follows from this that if one can show that the shock goes through some a point at the wall and if the corresponding adjustment constant on the end pressure is zero, then Theorem 1.1 will be proved.

Theorem 3.1. *Let the three-dimensional nozzle and the supersonic incoming flow be described as in Section 1. Assume further that*

$$\xi(1) = \tilde{r}_0, \tag{3-1}$$

where $\tilde{r}_0 \in (r_0 - \tilde{C} X_0^{3/2} \varepsilon, r_0 + \tilde{C} X_0^{3/2} \varepsilon)$ *with* $\tilde{C} > 0$ *some fixed constant. Then for* $\varepsilon < 1/X_0^3$ *and large* X_0, *there exists a constant* C_0 *such that the problem* (2-26)–(2-28) *and* (2-32) *with conditions* (2-29), (2-31) *and* (3-1) *has a* $C^{2,\alpha}(E_+)$ *transonic solution* $(S(z), P(z), U_1(z), U_2(z); \xi(z_2))$ *when* (2-30) *is replaced by*

$$P = \tilde{P}_e + \varepsilon \tilde{P}(z_2/X_0) + C_0 \quad \text{on } r = X_0 + 1. \tag{3-2}$$

Moreover,

$$\|\xi - \tilde{r}_0\|_{C^{3\alpha}[0,1]} \le C\varepsilon \tag{3-3}$$

and

$$\|(S, P, U_1) - (S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1))\|_{C^{2,\alpha}(E_+)} + \|U_2\|_{C^{2,\alpha}(E_+)} + |C_0| \le C\varepsilon. \tag{3-4}$$

Here C *is a generic nonnegative constant that is independent of* X_0 *and* ε, *and* $(S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1))$ *is the background solution representing a radially symmetric transonic shock at position* $\tilde{r}_0$ *with exit pressure* $\tilde{P}_e$.

Due to singular terms in (2-26) on $\{z_2 = 0\}$, special attention must be paid to handle the possible nearby singularities of the solution. Fortunately, this difficulty can be overcome and $C^{2,\alpha}$ regularity of the subsonic flow can be established.

We define iteration spaces as follows:

$$S_\sigma = \{\xi(z_2) \in C^{3,\alpha}[0,1] : \|\xi - \tilde{r}_0\|_{C^{3,\alpha}[0,1]} \le \sigma,\ \xi'(0) = \xi'(1) = 0,\ \xi^{(3)}(0) = 0\} \tag{3-5}$$

and

$$\begin{aligned} \Xi_\delta = \big\{(S, P, U_1, U_2) : \|(S, P, U_1, U_2) - (S_a^+, \hat{P}_a^+, \hat{U}_a^+, 0)\|_{C^{2,\alpha}(\overline{E_+})} \le \delta, \\ \partial_{z_2}(S, P, U_1)(z_1, 0) = \partial_{z_2}(S, P, U_1)(z_1, 1) = (0,0,0), \\ U_2(z_1, 0) = U_2(z_1, 1) = \partial_{z_2}^2 U_2(z_1, 0) = 0\big\}, \end{aligned} \tag{3-6}$$

with $\sigma > 0$ and $\delta > 0$ to be determined.

The proof of Theorem 3.1 is divided into four steps.

Step 1 (approximating shock). For $(\tilde{S}, P(q, \tilde{S}), V_1, V_2) \in \Xi_\delta$, we may by (2-32) define the approximating shock location as

$$\begin{aligned} \xi'(z_2) &= -\frac{\xi(z_2)}{X_0} \frac{(qV_1V_2)(0, z_2)}{P(q, \tilde{S})(0, z_2) - P_0^-(\xi(z_2)) + (qV_2^2)(0, z_2)}, \\ \xi(1) &= \tilde{r}_0, \end{aligned} \tag{3-7}$$

which has a unique solution $\xi(z_2) \in C^{3,\alpha}([0,1])$. It follows from the compatibility conditions in (3-6) that $\xi(z_2)$ satisfies at $z_2 = 0, 1$ the last two conditions in (3-5), and

$$\|\xi(z_2) - \tilde{r}_0\|_{C^{3,\alpha}} \le C\|V_2\|_{C^{2,\alpha}} \le C\delta. \tag{3-8}$$

In addition, as in (2-29), on $z_1 = \xi(z_2)$ we may require that

$$\begin{aligned} &(S, P, U_1)(0, z_2) - (S_a^+, \hat{P}_a^+(\tilde{r}_0), \hat{U}_a^+(\tilde{r}_0)) \\ &\qquad = (\check{g}_1, \check{g}_2, \check{g}_3)((V_2)^2, P_0^- - P_0^-(\tilde{r}_0), U_0^- - U_0^-(\tilde{r}_0)). \end{aligned} \tag{3-9}$$

It can be verified directly that $\partial_{z_2}(S, P, U_1)(0,0) = \partial_{z_2}(S, P, U_1)(0,1) = 0$.

Step 2 (approximating the specific entropy S). By (2-28), we approximate S by solving the problem

$$\begin{aligned} &\left(\left(V_1 + \frac{X_0(1 - z_1)\xi'(z_2)V_2}{\xi(z_2) + z_1(X_0 + 1 - \xi(z_2))}\right)\partial_{z_1} - \frac{X_0(X_0 + 1 - \xi(z_2))V_2}{\xi(z_2) + z_1(X_0 + 1 - \xi(z_2))}\partial_{z_2}\right) S = 0 \\ &\hspace{30em} \text{in } E_+, \\ &S_a^+ + \tilde{g}_1((V_2)^2(0, z_2), P_0^-(\xi(z_2)) - P_0^-(\tilde{r}_0), U_0^-(\xi(z_2)) - U_0^-(\tilde{r}_0)) = S \\ &\hspace{30em} \text{at } z_1 = 0. \end{aligned} \tag{3-10}$$

Due to (3-6), this problem has a unique solution $S \in C^{2,\alpha}(E_+)$. Moreover, by Remarks A.2 and A.3, we have

$$\begin{aligned}\|S - S_a^+\|_{C^{2,\alpha}} &\le C\|V_2\|^2_{C^{2,\alpha}} + \frac{C}{X_0}\|\xi - \tilde{r}_0\|_{C^{3,\alpha}} \\ &\le C\Big(\|V_2\|_{C^{2,\alpha}} + \frac{1}{X_0}\Big)\|V_2\|_{C^{2,\alpha}} \le C\Big(\delta + \frac{1}{X_0}\Big)\delta.\end{aligned} \tag{3-11}$$

Differentiating (3-10) with respect to z_2 and noting $\xi'(1) = V_2(z_1, 1) = 0$, we have

$$V_1\partial_{z_1}(\partial_{z_2}S) - \frac{X_0(X_0+1-\xi(z_2))\partial_{z_2}V}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}(\partial_{z_2}S) = 0 \text{ along } z_2 = 0 \text{ or } z_2 = 1,$$
$$\partial_{z_2}S(0,0) = \partial_{z_2}S(0,1) = 0,$$

which implies that

$$\partial_{z_2}S(z_1,0) = \partial_{z_2}S(z_1,1) = 0. \tag{3-12}$$

Thus, S belongs to Ξ_δ for small δ.

Convention 3.2. The reader may have noticed that X_0 sets the length scale for many quantities here. Since this trend will continue, we now declare that any symbol with check above it is that symbol divided by X_0. For example, $\check{z}_2 = z_2/X_2$, and $\check{1} = 1/X_0$.

Step 3 (approximating P and w). By (2-26), the second equality in (3-9) and (2-30)–(2-31), the approximate pressure P and w can be obtained from the boundary value problem

$$\begin{aligned} &\partial_1 w - \bar{a}_1\partial_2 P = F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2; \xi), \\ &\partial_2 w + \check{1}\cot\check{z}_2 w + \bar{a}_2\partial_1 P = F_2(\tilde{S}, P(q,\tilde{S}), V_1, V_2; \xi), \\ &P(0,z_2) - \hat{P}_a^+(\tilde{r}_0) \\ &\qquad = \tilde{g}_2(V_2^2(0,z_2), P_0^-(\xi(z_2)) - P_0^-(\tilde{r}_0), U_0^-(\xi(z_2)) - U_0^-(\tilde{r}_0)), \\ &P(1,z_2) = \tilde{P}_e + \varepsilon\tilde{P}(\check{z}_2) + C_0, \\ &w(z_1,0) = 0, \quad w(z_1,1) = 0. \end{aligned} \tag{3-13}$$

Here $\bar{a}_1$ and $\bar{a}_2$ are defined as a_1 and a_2 in (2-26), but with $(\hat{\rho}_0^+, \hat{U}_0^+, \hat{P}_0^+; r_0)$ replaced by $(\hat{\rho}_a^+, \hat{U}_a^+, \hat{P}_a^+; \tilde{r}_0)$. Note that the boundary condition $w(z_1,0) = 0$ comes essentially from requiring $C^{2,\alpha}$ regularity of the solution (P, w), by assuming $\tilde{P}'(0) = 0$ in (1-4). The constant C_0 will be chosen so that the solvability condition in (3-13) can be fulfilled. More concretely, it follows from the second equation in (3-13) and $w(z_1, 0) = 0$ that

$$w(z) = \frac{1}{\sin \check{z}_2} \int_0^{z_2} \sin \check{s}(F_2 - \bar{a}_2\partial_1 P)(z_1, s)ds.$$

Since $w(z_1, 1) = 0$, we have

$$\int_0^1 \sin \check{s}\big(F_2(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi) - \bar{a}_2\partial_1 P\big)(z_1, s)ds = 0.$$

In particular,

$$\int_0^1 \sin \check{s}\big(F_2(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi) - \bar{a}_2\partial_1 P\big)(1, s)ds = 0. \tag{3-14}$$

We will take this as the solvability condition of (3-13) that determines the unknown constant C_0.

Next, since $\hat{P}_a^+(z_1)$ satisfies

$$\bar{a}_2\partial_1 \hat{P}_a^+(z_1) - F_2(S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1), 0; \tilde{r}_0) = 0 \quad \text{in } E_+ \quad \text{and} \quad \hat{P}_a^+(1) = \tilde{P}_e,$$

a direct computation yields

$$\begin{aligned}
\partial_1 w - \bar{a}_1\partial_2(P - \hat{P}_a^+) &= F_1(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi),\\
\partial_2 w + \check{1}\cot \check{z}_2 w + \bar{a}_2\partial_1(P - \hat{P}_a^+) &= F_2(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi)\\
&\quad - F_2(S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1), 0; \tilde{r}_0),\\
(P - \hat{P}_a^+)(0, z_2) = \tilde{g}_2(V_2^2(0, z_2), &\, P_0^-(\xi(z_2)) - P_a^-(\tilde{r}_0), U_0^-(\xi(z_2)) - U_a^-(\tilde{r}_0)),\\
(P - \hat{P}_a^+)(1, z_2) = \varepsilon\tilde{P}(\check{z}_2) + C_0, &\\
w(z_1, 0) = 0, \quad w(z_1, 1) = 0. &
\end{aligned} \tag{3-15}$$

Next, we derive a second order elliptic equation for $P - \hat{P}_a^+$ from (3-15).

Applying ∂_{z_1} and $-(\partial_{z_2} + \check{1}\cot(\check{z}_2))$ to the first and second equation in (3-15) respectively and adding up yields

$$\begin{aligned}
&\partial_1\big(\bar{a}_2\partial_1(P - \hat{P}_a^+(z_1))\big) + \partial_2\big(\bar{a}_1\partial_2(P - \hat{P}_a^+(z_1))\big) + \check{a}_1\cot \check{z}_2\,\partial_2\big(P - \hat{P}_a^+(z_1)\big)\\
&\qquad = \partial_1\big(F_2(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi) - F_2(S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1), 0; \tilde{r}_0)\big)\\
&\qquad\quad - \partial_2\big(F_1(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi)\big) - \check{1}\cot \check{z}_2 F_1(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi) \quad \text{in } E_+,\\
&(P - \hat{P}_a^+)(0, z_2) = \tilde{g}_2\big(V_2^2(0, z_2), P_0^-(\xi(z_2)) - P_0^-(\tilde{r}_0), U_0^-(\xi(z_2)) - U_0^-(\tilde{r}_0)\big),\\
&(P - \hat{P}_a^+)(1, z_2) = \varepsilon\tilde{P}(\check{z}_2) + C_0,\\
&\partial_2(P - \hat{P}_a^+(z_1)) = 0 \quad \text{on } z_2 = 0 \text{ or } z_2 = 1,
\end{aligned} \tag{3-16}$$

where the fact that $\partial_{z_2}(P-\hat{P}_a^+)(z_1,0)=\partial_{z_2}(P-\hat{P}_a^+)(z_1,1)=0$ comes from (3-15) and $F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2; \xi)(z_1,0) = F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2; \xi)(z_1,1) = 0$.

We now decompose the problem (3-16) as $P(z) = P_1(z) + P_2(z)$, with

$$\begin{aligned}
&\partial_1(\bar{a}_2\partial_1(P_1-\hat{P}_a^+(z_1)))+\partial_2(\bar{a}_1\partial_2(P_1-\hat{P}_a^+(z_1)))+\check{\bar{a}}_1\cot\check{z}_2\partial_2(P_1-\hat{P}_a^+(z_1))\\
&\qquad=\partial_1\big(F_2(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)-F_2(S_a^+,\hat{P}_a^+(z_1),\hat{U}_a^+(z_1),0;\tilde{r}_0)\big)\\
&\qquad\qquad\qquad-\partial_2\big(F_1(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)\big)-\check{1}\cot\check{z}_2F_1(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi),\\
&P_1(0,z_2)-\hat{P}_a^+(0)=\tilde{g}_2(V_2^2(0,z_2),P_0^-(\xi(z_2))-P_a^-(\tilde{r}_0),U_0^-(\xi(z_2))-U_a^-(\tilde{r}_0)),\\
&P_1(1,z_2)-\hat{P}_a^+(1)=\varepsilon\tilde{P}(\check{z}_2),\\
&\partial_2(P_1-\hat{P}_a^+(z_1))=0\quad\text{on } z_2=0 \text{ or } z_2=1,
\end{aligned}\tag{3-17}$$

and

$$\begin{aligned}
&\bar{a}_2\partial_1^2P_2+\bar{a}_1\partial_2^2P_2+\check{\bar{a}}_1\cot\check{z}_2\partial_2P_2=0\quad\text{in } E_+,\\
&P_2(0,z_2)=0,\\
&P_2(1,z_2)=C_0,\\
&\quad\partial_2P_2=0\quad\text{on } z_2=0 \text{ or } z_2=1.
\end{aligned}\tag{3-18}$$

We first treat the problem (3-17).

It follows from Lemma B.5 (for the case of $k=1$) that (3-17) has a unique $C^{2,\alpha}(E_+)$ solution $P_1(z)$ satisfying

$$\begin{aligned}
&\|P_1(z)-\hat{P}_a^+(z_1)\|_{C^{2,\alpha}}\\
&\qquad\le C\|F_2(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)-F_2(S_a^+,\hat{P}_a^+(z_1),\hat{U}_a^+(z_1),0;\tilde{r}_0)\|_{C^{1,\alpha}}\\
&\qquad\quad+C\|F_1(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)\|_{C^{1,\alpha}}+C\varepsilon\|\tilde{P}(\check{z}_2)\|_{C^{2,\alpha}}\\
&\qquad\quad+C\|\tilde{g}_2(V_2^2(0,z_2),P_0^-(\xi(z_2))-P_0^-(\tilde{r}_0),U_0^-(\xi(z_2))-U_0^-(\tilde{r}_0))\|_{C^{2,\alpha}}.
\end{aligned}$$

Though $(V_2^2(X_0+1-\xi(z_2)))/(\xi(z_2)+z_1(X_0+1-\xi(z_2)))\cot\check{z}_2$ may be singular in F_1, it follows from Lemma B.3 that

$$\begin{aligned}
\left\|\frac{V_2^2(X_0+1-\xi(z_2))}{\xi(z_2)+z_1(X_0+1-\xi(z_2))}\cot\check{z}_2\right\|_{C^{1,\alpha}}&\le C\|V_2\|_{C^{1,\alpha}}\left\|\check{1}\cot\check{z}_2V_2\right\|_{C^{1,\alpha}(E_+)}\\
&\le C\delta\|V_2\|_{C^{2,\alpha}}.
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\|P_1(z)-\hat{P}_a^+(z_1)\|_{C^{2,\alpha}}\\
&\quad\le O(\check{1})\|\tilde{S}-S_a^+\|_{C^{2,\alpha}}+O(\check{1})\|P(q,\tilde{S})-\hat{P}_a^+\|_{C^{2,\alpha}}\\
&\qquad+O(\check{1}+\delta)\|V_1-\hat{U}_a^+\|_{C^{2,\alpha}}+O(\check{1}+\delta+\varepsilon)\|V_2\|_{C^{2,\alpha}}\\
&\qquad\qquad+O(\check{1}+\delta)\|\xi-\tilde{r}_0\|_{C^{2,\alpha}}+O(\varepsilon)\\
&\quad\le C(\check{\delta}+\delta^2+\varepsilon).
\end{aligned}
\tag{3-19}
$$

Next, note that the problem (3-18) has a solution

$$
P_2(z)=C_0z_1,
\tag{3-20}
$$

which is unique by Lemma B.5.

In this case, by the second equation in (3-15), (3-14) can be changed into

$$
\begin{aligned}
\int_0^1 \sin\check{s}\big(&F_2(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)-F_2(S_a^+,\hat{P}_a^+(z_1),\hat{U}_a^+(z_1),0;\tilde{r}_0)\\
&-\bar{a}_2(\partial_1P_1-\partial_1\hat{P}_a^+(z_1))-\bar{a}_2C_0\big)(1,s)ds=0.
\end{aligned}
\tag{3-21}
$$

Note that $\bar{a}_2=O(1)>0$ since $(S_a^+,\hat{P}_a^+,\hat{U}_a^+)$ is subsonic. Then we can choose a unique constant C_0 such that (3-21) holds. Moreover, it follows from (3-19) and the expression of F_2 that C_0 admits the estimate

$$
\begin{aligned}
&|C_0|\\
&=\frac{1}{2\bar{a}_2X_0\sin^2\frac{1}{2X_0}}\bigg|\int_0^1\sin\check{s}\big(F_2(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)\\
&\qquad\qquad-F_2(S_a^+,\hat{P}_a^+(z_1),\hat{U}_a^+(z_1),0;\tilde{r}_0)\\
&\qquad\qquad\qquad-\bar{a}_2(\partial_1P_1-\partial_1\hat{P}_a^+(z_1))\big)(1,s)ds\bigg|\\
&\le\|P_1(z)-\hat{P}_a^+(z_1)\|_{C^{2,\alpha}}\\
&\qquad+\frac{1}{\bar{a}_2}\big\|F_2(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)-F_2(S_a^+,\hat{P}_a^+(z_1),\hat{U}_a^+(z_1),0;\tilde{r}_0)\big\|_{C^{1,\alpha}}\\
&\le C(\check{\delta}+\delta^2+\varepsilon).
\end{aligned}
\tag{3-22}
$$

Collecting all the estimates (3-17)–(3-22) shows that there exists a unique constant C_0 such that the second order elliptic equation (3-16) with mixed boundary conditions has a unique solution $P(z)$ satisfying

$$
\|P-\hat{P}_a^+\|_{C^{2,\alpha}}+|C_0|\le\|P_1-\hat{P}_a^+\|_{C^{2,\alpha}}+C|C_0|\le C(\check{\delta}+\delta^2+\varepsilon).
\tag{3-23}
$$

With $P(z)$ so determined, we can obtain w in E_+ by solving the problem

$$\begin{aligned} &\partial_1 w = \bar{a}_1 \partial_2 P + F_1(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi), \\ &\partial_2 w + \check{1} \cot \check{z}_2 w = F_2(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi) - \bar{a}_2 \partial_1 P, \\ &w(z_1, 0) = 0. \end{aligned} \tag{3-24}$$

It follows from Lemma B.7 that (3-24) has a unique solution w due to (3-13). On the other hand, by $w(z_1, 0) = 0$, we arrive at

$$\|w\|_{C^{2,\alpha}} \leq C(\|\partial_1 w\|_{C^{1,\alpha}} + \|\partial_2 w\|_{C^{1,\alpha}}). \tag{3-25}$$

We now estimate $\|\partial_1 w\|_{C^{1,\alpha}(E_+)}$ and $\|\partial_2 w\|_{C^{1,\alpha}(E_+)}$.

By the first equation in (3-15) and (3-23), we have

$$\|\partial_1 w\|_{C^{1,\alpha}} \leq C\big(\|P - \hat{P}_a^+\|_{C^{2,\alpha}} + \|F_1\|_{C^{1,\alpha}}\big) \leq C(\check{\delta} + \delta^2 + \varepsilon). \tag{3-26}$$

Next, it follows from the second equation in (3-15) that

$$\begin{aligned} w(z) = \frac{1}{\sin \check{z}_2} \int_0^{z_2} \sin \check{s}\big(&F_2(\tilde{S}, P(q, \tilde{S}), V_1, V_2; \xi) \\ &- F_2(S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1), 0; \tilde{r}_0) \\ &- \bar{a}_2(\partial_1 P - \partial_1 \hat{P}_a^+(z_1))\big) ds. \end{aligned} \tag{3-27}$$

Furthermore, a direct but careful computation using (3-27) and (3-21) yields

$$w(z_1, 0) = \partial_{z_2}^2 w(z_1, 0) = w(1, 1) = 0. \tag{3-28}$$

Indeed, $w(z_1, 0) = w(1, 1) = 0$ comes directly from (3-21), (3-24) and (3-27), while $\partial_{z_2}^2 w(z_1, 0) = 0$ follows from the following computations:

Applying ∂_{z_2} two both sides of the second equation in (3-24) yields

$$\partial_{z_2}^2 w + \check{1} \cot \check{z}_2 \partial_{z_2} w - \frac{1}{X_0^2 \sin^2 \check{z}_2} w = \partial_{z_2} F_2 - \bar{a}_2 \partial_{z_1 z_2}^2 P. \tag{3-29}$$

Note that for small z_2,

$$\begin{aligned} &\partial_{z_2}^2 w + \check{1} \cot \check{z}_2 \partial_{z_2} w - \frac{1}{X_0^2 \sin^2 \check{z}_2} w \\ &\quad = \partial_{z_2}^2 w + \frac{1}{X_0^2 \sin^2 \check{z}_2} (\partial_{z_2} w X_0 \sin \check{z}_2 \cos \check{z}_2 - w) \\ &\quad = \partial_{z_2}^2 w + \frac{1}{X_0^2 \sin^2 \check{z}_2} \Big(\partial_{z_2} w X_0 (\check{z}_2 + o(\check{z}_2^2))\big(1 - \tfrac{1}{2} \check{z}_2^2 + o(\check{z}_2^3)\big) \\ &\qquad\qquad\qquad\qquad\qquad\qquad - z_2 \int_0^1 \partial_{z_2} w(z_1, \theta z_2) d\theta\Big) \\ &\quad = \partial_{z_2}^2 w + \frac{1}{X_0^2 \sin^2 \check{z}_2} \big(\partial_{z_2} w z_2 - \partial_{z_2} w(z_1, 0) z_2 - \tfrac{1}{2} \partial_{z_2}^2 w(z_1, 0) z_2^2 + o(z_2^2)\big) \end{aligned}$$

$$= \tfrac{3}{2}\partial^2_{z_2} w(z_1,0) + o(z_2),$$

and it follows from $\partial_{z_2} P(z_1,0)=0$ and the expression of F_2 that $\partial^2_{z_1z_2}P(z_1,0)=0$ and $\partial_{z_2}F_2(z_1,0)=0$. Consequently, (3-29) shows that $\partial^2_{z_2}w(z_1,0)=0$.

In addition, because $\partial_{z_1}w(z_1,1)=0$, which comes from $\partial_{z_2}P(z_1,1)=0$ and $F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2; \xi)(z_1,1)=0$, and $w(1,1)=0$, we have

$$w(z_1,1)=0. \tag{3-30}$$

Finally, it follows from the second equation in (3-15) and Lemma B.6 that

$$\begin{aligned}
&\|\partial_2 w\|_{C^\alpha} + \|\partial_2^2 w\|_{C^\alpha} \\
&\quad \le C\big(\|F_2(\tilde{S}, P(q,\tilde{S}), V_1, V_2; \xi) - F_2(S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1), 0; \tilde{r}_0)\|_{C^{1,\alpha}} \\
&\qquad\qquad + \|P - \hat{P}_a^+(z_1)\|_{C^{2,\alpha}}\big) \\
&\quad \le C(\check{\delta} + \delta^2 + \varepsilon).
\end{aligned}$$

This, together with (3-26), yields

$$\begin{aligned}
&\|w\|_{C^{2,\alpha}} \\
&\quad \le C\big(\|F_2(\tilde{S}, P(q,\tilde{S}), V_1, V_2; \xi) - F_2(S_a^+, \hat{P}_a^+(z_1), \hat{U}_a^+(z_1), 0; \tilde{r}_0)\|_{C^{1,\alpha}} \\
&\qquad\qquad + \|P - \hat{P}_a^+\|_{C^{2,\alpha}} + \|F_1\|_{C^{1,\alpha}}\big) \\
&\quad \le C(\check{\delta} + \delta^2 + \varepsilon).
\end{aligned} \tag{3-31}$$

Thus, it follows from (3-16), (3-22)–(3-24), (3-28), (3-30) and (3-31) that there exists a unique constant C_0 such that the first order elliptic system (3-13) has a unique solution $(P(z), w(z))$ satisfying the estimates

$$\|P - \hat{P}_0^+\|_{C^{2,\alpha}} + \|w\|_{C^{2,\alpha}} + |C_0| \le C(\check{\delta} + \delta^2 + \varepsilon). \tag{3-32}$$

and

$$\partial_2 P(z_1,0) = \partial_2 P(z_1,1) = w(z_1,0) = w(z_1,1) = \partial_2^2 w(z_1,0) = 0. \tag{3-33}$$

Step 4 (approximating the radial velocity U_1). Due to (2-27), the radial velocity U_1 can be uniquely determined from

$$\begin{aligned}
&U_1^2(1+w^2) + \frac{2\gamma}{\gamma-1}\frac{P}{\rho} - \big(\hat{U}_a^+\big)^2 - \frac{2\gamma}{\gamma-1}\frac{\hat{P}_a^+}{\hat{\rho}_a^+} = 0, \\
&U_1(z) > 0.
\end{aligned} \tag{3-34}$$

It follows from (3-11) and (3-32) that $U_1(z)$ satisfies

$$\begin{aligned}
\|U_1 - \hat{U}_a^+\|_{C^{2,\alpha}} &\le C\big(\delta\|w\|_{C^{2,\alpha}} + \|S - S_a^+\|_{C^{2,\alpha}} + \|P - \hat{P}_a^+\|_{C^{2,\alpha}}\big) \\
&\le C(\check{\delta} + \delta^2 + \varepsilon).
\end{aligned} \tag{3-35}$$

By (3-12), (3-28) and (3-30), a direct computation yields

$$\partial_{z_2} U_1(z_1, 0) = \partial_{z_2} U_1(z_1, 1) = 0. \tag{3-36}$$

All the constants C in (3-8), (3-11), (3-32) and (3-35) depend only on the supersonic incoming flow and $\|\tilde{P}(\check{z}_2)\|_{C^{2,\alpha}}$, so we can choose $\sigma = O(1)\varepsilon > 0$ and $\delta = O(1)\varepsilon > 0$ such that $(S, P, U_1, U_2; \xi)$ obtained in Steps 1–4 belongs to the space Ξ_δ. Consequently, we can define a map T from Ξ_δ to itself by

$$T(\tilde{S}, P(q, \tilde{S}), V_1, V_2) = (S, P, U_1, U_2). \tag{3-37}$$

Proof of Theorem 3.1. It suffices to prove that the mapping T defined in (3-37) is contractible in $C^{1,\alpha}(E_+)$.

For any two given elements $(\tilde{S}_1, \tilde{P}_1, V_{11}, V_{21})$ and $(\tilde{S}_2, \tilde{P}_2, V_{12}, V_{22})$ in Ξ_δ, set

$$\begin{aligned} T(\tilde{S}_1, \tilde{P}_1, V_{11}, V_{21}) &= (S_1, P_1, U_{11}, U_{21}),\\ T(\tilde{S}_2, \tilde{P}_2, V_{12}, V_{22}) &= (S_2, P_2, U_{12}, U_{22}), \end{aligned}$$

and denote the corresponding approximating shocks (obtained by solving (3-7)) by $\xi_1(z_2)$ and $\xi_2(z_2)$, respectively. Below we will use the fact that $\sigma = O(1)\varepsilon > 0$ and $\delta = O(1)\varepsilon > 0$ in (3-5) and (3-6).

Define

$$\begin{aligned} (W_1, W_2, W_3, W_4) &= (S_1 - S_2, P_1 - P_2, U_{11} - U_{12}, U_{21} - U_{22}),\\ (\widetilde{W}_1, \widetilde{W}_2, \widetilde{W}_3, \widetilde{W}_4) &= (\widetilde{S}_1 - \widetilde{S}_2, \widetilde{P}_1 - \widetilde{P}_2, V_{11} - V_{12}, V_{21} - V_{22}). \end{aligned}$$

For convenience, we set also

$$W_5 = \frac{U_{21}}{U_{11}} - \frac{U_{22}}{U_{12}}, \quad \widetilde{W}_5 = \frac{V_{21}}{V_{11}} - \frac{V_{22}}{V_{12}}, \quad W_6 = \xi_1(z_2) - \xi_2(z_2).$$

Next, we derive some useful estimates on W_i for $i = 1, 2, \cdots, 6$, so that the contractible property of T can be established.

First, it follows from (3-7) and a simple computation that

$$\begin{aligned} W_6'(z_2) = O(\varepsilon)\widetilde{W}_1 + O(\varepsilon)\widetilde{W}_2 + O(\varepsilon)\widetilde{W}_3 &\\ + O(1)\widetilde{W}_4 + O(\check{\varepsilon})W_6 &\quad \text{in } [0, 1],\\ W_6(1) = 0. & \end{aligned} \tag{3-38}$$

This yields

$$\|W_6\|_{C^{2,\alpha}[0,1]} \le C\Big(\varepsilon \sum_{i=1}^{3} \|\widetilde{W}_i\|_{C^{1,\alpha}} + \|\widetilde{W}_4\|_{C^{1,\alpha}}\Big). \tag{3-39}$$

Second, it follows from (2-28) and Lemma B.8 that

$$\|W_1\|_{C^{1,\alpha}} \le C\Big(\varepsilon \sum_{i=2}^{4} \|\widetilde{W}_i\|_{C^{1,\alpha}} + \check{1}\|W_6\|_{C^{2,\alpha}}\Big). \tag{3-40}$$

Next, it follows from (3-13) and (3-21) that

$$\begin{aligned}
\partial_1 W_5 - \bar{a}_1\partial_2 W_2 &= O(\varepsilon)\widetilde{W}_1 + O(\varepsilon)\widetilde{W}_2 + O(\varepsilon)\widetilde{W}_3\\
&\qquad + O(\check{1})\widetilde{W}_5 + O(\varepsilon)W_6 + O(\varepsilon)\partial_1\widetilde{W}_2\\
&\qquad\qquad + O(\check{1})\partial_2\widetilde{W}_2 + O(\check{1})W_6'(z_2),\\
\partial_2 W_5 + \check{1}\cot(\check{z}_2)W_5 &+ \bar{a}_2\partial_1 W_2\\
&= O(\check{1})\widetilde{W}_1 + O(\check{1})\widetilde{W}_2 + O(\check{1})\widetilde{W}_3\\
&\qquad + O(\varepsilon)\widetilde{W}_5 + O(\check{1})W_6 + O(\check{1})\partial_1\widetilde{W}_2\\
&\qquad + O(\varepsilon)\partial_2\widetilde{W}_2 + O(\varepsilon)\partial_1\widetilde{W}_5 + O(\varepsilon)W_6'(z_2),\\
W_2(0, z_2) &= O(\varepsilon)\widetilde{W}_4(0, z_2) + O(\check{1})W_6(z_2),\\
W_2(1, z_2) &= \text{constant},\\
W_5(z_1, 0) &= 0, \quad W_5(z_1, 1) = 0.
\end{aligned} \tag{3-41}$$

Then it follows from Lemma B.5 for the case $k = 0$ and (B-31) of Lemma B.6 that

$$\|W_2\|_{C^{1,\alpha}} + \|W_5\|_{C^{1,\alpha}} + |\text{constant}| \le \check{C}\Big(\sum_{i=1}^{5} \|\widetilde{W}_i\|_{C^{1,\alpha}} + \|W_6\|_{C^{2,\alpha}}\Big). \tag{3-42}$$

Finally, it follows from the algebraic equation (2-27) that

$$W_3 = O(1)W_1 + O(1)W_2 + O(\varepsilon)W_5. \tag{3-43}$$

This yields

$$\|W_3\|_{C^{1,\alpha}} \le C(\|W_1\|_{C^{1,\alpha}} + \|W_2\|_{C^{1,\alpha}} + \varepsilon\|W_5\|_{C^{1,\alpha}}). \tag{3-44}$$

Collecting all the estimates (3-39), (3-40), (3-42) and (3-44) obtained thus far, we arrive at

$$\sum_{i=1}^{3} \|W_i\|_{C^{1,\alpha}} + \|W_5\|_{C^{1,\alpha}} \le C(\check{1} + \varepsilon)\sum_{j=1}^{5} \|\widetilde{W}_i\|_{C^{1,\alpha}}. \tag{3-45}$$

In terms of the definitions of W_4, W_5, $\widetilde{W}_4$ and $\widetilde{W}_5$, one deduces from (3-45) that

$$\sum_{i=1}^{4} \|W_i\|_{C^{1,\alpha}} \le C(\check{1} + \varepsilon)\sum_{j=1}^{4} \|\widetilde{W}_i\|_{C^{1,\alpha}}. \tag{3-46}$$

Since X_0 is large and ε is small, $C(\check{1}+\varepsilon) < 1$ holds true in (3-46). This implies that the mapping T from Ξ_δ into itself is contractible in $C^{1,\alpha}(E_+)$. Therefore, it follows from the contractible mapping theorem that there exists a unique fixed point of T in the function space Ξ_δ, which completes the proof of Theorem 3.1. □

We complete this section by pointing out some refined estimates on the solution obtained in Theorem 3.1. First, we note that by some elementary analysis for ordinary differential systems, one can verify the following fact, which has been given in [Li et al. 2009b, Proposition 5.3]:

Suppose $(S^+_{0,1}, \hat{P}^+_{0,1}(r), \hat{U}^+_{0,1}(r)$ *and* $(S^+_{0,2}, \hat{P}^+_{0,2}(r), \hat{U}^+_{0,2}(r))$, *with* $r \in [X_0, X_0+1]$ *given in Remark A.3, are two extended subsonic flows that correspond to the shock positions* $r_{0,1}$ *and* $r_{0,2}$ *with* $r_{0,i} \in (X_0, X_0+1)$, *and constant end pressures* $P_{1,e}$ *and* $P_{2,e}$ *respectively. Then there exists a uniform constant* $C > 1$ *independent of* X_0 *such that for large* X_0

$$\begin{aligned} &\|(S^+_{0,1}, \hat{P}^+_{0,2}(r), \hat{U}^+_{0,2}(r)) - (S^+_{0,2}, \hat{P}^+_{0,1}(r), \hat{U}^+_{0,1}(r))\|_{C^{4,\alpha}[X_0,X_0+1]} \\ &\qquad\qquad \le C|P_{2,e} - P_{1,e}|, \\ &(X_0/C)|P_{2,e} - P_{1,e}| \le |r_{0,2} - r_{0,1}| \le CX_0|P_{2,e} - P_{1,e}|. \end{aligned} \tag{3-47}$$

This result combines with Theorem 3.1 to give another:

Theorem 3.1′. *Under the assumptions of Theorem 3.1, we have*

$$\|\xi - r_0\|_{L^\infty[0,1]} \le CX_0^{3/2}\varepsilon, \qquad \|\xi'\|_{C^{2,\alpha}[0,1]} \le C\varepsilon \tag{3-48}$$

and

$$\|(S, P, U_1) - (S^+_0, \hat{P}^+_0(z_1), \hat{U}^+_0(z_1))\|_{C^{2,\alpha}(E_+)} + |C_0| \le C\sqrt{X_0}\varepsilon, \tag{3-49}$$

$$\begin{aligned} \|\partial_{z_2}(S, P, U_1) - \partial_{z_2}(S^+_0, \hat{P}^+_0(z_1), \hat{U}^+_0(z_1))\|_{C^{1,\alpha}(E_+)} + \|U_2\|_{C^{2,\alpha}(E_+)} \\ \le C\varepsilon. \end{aligned} \tag{3-50}$$

Here the generic constant $C > 0$ *is independent of* X_0 *and* ε, *but may depend on* $\tilde{C}$.

Remark 3.3. In Theorems 3.1 and 3.1′ or the problem (1-1) with (1-2)–(1-5), it seems difficult to find higher order compatibility conditions at the nozzle wall so that the solutions will achieve $C^{3,\alpha}$ regularity; this is due to the source terms in (2-8). For more details, see Appendix C.

4. The monotonic dependence of the shock position on the exit pressure

The key to proving Theorem 1.1, as in [Li et al. 2009b], establishing the monotonic dependence of the shock position on the end pressure. For this end, we assume that the problem (2-26)–(2-28), (2-32) with (2-29) and (2-31), has two solutions

$$(S, P, U_1, U_2; \xi_1) \in C^{2,\alpha}(E_+) \times C^{3,\alpha}([0, 1]),$$
$$(\tilde{S}, \tilde{P}, V_1, V_2; \xi_2) \in C^{2,\alpha}(E_+) \times C^{3,\alpha}([0, 1])$$

when the exit pressure boundary condition (2-30) is replaced respectively by

$$P(1, z_2) = P_e + \varepsilon \tilde{P}_1(\check{z}_2), \tag{4-1}$$

$$\tilde{P}(1, z_2) = P_e + \varepsilon \tilde{P}_2(\check{z}_2). \tag{4-2}$$

Theorem 4.1. *If* $(P, \rho, U_1, U_2, S; \xi_1)$ *and* $(\tilde{P}, q, V_1, V_2, \tilde{S}; \xi_2)$ *both satisfy the estimates* (3-48)–(3-50), *and*

$$M_0^-(X_0) \equiv \frac{U_0^-(X_0)}{c(\rho_0^-(X_0))} > \sqrt{\frac{\gamma+3}{2}},$$

then

$$|\xi_2(1) - \xi_1(1)| \le C X_0 \varepsilon \| \tilde{P}_1(\check{z}_2) - \tilde{P}_2(\check{z}_2) \|_{C^{1,\alpha}[0,1]}, \tag{4-3}$$

and

$$\begin{aligned} \|(S, P, U_1, U_2) - (\tilde{S}, \tilde{P}, V_1, V_2)\|_{C^{1,\alpha}(E_+)} + \|\xi_1' - \xi_2'\|_{C^{1,\alpha}[0,1]} \\ \le C\varepsilon \| \tilde{P}_1(\check{z}_2) - \tilde{P}_2(\check{z}_2) \|_{C^{1,\alpha}[0,1]}. \end{aligned} \tag{4-4}$$

Furthermore, if $P_1(1, z_2) - P_2(1, z_2) = \tilde{C} = O(\sqrt{X_0}\varepsilon)$ *and* $\xi_1(1) < \xi_2(1)$, *then* $\xi_1(z_2) < \xi_2(z_2)$ *and the constant* $\tilde{C}$ *is positive. Moreover, there exists a generic constant* $C > 1$ *such that*

$$\frac{\check{1}}{C}(\xi_2(1) - \xi_1(1)) \le \tilde{C} \le \check{C}(\xi_2(1) - \xi_1(1)). \tag{4-5}$$

Proof. Without loss of generality, we assume

$$\xi_1(1) < \xi_2(1). \tag{4-6}$$

With some abuse of notation, we set

$$W_1(z) = S - \tilde{S}, \qquad W_2(z) = P - \tilde{P}, \qquad W_3(z) = U_1 - V_1,$$
$$W_4(z) = U_2 - V_2, \quad W_5(z) = \frac{U_2}{U_1} - \frac{V_2}{V_1}, \quad W_6(z_2) = \xi_1 - \xi_2.$$

The proof of Theorem 4.1 will be divided into five steps.

Step i (the estimate of W_6). It follows from (2-32) that $W_6(z_2)$ satisfies

$$\begin{aligned} W_6'(z_2) &= \sum_{i=1}^{3} O(\varepsilon)W_i + O(1)W_4 + O(\check{\varepsilon})W_6, \\ W_6(1) &= \xi_1(1) - \xi_2(1) \end{aligned} \tag{4-7}$$

and

$$\begin{aligned} W_6''(z_2) = \sum_{i=1}^{4} O(\varepsilon)W_i + O(\check{\varepsilon})W_6 + \sum_{i=1}^{3} O(\varepsilon)\partial_2 W_i& \\ + O(1)\partial_2 W_4 + O(\check{\varepsilon})W_6'(z_2)&, \\ W_6'(1) = 0.& \end{aligned} \tag{4-8}$$

By (4-6), we have

$$\|W_6'(z_2)\|_{C^{1,\alpha}} \le C(\varepsilon(\xi_2(1) - \xi_1(1)) + \|\partial_2 W_4\|_{C^\alpha}) + C\varepsilon\Big(\sum_{i=1}^{4}\|W_i\|_{C^{1,\alpha}}\Big) \tag{4-9}$$

and

$$\begin{aligned} \|W_6\|_{C^{2,\alpha}} &\le C((\xi_2(1) - \xi_1(1)) + \|W_6'(z_2)\|_{C^{1,\alpha}}) \\ &\le C((\xi_2(1) - \xi_1(1)) + \|\partial_2 W_4\|_{C^\alpha}) + C\varepsilon\Big(\sum_{i=1}^{4}\|W_i\|_{C^{1,\alpha}}\Big). \end{aligned} \tag{4-10}$$

Step ii (the estimate of W_1). First, we solve the first order system (2-28) in the coordinate $z = (z_1, z_2)$. Let $z_2^1(s; z)(z_2^2(s; z))$ be the characteristic going through $z = (z_1, z_2)$ and reaching $(0, \beta)((0, \tilde{\beta}))$ at $s = 0$ corresponding to the vector field $(U_1, U_2)((V_1, V_2))$, that is,

$$\begin{aligned} &\frac{dz_2^1(s; z)}{ds} = -\frac{X_0(X_0 + 1 - \xi_1(z_2^1))}{A_1} U_2(\xi_1(z_2^1) + s(X_0 + 1 - \xi_1(z_2^1)), z_2^1), \\ &z_2^1(z_1; z) = z_2, \quad z_2^1(0; z) = \beta, \end{aligned}$$

where

$$A_1 = (\xi_1(z_2^1) + s(X_0 + 1 - \xi_1(z_2^1)))U_1 + U_2 X_0(1 - s)\xi_1'(z_2^1).$$

Set $l(s; z) = z_2^1(s; z) - z_2^2(s; z)$, and note that $z_2^1(0; z) = \beta$ and $z_2^2(0; z) = \tilde{\beta}$. Then we have

$$\begin{aligned} \frac{dl}{ds} = O(\varepsilon)l + O(\varepsilon)W_3(s, z_2^1) + O(1)W_4(s, z_2^1)& \\ + O(\varepsilon)W_6(z_2^1) + O(\varepsilon^2)W_6'(z_2^1)& \\ l(0; z) = \beta - \tilde{\beta}, \quad l(z_1; z) = 0.& \end{aligned} \tag{4-11}$$

By the $C^{2,\alpha}$ regularity of solutions, we can check that the coefficients of $l(t;z)$ in (4-11) are in $C^{1,\alpha}$. Based on this, we intend to derive the $C^{1,\alpha}$ estimate of $\beta-\tilde{\beta}$.

Indeed, by (4-11), we can arrive at

$$\|\beta-\tilde{\beta}\|_{L^\infty}\le C(\varepsilon\|W_3\|_{L^\infty}+\|W_4\|_{L^\infty}+\varepsilon\|W_6\|_{L^\infty}+\varepsilon^2\|W_6'(z_2)\|_{L^\infty}).$$

On the other hand,

$$z_2^1(s;z)=-\int_0^s\frac{X_0(X_0+1-\xi_1(z_2^1))}{A_1}U_2(\xi_1(z_2^1)+t(X_0+1-\xi_1(z_2^1)),z_2^1)dt+\beta,$$

and

$$z_2=-\int_0^{z_1}\frac{X_0(X_0+1-\xi_1(z_{21}))}{A_1}U_2(\xi_1(z_2^1)+t(X_0+1-\xi_1(z_{21})),z_2^1)dt+\beta.$$

Similar relations hold for $z_2^2(s;z)$, z_2, and $\tilde{\beta}$ corresponding to (V_1,V_2).

Hence, one can obtain

$$\begin{aligned}\beta-\tilde{\beta}&=\int_0^{z_1}(O(\varepsilon)W_3(t,z_2^1)+O(1)W_4(t,z_2^1)\\&\qquad\qquad+O(\varepsilon)W_6(z_2^1)+O(\varepsilon^2)W_6'(z_2^1)+O(\varepsilon)l(t;z))dt,\\ l(s;z)&=\int_{z_1}^{s}(O(\varepsilon)W_3(t,z_2^1)+O(1)W_4(t,z_2^1)\\&\qquad\qquad+O(\varepsilon)W_6(z_2^1)+O(\varepsilon^2)W_6'(z_2^1)+O(\varepsilon)l(t;z))dt\end{aligned}\tag{4-12}$$

and

$$\|\partial_{z_1}(\beta,\tilde{\beta})\|_{C^{1,\alpha}}\le C\varepsilon,\quad \|\partial_{z_2}(\beta,\tilde{\beta})\|_{C^{1,\alpha}}\le C.\tag{4-13}$$

It follows from (4-12) and (4-13) that

$$\|\beta-\tilde{\beta}\|_{C^{1,\alpha}}\le C(\varepsilon\|W_3\|_{C^{1,\alpha}}+\|W_4\|_{C^{1,\alpha}}+\varepsilon\|W_6\|_{C^{2,\alpha}}).\tag{4-14}$$

In addition, by (2-28) and the characteristics method, we have

$$\begin{aligned}W_1(z)&=W_1(0,\beta(z_1,z_2))+O(\varepsilon)\big(\beta(z_1,z_2)-\tilde{\beta}(z_1,z_2)\big),\\ W_1(0,z_2)&=O(\varepsilon)W_4(0,z_2)+O(\check{1})W_6(z_2).\end{aligned}\tag{4-15}$$

Combining (4-15) with (4-14) yields

$$\begin{aligned}\|W_1\|_{C^{1,\alpha}}&\le C\big(\varepsilon\|(\varepsilon W_2,\varepsilon W_3,W_4)\|_{C^{1,\alpha}}+\check{1}\|W_6\|_{C^{2,\alpha}}+\varepsilon\|\beta-\tilde{\beta}\|_{C^{1,\alpha}}\big)\\&\le C\big(\check{1}(\xi_2(1)-\xi_1(1))+\varepsilon\|(\varepsilon W_2,W_3,W_4)\|_{C^{1,\alpha}}\\&\qquad\qquad+\check{1}\|W_6'(z_2)\|_{C^{1,\alpha}}\big).\end{aligned}\tag{4-16}$$

Step iii (the estimates of W_2 and W_5). By the system (2-26) and the related boundary conditions, a direct computation yields

$$\begin{aligned}
&\partial_1 W_5 - \tilde{a}_1 \partial_2 W_2 = O(\varepsilon)\cdot(W_1, W_2, W_3, W_6) + O(\check{1}) W_5 + O(\varepsilon)\partial_1 W_2\\
&\qquad\qquad\qquad\qquad\qquad\qquad + O(\check{1})\partial_2 W_2 + O(\check{1}) W_6'(z_2),\\
&\partial_2 W_5 + \check{1}\cot(\check{z}_2) W_5 + \tilde{a}_2 \partial_1 W_2\\
&\qquad\qquad = O(\check{1})\cdot(W_1, W_2, W_3, W_6, \partial_1 W_2)\\
&\qquad\qquad\qquad\qquad + O(\varepsilon)\cdot(W_5, \partial_2 W_2, \partial_1 W_5, W_6'),\\
&\quad W_2(0, z_2) = O(\varepsilon) W_4(0, z_2) + O(\check{1}) W_6(z_2),\\
&\quad W_2(1, z_2) = \varepsilon \tilde{P}_1(\check{z}_2) - \varepsilon \tilde{P}_2(\check{z}_2),\\
&\quad W_5(z_1, 0) = 0,\\
&\quad W_5(z_1, 1) = 0,
\end{aligned}\tag{4-17}$$

where $\tilde{a}_1$ and $\tilde{a}_2$ are positive constants that are defined like a_1 and a_2 respectively in (2-26) for the background solution, but with shock position at $r = \xi_1(1)$ rather than at $r = r_0$.

As in (3-16)–(3-18) and (3-21), we decompose $W_2 = W_{21} + W_{22}$ so that

$$\begin{aligned}
&\tilde{a}_2 \partial_1^2 W_{21} + \tilde{a}_1 \partial_2^2 W_{21} + (\tilde{a}_1/X_0)\cot(\check{z}_2)\partial_2 W_{21}\\
&\qquad = \partial_1\big(O(\check{1})\cdot(W_1, W_2, W_3, \partial_1 W_2)\\
&\qquad\qquad\qquad + O(\varepsilon)\cdot(W_5, \partial_2 W_2, \partial_1 W_5, W_6') + a_3(z) W_6\big)\\
&\qquad - \partial_2\big(O(\varepsilon)\cdot(W_1, W_2, W_3, W_6, \partial_1 W_2) + O(\check{1})\cdot(W_5, \partial_2 W_2, W_6')\big)\\
&\qquad - X_0^{-1}\cot(\check{z}_2)\\
&\qquad\quad \times (O(\varepsilon)\cdot(W_1, W_2, W_3, W_6, \partial_1 W_2) + O(\check{1})\cdot(W_5, \partial_2 W_2, W_6')),\\
&\quad W_{21}(0, z_2) = O(\varepsilon) W_4(0, z_2) + O(\check{1}) W_6(z_2),\\
&\quad W_{21}(1, z_2) = 0,\\
&\partial_2 W_{21}(z_1, 0) = 0,\\
&\partial_2 W_{21}(z_1, 1) = 0
\end{aligned}\tag{4-18}$$

and

$$\begin{aligned}
&\tilde{a}_2 \partial_1^2 W_{22} + \tilde{a}_1 \partial_2^2 W_{22} + (\tilde{a}_1/X_0)\cot(\check{z}_2)\partial_2 W_{22} = 0,\\
&W_{22}(0, z_2) = 0,\\
&W_{22}(1, z_2) = \varepsilon \tilde{P}_1(\check{z}_2) - \varepsilon \tilde{P}_2(\check{z}_2),\\
&\partial_2 W_{22}(z_1, 0) = 0,\\
&\partial_2 W_{22}(z_1, 1) = 0
\end{aligned}\tag{4-19}$$

and

$$\tag{4-20} \begin{aligned}\int_0^1 \sin\check{s}\big(O(\check{1})\cdot(W_1, W_2, W_3, \partial_1 W_2) + O(\varepsilon)\cdot(W_5, \partial_2 W_2, \partial_1 W_5, W_6')\\ + a_3(z)W_6 - \tilde{a}_2\partial_1 W_{21} - \tilde{a}_2\partial_1 W_{22}\big)(1, s)ds = 0,\end{aligned}$$

where $a_3(z_2) = O(\check{1})$. In particular, due to the estimates (3-48)–(3-50), we have

$$\tag{4-21} \begin{aligned} a_3(z) = -\Big(\frac{1}{\rho U_1^2} - \frac{1}{\gamma P}\Big)\\ \times \partial_1 P\bigg(\frac{1-z_1}{X_0(X_0+1-\xi_1(z_2))} + \frac{\xi_2(z_2)+z_1(X_0+1-\xi_2(z_2))}{X_0(X_0+1-\xi_1(z_2))(X_0+1-\xi_2(z_2))}\bigg)\\ + O(\varepsilon) < 0.\end{aligned}$$

Similar to the estimates in (3-42), by (B-20) in Lemma B.5 for the case $k = 0$, we have

$$\tag{4-22} \|W_{21}\|_{C^{1,\alpha}(E_1)} \le \check{C}\sum_{i=1}^{6}\|W_i\|_{C^{1,\alpha}(E_1)},$$

$$\tag{4-23} \|W_{22}\|_{C^{1,\alpha}(E_1)} \le C\varepsilon\|\tilde{P}_1(\check{z}_2) - \tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}[0,1]}.$$

In particular, for the case of $P(1, z_2) - \tilde{P}(1, z_2) = \tilde{C}$, we can determine $W_{22} = \tilde{C}z_1$ as in Section 3. Thus it follows from (4-20) and $\tilde{a}_2(z) = O(1) > 0$ that

$$\tag{4-24} \begin{aligned}\tilde{C} \le C\big(\check{1}(\xi_2(1) - \xi_1(1)) + \check{1}\|W_1\|_{C^{1,\alpha}} + \|W_{21}\|_{C^{1,\alpha}} + \check{1}\|W_3\|_{C^{1,\alpha}}\\ + \varepsilon\|W_5\|_{C^{1,\alpha}} + \check{1}\|W_6'(z_2)\|_{C^{1,\alpha}}\big).\end{aligned}$$

Similar to the estimates for (3-21), (3-26) and (3-31), together with (4-9) and (4-22)–(4-23), we get

$$\tag{4-25} \begin{aligned}\|W_{21}\|_{C^{1,\alpha}} \le \check{C}\big((\xi_2(1) - \xi_1(1)) + \|(W_1, W_3, W_5, W_6')\|_{C^{1,\alpha}}\big)\\ + \check{C}\varepsilon\|\tilde{P}_1(\check{z}_2) - \tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}},\end{aligned}$$

$$\tag{4-26} \begin{aligned}\|W_5\|_{C^{1,\alpha}} \le \check{C}\big((\xi_2(1) - \xi_1(1)) + \|(W_1, W_{21}, W_3, W_6')\|_{C^{1,\alpha}}\big)\\ + \check{C}\varepsilon\|\tilde{P}_1(\check{z}_2) - \tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}}.\end{aligned}$$

Thus, combining (4-25) and (4-26) with (4-23) yields

$$\tag{4-27} \begin{aligned}\|W_2\|_{C^{1,\alpha}} + \|W_5\|_{C^{1,\alpha}} \le \check{C}\big((\xi_2(1) - \xi_1(1)) + \|(W_1, W_3, W_6')\|_{C^{1,\alpha}}\big)\\ + C\varepsilon\|\tilde{P}_1(\check{z}_2) - \tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}}.\end{aligned}$$

Step iv (the estimate of W_3). It follows from (2-27) that

$$\tag{4-28} W_3 = O(\varepsilon)W_5 + O(1)W_1 + O(1)W_2.$$

This yields

$$\|W_3\|_{C^{1,\alpha}} \le C\big(\|W_1\|_{C^{1,\alpha}} + \|W_2\|_{C^{1,\alpha}} + \varepsilon\|W_5\|_{C^{1,\alpha}}\big). \tag{4-29}$$

Step v (the estimate of $W_2(0, z_2)$). Note that the supersonic background solution (ρ_0^-, P_0^-, U_0^-) satisfies the system (2-11), that is,

$$\begin{aligned} \frac{d(\rho_0^- U_0^-)}{dy_1} &= -\frac{2\rho_0^- U_0^-}{y_1}, \\ \frac{d(\rho_0^-(U_0^-)^2 + P_0^-)}{dy_1} &= -\frac{2\rho_0^-(U_0^-)^2}{y_1}. \end{aligned} \tag{4-30}$$

Set

$$\begin{aligned} &m_0(y_1) = (\rho_0^- U_0^-)^2, \\ &m_1(y_1) = \rho_0^-(U_0^-)^2 + P_0^-, \quad m_2 = \frac{\gamma}{\gamma-1}\frac{P_0^-}{\rho_0^-} + \tfrac{1}{2}(U_0^-)^2. \end{aligned}$$

It follows from Bernoulli's law, (2-27), that m_2 is a constant.

In addition, by (2-16) and (2-27), we have on $z_1 = 0$

$$\begin{aligned} \rho U_1 &= \sqrt{m_0} + \frac{\rho^2 U_1 U_2^2}{[\rho U_2^2 + P]}, \\ \rho U_1^2 + P &= m_1 + \frac{(\rho U_1 U_2)^2}{[\rho U_2^2 + P]}, \qquad m_2 = \frac{\gamma}{\gamma-1}\frac{P}{\rho} + \tfrac{1}{2}(U_1^2 + U_2^2). \end{aligned} \tag{4-31}$$

This implies

$$\begin{aligned} \rho &= \frac{\big(\sqrt{m_0} + \rho^2 U_1 U_2^2/[\rho U_2^2 + P]\big)^2}{m_1 - P + (\rho U_1 U_2)^2/[\rho U_2^2 + P]}, \\ U_1 &= \frac{m_1 - P}{\sqrt{m_0}}, \qquad m_2 = \frac{\gamma}{\gamma-1}\frac{P}{\rho} + \tfrac{1}{2}(U_1^2 + U_2^2). \end{aligned} \tag{4-32}$$

Substituting the first two expressions in (4-32) into the third equality in (4-32) yields on $z_1 = 0$

$$\begin{aligned} &\tfrac{1}{2}(m_1 - P)^2 + \frac{\gamma}{\gamma-1}P(m_1 - P) - m_2 m_0 \\ &= m_2\sqrt{m_0}\frac{\rho^2 U_1 U_2^2}{[\rho U_2^2 + P]} - \tfrac{1}{2}m_0 U_2^2 - \tfrac{1}{2}\sqrt{m_0}\frac{\rho^2 U_1 U_2^2}{[\rho U_2^2 + P]}\bigg(\frac{(m_1 - P)^2}{m_0} + U_2^2\bigg). \end{aligned} \tag{4-33}$$

Since $(S, P, U_1, U_2; \xi_1)$ and $(\tilde{S}, \tilde{P}, V_1, V_2; \xi_2)$ both satisfy (4-33), it follows from a direct computation and the estimates (3-48)–(3-50) for $(S, P, U_1, U_2; \xi_1)$

and $(\tilde{S}, \tilde{P}, V_1, V_2; \xi_2)$ that

$$\begin{aligned}a_4(z_2)W_2 = a_5(z_2)W_6(z_2) + O(\varepsilon^2)W_1 + O(\varepsilon^2)W_2 + O(\varepsilon^2)W_3 \\ + O(\varepsilon)W_4 + O(\varepsilon^2 X_0^{-1})W_6,\end{aligned} \tag{4-34}$$

where

$$\begin{aligned}a_4(z_2) &= \frac{\gamma}{\gamma-1}m_1(\xi_1) - \tfrac{1}{2}(m_1(\xi_1) + m_1(\xi_2) - P - \tilde{P}) - \frac{\gamma}{\gamma-1}(P + \tilde{P}) \\ &= \frac{\gamma}{\gamma-1}m_1(r_0) - (m_1(r_0) - \hat{P}_0^+(r_0)) - \frac{2\gamma}{\gamma-1}\hat{P}_0^+(r_0) + O(\sqrt{X_0}\varepsilon) \\ &= \frac{1}{\gamma-1}\hat{\rho}_0^+(r_0)(\hat{U}_0^+(r_0))^2 - \frac{\gamma}{\gamma-1}\hat{P}_0^+(r_0) + O(\sqrt{X_0}\varepsilon) \\ &= \frac{1}{\gamma-1}\hat{\rho}_0^+(r_0)\big((\hat{U}_0^+(r_0))^2 - c^2(\hat{\rho}_0^+(r_0))\big) + O(\sqrt{X_0}\varepsilon) < 0\end{aligned} \tag{4-35}$$

and

$$\begin{aligned}a_5(z_2) &= m_2\int_0^1 m_0'(\xi_2 + s(\xi_1 - \xi_2))ds \\ &\qquad - \tfrac{1}{2}(m_1(\xi_1) + m_1(\xi_2) - P - \tilde{P})\int_0^1 m_1'(\xi_2 + s(\xi_1 - \xi_2))ds \\ &\qquad - \frac{\gamma}{\gamma-1}\tilde{P}\int_0^1 m_1'(\xi_2 + s(\xi_1 - \xi_2))ds \\ &= m_2 m_0'(r_0) - (m_1(r_0) - \hat{P}_0^+(r_0))m_1'(r_0) - \frac{\gamma}{\gamma-1}\hat{P}_0^+(r_0)m_1'(r_0) \\ &\qquad + O(\sqrt{X_0}\varepsilon) \\ &= -2\frac{(\rho_0^-(U_0^-)^2)(r_0)}{(\gamma-1)r_0}((\gamma+1)P_0^-(r_0) - \hat{P}_0^+(r_0)) + O(\sqrt{X_0}\varepsilon).\end{aligned} \tag{4-36}$$

Next, we analyze the sign of $a_5(z_2)$ for small ε and especially the sign of $(\gamma+1)P_0^-(r_0) - \hat{P}_0^+(r_0)$.

In fact, by (4-32), $\hat{P}_0^+(r_0)$ is a solution of the algebraic equation

$$F(s) = \tfrac{1}{2}(m_1(r_0) - s)^2 + \frac{\gamma}{\gamma-1}s(m_1(r_0) - s) - m_2 m_0(r_0) = 0. \tag{4-37}$$

Since

$$\begin{aligned}F(P_0^-(r_0)) = 0, \qquad F''(s) &= -\frac{\gamma+1}{\gamma-1} < 0, \\ F'(P_0^-(r_0)) &= \frac{1}{\gamma-1}\big((\rho_0^-(U_0^-)^2)(r_0) - \gamma P_0^-(r_0)\big) \\ &= \frac{\rho_0^-(r_0)}{\gamma-1}((U_0^-(r_0))^2 - c^2(\rho_0^-(r_0))) > 0,\end{aligned}$$

which follows from direct computations, $F(s)$ is a concave function and $P_0^-(r_0)$ is a left zero point of $F(s)$.

Using the assumption $M_0^-(X_0) > \sqrt{(\gamma+3)/2}$ on the Mach number for the supersonic incoming flow, we have

$$F((\gamma+1)P_0^-(r_0)) = \frac{(\rho_0^-(r_0))^2c^2(\rho_0^-(r_0))}{2(\gamma-1)}\big(2(U_0^-(r_0))^2-(\gamma+3)c^2(\rho_0^-(r_0))\big) > 0.$$

This shows that

$$\hat{P}_0^+(r_0) > (\gamma+1)P_0^-(r_0). \tag{4-38}$$

Combining (4-38) with (4-36), we have

$$a_5(z_2) = O(\check{1}) \quad \text{and} \quad a_5(z_2) > 0. \tag{4-39}$$

On the other hand, by the estimates (4-9), (4-10), (4-16), (4-27) and (4-29), we have

$$\begin{aligned}\sum_{i=1}^{4}\|W_i\|_{C^{1,\alpha}} + \|W_6'(z_2)\|_{C^{1,\alpha}}&\\ \le \check{C}|\xi_1(1)-\xi_2(1)| + C\varepsilon\|\tilde{P}_1(\check{z}_2)-\tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}}&.\end{aligned} \tag{4-40}$$

This, together with (4-34)–(4-36), yields

$$W_2(0,z_2) \ge \check{b}_1(\xi_2(1)-\xi_1(1)) - b_2\varepsilon\|\tilde{P}_1(\check{z}_2)-\tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}[0,1]}, \tag{4-41}$$

where b_i for $i=1,2$ is a generic positive constant of order $O(1)$.

Based on Steps i–v, we can prove Theorem 4.1.

Using (4-21) and substituting (4-40) into (4-20) (noting that (4-20) holds for all $z_1 \in [0,1]$), we have, for all $z_1 \in [0,1]$,

$$\begin{aligned}\int_0^1 \sin\check{s}\big(\check{b}_3(\xi_2(1)-\xi_1(1))&\\ - b_4\varepsilon\|\tilde{P}_1(\check{z}_2)-\tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}} - \partial_1 W_{21}\big)(z_1,s)ds \le 0&,\end{aligned} \tag{4-42}$$

where b_i for $i=3,4$ is a generic positive constant. In particular,

$$b_3 \ge C(-X_0a_3(z)+\check{C}) = O(1) > 0$$

because $a_3(z) = O(\check{1}) < 0$ in (4-21).

If we assume

$$\varepsilon\|\tilde{P}_1(\check{z}_2)-\tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}} < \min\left\{\frac{\check{b}_1}{2b_2}, \frac{\check{b}_3}{2b_4}\right\}(\xi_2(1)-\xi_1(1)), \tag{4-43}$$

is false (that is, that this statement is true with "≥" instead of "<"), then (4-3) has been shown. If we assume (4-43) is true, then this means $W_2(0, z_2) > 0$. Due to $W_2(0, z_2) = W_{21}(0, z_2) + W_{22}(0, z_2)$ and $W_{22}(0, z_2) = 0$ in (4-18), we then get

$$W_{21}(0, z_2) > 0. \tag{4-44}$$

On the other hand, it follows from (4-42) and (4-43) that for $z_1 \in [0, 1]$

$$\partial_1\Big(\int_0^1 sW_{21}(z_1, s)\sin\check{s}ds\Big) > 0. \tag{4-45}$$

Combining (4-44) with (4-45) yields

$$\int_0^1 W_{21}(1, s)\sin\check{s}ds > 0.$$

However, this contradicts that $W_{21}(1, z_2) = 0$ in (4-18). Thus (4-43) does not hold, that is, we have shown that there exists a constant $C > 0$ such that

$$|\xi_2(1) - \xi_1(1)| \le CX_0\varepsilon\|\tilde{P}_1(\check{z}_2) - \tilde{P}_2(\check{z}_2)\|_{C^{1,\alpha}}.$$

Combining this with (4-40), we complete the proof of (4-3) and (4-4).

Finally, by (4-24) and (4-25) and an argument analogous to the one for (4-3) and (4-4), we can also show (4-5). We omit the details. □

Remark 4.2. From (4-3) of Theorem 4.1, we have established that the position of the shock depends continuously on the exit pressure. If the condition (2-30) is replaced by $P(1, z_2) = P_e + \varepsilon\tilde{P}(\check{z}_2) + C$, then (4-5) establishes that the corresponding position of the shock depends monotonically on the exit pressure. Thus, the constant C_0 in Theorem 3.1′ can be considered as a function of the variable $y_1 \in (X_0, X_0 + 1)$, which is denoted by $C_0(y_1)$. Furthermore, it follows from (4-5) that the function $C_0(y_1)$ is Lipschitz continuous and decreasing.

5. Proof of Theorem 1.1.

First, we prove that the system (2-26)–(2-28), (2-32) with (2-29)–(2-31) has a solution.

Denote by $\bar{P}_1 = P_e - \sqrt{X_0}\varepsilon$ and $\bar{P}_2 = P_e + \sqrt{X_0}\varepsilon$ the exit pressures of the symmetric transonic shock solutions with corresponding shock positions at $y_1 = r_1$ and $y_1 = r_2$, respectively. Then it follows from (4-5) in Theorem 4.1 that $r_1 > r_2$ holds true.

For each fixed point $(y_1^*, 1)$ with $y_1^* \in [r_2, r_1]$, it follows from Theorem 3.1′ and Remark 4.2 that there exists a constant $C_0(y_1^*)$ such that problem (2-26)–(2-28), (2-32) with (2-29), (2-31) and the exit pressure $P = P_e + \varepsilon P_0(\theta) + C_0(y_1^*)$ has a unique solution $(S, P, U_1, U_2; \xi(z_2))$ that admits the estimates in Theorem 3.1′.

If $y_1^* = r_2$, it follows from (3-4) and (3-47) that

$$|C_0(r_2) - \sqrt{X_0}\varepsilon| \leq C\varepsilon. \tag{5-1}$$

This implies that $C_0(r_2) > 0$. Analogously, we have $C_0(r_1) < 0$. Therefore, in terms of Theorem 4.1 and Remark 4.2, there exists a unique point $y_1^0 \in (r_2, r_1)$ such that $C_0(y_1^0) = 0$, that is, the system (2-26)–(2-28), (2-32) with (2-29)–(2-31) has a unique angular symmetric solution $(S, P, \rho, U_1, U_2; \xi)$. Also, by Theorem 3.1, this solution also satisfies the estimates

$$\|\xi - r_0\|_{L^\infty[0,1]} \leq CX_0\varepsilon, \qquad \|\xi'\|_{C^{2,\alpha}[0,1]} \leq C\varepsilon \tag{5-2}$$

and

$$\|(S, P, U_1) - (S_0^+, \hat{P}_0^+(z_1), \hat{U}_0^+(z_1))\|_{C^{2,\alpha}(E_+)} \leq C\varepsilon. \tag{5-3}$$

According to the constructions of the spaces of S_σ and Ξ_δ in Section 3, we can derive that

$$\begin{aligned} \partial_{z_2}S(z_1, 0) = \partial_{z_2}P(z_1, 0) = \partial_{z_2}U_1(z_1, 0) = 0,\\ \partial_{z_2}S(z_1, 1) = \partial_{z_2}P(z_1, 1) = \partial_{z_2}U_1(z_1, 1) = 0,\\ U_2(z_1, 0) = \partial_{z_2}^2 U_2(z_1, 0) = U_2(z_1, 1) = 0,\\ \xi'(0) = \xi^{(3)}(0) = \xi'(1) = 0. \end{aligned} \tag{5-4}$$

Next, we verify that the axisymmetric solution $(S, P, U_1, U_2; \xi)$ satisfies all the estimates in Theorem 1.1 in the (x_1, x_2, x_3) coordinate system.

The transformation (2-20) keeps the equivalence of $C^{2,\alpha}$ norms between the coordinates (y_1, y_2) and (z_1, z_2). Denoting the solution by $((S, P, U_1, U_2)(y); \xi(y_2))$ in the coordinates (y_1, y_2), we have

$$|\xi(y_2) - r_0| \leq CX_0\varepsilon, \quad \|\xi'(y_2)\|_{C^{2,\alpha}[0,1]} \leq C\varepsilon \tag{5-5}$$

and

$$\|(S, P, U_1, U_2) - (S_0^+, \hat{P}_0^+(y_1), \hat{U}_0^+(y_1), 0)\|_{C^{2,\alpha}(R_+)} \leq C\varepsilon. \tag{5-6}$$

In addition, it follows from (5-4) and a direct computation that

$$\begin{aligned} \partial_{y_2}S(y_1, 0) = \partial_{y_2}P(y_1, 0) = \partial_{y_2}U_1(y_1, 0) = 0,\\ \partial_{y_2}S(y_1, 1) = \partial_{y_2}P(y_1, 1) = \partial_{y_2}U_1(y_1, 1) = 0,\\ U_2(y_1, 0) = \partial_{y_2}^2 U_2(y_1, 0) = U_2(y_1, 1) = 0,\\ \xi'(0) = \xi^{(3)}(0) = \xi'(1) = 0. \end{aligned} \tag{5-7}$$

Therefore, by the inverse transformations of (2-1) and (2-2), the solution to the problem (1-1) with (1-2)–(1-5) has the form

$$(S, P)(x_1, x_2, x_3) = (S, P)\biggl((x_1^2+x_2^2+x_3^2)^{1/2}, X_0 \arcsin\biggl(\frac{(x_2^2+x_3^2)^{1/2}}{(x_1^2+x_2^2+x_3^2)^{1/2}}\biggr)\biggr),$$

and

$$\begin{pmatrix} u_1\\ u_2\\ u_3\end{pmatrix}(x_1, x_2, x_3) = \frac{1}{(x_1^2+x_2^2+x_3^2)^{1/2}}\begin{bmatrix} x_1 & (x_2^2+x_3^2)^{1/2}\\ x_2 & -x_1x_2/(x_2^2+x_3^2)^{1/2}\\ x_3 & -x_1x_3/(x_2^2+x_3^2)^{1/2}\end{bmatrix}$$
$$\cdot\begin{pmatrix} U_1\\ U_2\end{pmatrix}\biggl((x_1^2+x_2^2+x_3^2)^{1/2}, X_0 \arcsin\bigl(\frac{(x_2^2+x_3^2)^{1/2}}{(x_1^2+x_2^2+x_3^2)^{1/2}}\bigr)\biggr),$$

and the shock position $x_1 = \eta(x_2, x_3)$ is given by the implicit function

$$G(x_1, x_2, x_3) \equiv (x_1^2+x_2^2+x_3^2)^{1/2} - \xi\biggl(X_0 \arcsin\biggl(\frac{(x_2^2+x_3^2)^{1/2}}{(x_1^2+x_2^2+x_3^2)^{1/2}}\biggr)\biggr) = 0, \tag{5-8}$$

where we have for small ε

$$\partial_{x_1} G = \frac{x_1}{(x_1^2+x_2^2+x_3^2)^{1/2}} + \xi'\biggl(X_0 \arcsin\bigl(\frac{(x_2^2+x_3^2)^{1/2}}{(x_1^2+x_2^2+x_3^2)^{1/2}}\bigr)\biggr)\frac{X_0(x_2^2+x_3^2)^{1/2}}{x_1^2+x_2^2+x_3^2} > 0$$

because $|\xi'| \le C\varepsilon$.

Thanks to (5-7) and Lemmas B.3 and B.4, we know that

$$(S^+(x), P^+(x), u_1^+(x), u_2^+(x), u_3^+(x))$$

belongs to $C^{2,\alpha}(\overline{\Omega}_+)$ and satisfies the estimates in Theorem 1.1.

Finally, we show that $\eta(x_2, x_3) \in C^{3,\alpha}(\bar{S}_e)$ and satisfies Theorem 1.1(i).

Since the shock surface $x_1 = \eta(x_2, x_3)$ is determined by (5-8),

$$\|\eta - (r_0^2 - x_2^2 - x_3^2)^{1/2}\|_{L^\infty(S_e)} \le C\|\xi - r_0\|_{L^\infty[0,1]} \le CX_0\varepsilon. \tag{5-9}$$

In addition, $\eta(x_2, x_3)$ satisfies the Rankine–Hugoniot conditions (1-2), so we have

$$\begin{aligned} \partial_{x_2}\eta &= \frac{[\rho u_1u_2][P+\rho u_3^2] - [\rho u_1u_3][\rho u_2u_3]}{[P+\rho u_2^2][P+\rho u_3^2] - [\rho u_2u_3]^2},\\ \partial_{x_3}\eta &= \frac{[\rho u_1u_3][P+\rho u_2^2] - [\rho u_1u_2][\rho u_2u_3]}{[P+\rho u_2^2][P+\rho u_3^2] - [\rho u_2u_3]^2}. \end{aligned} \tag{5-10}$$

Similarly, $\eta_0(x_2, x_3) = (r_0^2 - x_2^2 - x_3^2)^{1/2}$ also satisfies (5-10) when the solution $(\rho^\pm, P^\pm, u^\pm)$ is replaced by the corresponding background solution in (5-10).

Therefore, by Remark A.2, (5-9) and the interpolation theorem in Hölder space, we have

$$\begin{aligned}&\big\|\nabla_{x_2,x_3}\big(\eta(x_2,x_3)-(r_0^2-x_2^2-x_3^2)^{1/2}\big)\big\|_{C^{2,\alpha}(\bar{S}_e)}\\&\quad\le C(\varepsilon+\|\nabla_x(S_0^+,\hat{P}_0^+,\hat{u}_{1,0}^+,\hat{u}_{2,0}^+,\hat{u}_{3,0}^+)\|_{C^{2,\alpha}}\|\eta-(r_0^2-x_2^2-x_3^2)^{1/2}\|_{C^{2,\alpha}(\bar{S}_e)})\\&\quad\le C\varepsilon.\end{aligned}$$

This completes the proof of Theorem 1.1. □

Appendix A.

In this appendix, we will describe the transonic solution of the problem (1-1) with (1-2)–(1-5), when the exit pressure is a suitable constant P_e under the assumptions given in Section 1 on the nozzle walls and the supersonic incoming flow. Such a solution, called a background solution, can be obtained by solving the related ordinary differential equations. Related analysis has been given in [Courant and Friedrichs 1948, Section 147] and the details can be seen in [Xin and Yin 2008b]. For the reader's convenience and because it's needed for the computations in this paper, we will give a detailed statement.

Theorem A.1. *If the three-dimensional nozzle wall Γ and the supersonic incoming flow are as defined in Section 1, then there exist two constant pressures $\tilde{P}_1$ and $\tilde{P}_2$ with $\tilde{P}_1 < \tilde{P}_2$ such that if the exit pressure $\tilde{P}_e \in (\tilde{P}_1, \tilde{P}_2)$, then the system* (1-2) *has a symmetric transonic shock solution*

$$(P,u_1,u_2,u_3,S)=\begin{cases}(P_0^-(r),u_{1,0}^-(x),u_{2,0}^-(x),u_{3,0}^-(x),S_0^-) & \textit{for } r<r_0,\\ (P_0^+(r),u_{1,0}^+(x),u_{2,0}^+(x),u_{3,0}^+(x),S_0^+) & \textit{for } r>r_0,\end{cases}$$

where $u_{i,0}^+(x)=U_0^+(r)x_i/r$ for $i=1,2,3$, $X_0<r_0<X_0+1$, S_0^+ is a constant, and $(P_0^+(r),U_0^+(r))$ is C^3-smooth.

See Theorem 1.1 in [Xin and Yin 2008b] for the proof.

Next, we cite two useful remarks, which were stated in [Xin and Yin 2008b].

Remark A.2. By the assumption (1-6), we have for $r_0 \le r \le X_0+1$

$$\left|\frac{d^kU_0^+(r)}{dr^k}\right|+\left|\frac{d^kP_0^+(r)}{dr^k}\right|\le\frac{C_k}{X_0^k}\quad\text{for } k=1,2,3.$$

Remark A.3. One can obtain an extension $(\hat{\rho}_0^+(r),\hat{U}_0^+(r))$ of $(\rho_0^+(r),U_0^+(r))$ for $r\in(X_0,X_0+1)$ by solving the Euler system.

Appendix B.

We now give some elementary facts and computations often used in Section 5. Compared with the similar results in [Li et al. 2010a, Appendix B], the estimates here are more delicate since we require them to be independent of X_0. Here and in what follows, X_0 is defined as in Section 1 and C stands for a generic positive constant that is independent of X_0.

For the convenience, we set

$$\begin{aligned}
&E_1 = \{(z_1, z_2) \in R^2 : 0 < z_1 < 1,\ 0 < z_2 < 1\},\\
&E_2 = \{(x_1, x_2, x_3) \in R^3 : 0 < x_1 < 1,\ x_2^2 + x_3^2 < 1\},\\
&E_3 = \{(z_1, z_2, z_3) \in R^3 : 0 < z_1 < 1,\ 0 < z_2 < 1,\ 0 \le z_3 < 2\pi\},\\
&E_4 = \{(x_1, x_2) \in R^2 : x_1^2 + x_2^2 \le 1\}.
\end{aligned}$$

Lemma B.1. *Let*

$$\phi(x_1, x_2) = \left(\frac{1}{X_0}\cot\!\left(\frac{(x_1^2+x_2^2)^{1/2}}{X_0}\right) - \frac{1}{(x_1^2+x_2^2)^{1/2}}\right)\frac{x_2}{(x_1^2+x_2^2)^{1/2}}.$$

Then we have

$$\|\phi\|_{C^{0,1}(E_4)} \le C. \tag{B-1}$$

Proof. Note that $\phi(x_1, x_2)$ can be rewritten as

$$\begin{aligned}
\phi(x_1, x_2) &= \frac{\int_0^1 (\cos(\check{\rho}) - \cos(s\check{\rho}))ds}{X_0 \sin\check{\rho}}\,\frac{x_2}{\rho}\\
&= \frac{2x_2}{\rho}\,\frac{\int_0^1 (\sin(\frac{1}{2}(1+s)\check{\rho})\sin(\frac{1}{2}(s-1)\check{\rho}))ds}{X_0\sin(\check{\rho})},
\end{aligned}$$

where $\rho = (x_1^2 + x_2^2)^{1/2}$

It is easy to see that

$$\|\phi\|_{L^\infty(E_4)} \le C. \tag{B-2}$$

For any two distinct points (x_{11}, x_{21}) and (x_{12}, x_{22}) in E_4, it follows from a direct computation that

$$\phi(x_{11}, x_{21}) - \phi(x_{12}, x_{22}) = I_1 + I_2 + I_3, \tag{B-3}$$

where, with $a = (x_{11}^2 + x_{21}^2)^{1/2}$ and $b = (x_{12}^2 + x_{22}^2)^{1/2}$,

$$\begin{aligned}
I_1 &= \left(\check{1}\cot(\check{a}) - \frac{1}{a}\right)\frac{x_{21} - x_{22}}{a},\\
I_2 &= -\left(\check{1}\cot(\check{a}) - \frac{1}{a}\right)\frac{x_{22}\big((x_{11} - x_{12})(x_{11} + x_{12}) + (x_{21} - x_{22})(x_{21} + x_{22})\big)}{ab\big(a+b\big)},\\
I_3 &= \frac{x_{12}}{b}\left(\check{1}\cot(\check{a}) - \frac{1}{a} - \check{1}\cot(\check{b}) + \frac{1}{b}\right).
\end{aligned}$$

Now we only estimate I_3 since the treatments on I_1 and I_2 are analogous or even simpler.

Assume that $a \geq b$ without loss of generality. Then a direct computation yields

$$|I_3| \leq \left| \frac{ab\sin(\check{b}-a) - X_0\sin(\check{a})\sin(\check{b})(b-a)}{X_0\sin(\check{a})\sin(\check{b})ab} \right|.$$

Since

$$\begin{aligned} |ab\sin(\check{b}-a) - X_0\sin(\check{a})\sin(\check{b})(b-a)| \\ &= \left| \check{b} - aab \int_0^1 \int_0^1 (\cos(s\check{b}-a) - \cos(s\check{a})\cos(t\check{b}))dsdt \right| \\ &\leq \frac{|b-a|}{X_0} ab\big(\sin(\check{a})\sin(\check{b}) + 2\sin(\check{b})\sin(\check{2}b)\big), \end{aligned}$$

we have $|I_3| \leq C|a-b|$ and hence

$$|\phi(x_{11}, x_{21}) - \phi(x_{12}, x_{22})| \leq C|a-b|. \tag{B-4}$$

Combining (B-4) with (B-2) yields Lemma B.1. □

Remark B.2. By the computation of $|I_3|$, we show that

$$X_0^{-1}\cot((x_2^2+x_3^2)^{1/2}X_0^{-1}) - (x_2^2+x_3^2)^{-1/2}$$

is in $C^{0,1}(E_4)$ and is no greater than C.

Lemma B.3. (i) *For $\phi(z_1, z_2) \in C^{\alpha}(E_1)$ with $0 < \alpha < 1$, there exists a constant $C > 1$ such that*

$$\frac{1}{C}\|\phi(x_1, (x_2^2+x_3^2)^{1/2})\|_{C^{\alpha}(E_2)} \leq \|\phi\|_{C^{\alpha}(E_1)} \leq C\|\phi(x_1, (x_2^2+x_3^2)^{1/2})\|_{C^{\alpha}(E_2)}. \tag{B-5}$$

If $\phi(z_1, z_2) \in C^{k,\alpha}(E_1)$ for some $k \in \{1, 2\}$ and $\partial_{z_2}\phi(z_1, 0) = 0$, then there exists a constant $C > 1$ such that

$$\begin{aligned} \frac{1}{C}\|\phi(x_1, (x_2^2+x_3^2)^{1/2})\|_{C^{k,\alpha}(E_2)} &\leq \|\phi\|_{C^{k,\alpha}(E_1)} \\ &\leq C\|\phi(x_1, (x_2^2+x_3^2)^{1/2})\|_{C^{k,\alpha}(E_2)}. \end{aligned} \tag{B-6}$$

(ii) *If $\phi(z_1, z_2) \in C^{k,\alpha}(E_1)$ with some $k \in \{1, 2\}$ and $\phi(z_1, 0) = 0$, then there exists a constant $C_2 > 1$ such that*

$$\|\check{1}\cot(\check{z}_2)\phi\|_{C^{k-1,\alpha}(E_1)} \leq C\|\phi\|_{C^{k,\alpha}(E_1)}. \tag{B-7}$$

Proof. Since (B-5) and (B-6) can be verified directly, we omit the proof. Next we show (B-7).

Using $\phi(z_1, 0) = 0$, we have

$$\check{1}\cot(\check{z}_2)\phi(z_1, z_2) = \frac{\check{z}_2\cos(\check{z}_2)}{\sin\check{z}_2}\int_0^1 \partial_{z_2}\phi(z_1, sz_2)ds$$
$$= \cos(\check{z}_2)\biggl(1 + \frac{\check{z}_2 - \sin\check{z}_2}{\sin\check{z}_2}\biggr)\int_0^1 \partial_{z_2}\phi(z_1, sz_2)ds,$$

this yields for $k = 1$ or 2

(B-8)
$$\bigl\|\check{1}\cot(\check{z}_2)\phi(z_1, z_2)\bigr\|_{C^{k-1,\alpha}(E_1)} \le C\biggl(1 + \biggl\|\frac{\check{z}_2 - \sin\check{z}_2}{\sin\check{z}_2}\biggr\|_{C^{k-1,\alpha}[0,1]}\biggr)\|\partial_{z_2}\phi\|_{C^{k-1,\alpha}(E_1)}.$$

Since

$$\frac{\check{z}_2 - \sin\check{z}_2}{\sin\check{z}_2} = \frac{\check{z}_2}{\sin\check{z}_2}\int_0^1 (1 - \cos(s\check{z}_2))ds = \frac{2\check{z}_2}{\sin\check{z}_2}\int_0^1 \bigl(\sin(\tfrac{1}{2}s\check{z}_2)\bigr)^2 ds,$$

$$\frac{d}{dz_2}\biggl(\frac{\check{z}_2 - \sin(\check{z}_2)}{\sin\check{z}_2}\biggr) = \frac{2\sin\check{z}_2 - 2\check{z}_2\cos\check{z}_2}{X_0\sin^2\check{z}_2}\int_0^1 (\sin(\tfrac{1}{2}s\check{z}_2))^2 ds + \frac{\check{z}_2}{X_0\sin\check{z}_2}\int_0^1 \sin(s\check{z}_2)s\,ds,$$

$$\frac{d^2}{dz_2^2}\biggl(\frac{\check{z}_2 - \sin\check{z}_2}{\sin\check{z}_2}\biggr) = \frac{2\check{z}_2 + 2\check{z}_2\cos^2\check{z}_2 - \sin(2\check{z}_2)}{X_0^2\sin^3\check{z}_2}\int_0^1 (\sin(\tfrac{1}{2}s\check{z}_2))^2 ds$$
$$+ \frac{2\sin\check{z}_2 - 2\check{z}_2\cos(\check{z}_2)}{X_0\sin^2\check{z}_2}\int_0^1 \sin(s\check{z}_2)s\,ds + \frac{\check{z}_2}{X_0^2\sin\check{z}_2}\int_0^1 \cos(s\check{z}_2)s^2 ds,$$

and because

$$\int_0^1 (\sin(\tfrac{1}{2}s\check{z}_2))^2 ds \le \tfrac{1}{4}\check{z}_2^2, \quad \text{and} \quad \int_0^1 \sin(\tfrac{1}{2}s\check{z}_2)ds \le \tfrac{1}{2}\check{z}_2,$$

we have

$$\biggl\|\frac{z_2 - X_0\sin\check{z}_2}{X_0\sin\check{z}_2}\biggr\|_{C^{1,1}[0,1]} \le C.$$

Combining this with (B-8) yields (B-7) for $k = 1$ or $k = 2$. □

Lemma B.4. (i) *For* $\phi(z_1, z_2) \in C^{k,\alpha}(E_1)$ *with some* $k = \{0, 1\}$ *and* $\phi(z_1, 0) = 0$,

$$\sum_{i=2}^{3} \left\| \frac{x_i}{(x_2^2+x_3^2)^{1/2}} \phi(x_1, (x_2^2+x_3^2)^{1/2}) \right\|_{C^{k,\alpha}(E_2)} \leq C \|\phi(z_1, z_2)\|_{C^{k,\alpha}(E_1)}. \tag{B-9}$$

(ii) *For* $\phi \in C^{2,\alpha}(E_1)$ *and* $\phi(z_1, 0) = \partial_{z_2}^2 \phi(z_1, 0) = 0$,

$$\sum_{i=2}^{3} \left\| \frac{x_i}{(x_2^2+x_3^2)^{1/2}} \phi(x_1, (x_2^2+x_3^2)^{1/2}) \right\|_{C^{2,\alpha}(E_2)} \leq C \|\phi(z_1, z_2)\|_{C^{2,\alpha}(E_1)}. \tag{B-10}$$

Proof. Put $\rho = (x_2^2+x_3^2)^{1/2}$. Set

$$V_i(x_1, x_2, x_3) = (x_i/\rho)\phi(x_1, (x_2^2+x_3^2)^{1/2}) \quad \text{for } i = 2, 3.$$

Then

$$\|V_i\|_{L^\infty(E_2)} \leq \|\phi(r)\|_{L^\infty(E_1)} \quad \text{for } i = 2, 3. \tag{B-11}$$

Since V_2 and V_3 have the analogous forms, it suffices to treat V_2.

(i) First we show (B-9).

For any two distinct points (x_{11}, x_{21}, x_{31}) and (x_{12}, x_{22}, x_{32}) in E_2, we may assume without loss of generality that $|x_{21}| \geq |x_{22}|$. Put $a = (x_{21}^2 + x_{31}^2)^{1/2}$ and $b = (x_{22}^2+x_{32}^2)^{1/2}$. Then

$$\begin{aligned} V_2(x_{11}, x_{21}, x_{31}) - V_2(x_{12}, x_{22}, x_{32}) &= \frac{x_{21}}{a}\phi(x_{11}, a) - \frac{x_{22}}{b}\phi(x_{12}, b) \\ &= J_1 + J_2 + J_3. \end{aligned} \tag{B-12}$$

where

$$\begin{aligned} J_1 &= \frac{x_{21}-x_{22}}{a}\phi(x_{11}, a), \\ J_2 &= -\frac{x_{22}\big((x_{21}-x_{22})(x_{21}+x_{22}) + (x_{31}-x_{32})(x_{31}+x_{32})\big)}{ab(a+b)}\phi(x_{11}, a), \\ J_3 &= \frac{x_{22}}{b}\big(\phi(x_{11}, a) - \phi(x_{12}, b)\big). \end{aligned}$$

By $\phi(z_1, 0) = 0$ and the assumption $|x_{21}| \geq |x_{22}|$, a direct computation yields

$$\begin{aligned} |J_1| &\leq [\phi]_\alpha \frac{(|x_{21}|+|x_{22}|)^{1-\alpha}}{a^{1-\alpha}}|x_{21}-x_{22}|^\alpha \leq 2^{1-\alpha}[\phi]_\alpha |x_{21}-x_{22}|^\alpha, \\ |J_2| &\leq 2[\phi]_\alpha \frac{|x_{22}|\big(|x_{21}-x_{22}| + |x_{31}-x_{32}|\big)}{a^{1-\alpha}b} \\ &\leq 2^{2-\alpha}[\phi]_\alpha \big(|x_{21}-x_{22}|^\alpha + |x_{31}-x_{32}|^\alpha\big), \\ |J_3| &\leq [\phi]_\alpha \big((x_{11}-x_{12})^2 + (x_{21}-x_{22})^2 + (x_{31}-x_{32})^2\big)^{\alpha/2}. \end{aligned} \tag{B-13}$$

Here $[\phi]_\alpha$ denotes the Hölder seminorm with exponent α.

Combining (B-13) with (B-12) and (B-11) yields

$$\|V_2\|_{C^\alpha(E_2)} \le C\|\phi\|_{C^\alpha(E_1)}. \tag{B-14}$$

If $\phi \in C^{1,\alpha}(E_2)$ and $\phi(z_1, 0) = 0$, we have

$$\begin{aligned}
\partial_{x_1} V_2 &= (x_2/\rho)\partial_{z_1}\phi(x_1, \rho),\\
\partial_{x_2} V_2 &= \frac{x_3^2}{\rho^3}\phi(x_1, \rho) + \frac{x_2^2}{\rho^2}\partial_{z_2}\phi(x_1, \rho),\\
\partial_{x_3} V_2 &= -\frac{x_2x_3}{\rho^3}\phi(x_1, \rho) + \frac{x_2x_3}{\rho^2}\partial_{z_2}\phi(x_1, \rho),
\end{aligned}$$

Next, we only analyze $\partial_{x_2} V_2$ since the treatment of $\partial_{x_1} V_2$ and $\partial_{x_3} V_2$ is similar. Rewrite $\partial_{x_2} V_2$ as $\partial_{x_2} V_2 = J_5 + J_6$, where

$$J_5 = \frac{x_3^2}{\rho^2}\int_0^1 \big(\partial_{z_2}\phi(x_1, \theta\rho) - \partial_{z_2}\phi(x_1, \rho)\big)d\theta \quad \text{and} \quad J_6 = \partial_{z_2}\phi(x_1, \rho).$$

For convenience, we set

$$\overline{V}(x_1, \rho) = \int_0^1 (\partial_{z_2}\phi(x_1, \theta\rho) - \partial_{z_2}\phi(x_1, \rho))d\theta.$$

Then $\overline{V}(x_1, 0) = 0$. Applying the same argument as for (B-14) yields

$$\|J_5\|_{C^\alpha(E_2)} \le C\|\phi\|_{C^{1,\alpha}(E_1)}. \tag{B-15}$$

In addition, by (B-5), we have

$$\|J_6\|_{C^\alpha(E_2)} \le C\|\phi\|_{C^{1,\alpha}(E_1)}. \tag{B-16}$$

Thus, combining (B-15) and (B-16) with (B-14) yields (B-9).

(ii) We now show (B-10).

For $\phi(z) \in C^{2,\alpha}(E_1)$ with $\phi(z_1, 0) = \partial_{z_2}^2\phi(z_1, 0) = 0$, we have

$$\begin{aligned}
\partial_{x_1}^2 V_2 &= \frac{x_2}{\rho}\partial_{z_1}^2\phi(x_1, \rho),\\
\partial_{x_1x_2}^2 V_2 &= \frac{x_3^2}{\rho^3}\partial_{z_1}\phi(x_1, \rho) + \frac{x_2^2}{\rho^2}\partial_{z_1z_2}^2\phi(x_1, \rho)\\
&= \frac{x_3^2}{\rho^2}\int_0^1 (\partial_{z_1z_2}^2\phi(x_1, \theta\rho) - \partial_{z_1z_2}^2\phi(x_1, \rho))d\theta + \partial_{z_1z_2}^2\phi(x_1, \rho),\\
\partial_{x_1x_3}^2 V_2 &= -\frac{x_2x_3}{\rho^3}\partial_{z_1}\phi(x_1, \rho) + \frac{x_2x_3}{\rho^2}\partial_{z_1z_2}^2\phi(x_1, \rho).
\end{aligned}$$

It follows from $\phi(z_1, 0) = 0$, $\partial^2_{z_1}\phi(z_1, 0) = 0$ and (B-9) that

$$\|\partial^2_{x_1} V_2\|_{C^\alpha(E_1)} \le C\|\phi\|_{C^{2,\alpha}(E_1)}. \tag{B-17}$$

In a similar proof as for (B-15) and (B-16), we have

$$\sum_{i=2}^{3}\|\partial^2_{x_1x_i} V_2\|_{C^\alpha(E_1)} \le C\|\phi\|_{C^{2,\alpha}(E_1)}, \tag{B-18}$$

The quantities $\partial^2_{x_ix_j} V_2$ for $i, j = 2, 3$ can also be estimated in the same way. Therefore, due to (B-17), (B-18) and (B-9), we have proved (B-10). □

Lemma B.5. *Let $k = 0$ or $k = 1$. If $f_i(z_1, z_2) \in C^{k,\alpha}(E_1)$ and $g_i(z_2) \in C^{k+1,\alpha}[0, 1]$ with $g'_i(0) = g'_i(1) = 0$ for $i = 1, 2$ and $\partial_{z_2} f_1(z_1, 0) = f_2(z_1, 0) = 0$, then the problem*

$$\begin{aligned} \partial^2_{z_1}U + \partial^2_{z_2}U + \check{1}\cot(\check{z}_2)\partial_{z_2}U &= \partial_{z_1} f_1(z_1, z_2) + \partial_{z_2} f_2(z_1, z_2)\\ &\qquad + \check{1}\cot(\check{z}_2) f_2(z_1, z_2) \quad \text{in } E_1,\\ U(0, z_2) &= g_1(z_2),\\ U(1, z_2) &= g_2(z_2),\\ \partial_{z_2}U(z_1, 0) &= 0,\\ \partial_{z_2}U(z_1, 1) &= 0 \end{aligned} \tag{B-19}$$

has a unique solution $U(z) \in C^{k+1,\alpha}(E_1)$ that admits the estimate

$$\|U(z)\|_{C^{k+1,\alpha}(E_1)} \le C\sum_{i=1}^{2}\big(\|f_i(z)\|_{C^{k,\alpha}(E_1)} + \|g_i\|_{C^{k+1,\alpha}[0,1]}\big). \tag{B-20}$$

Proof. Again let $\rho = (x_2^2 + x_3^2)^{1/2}$. First, we consider the elliptic problem

$$\begin{aligned} (\partial^2_{x_1} + \partial^2_{x_2} + \partial^2_{x_3})U_1 + b_1(x_1, x_2, x_3)\partial_{x_2}U_1 &\\ + b_2(x_1, x_2, x_3)\partial_{x_3}U_1 &= \sum_{i=1}^{2} F_i(x_1, x_2, x_3) \quad \text{in } E_2,\\ U_1(0, x_2, x_3) &= g_1(\rho),\\ U_1(1, x_2, x_3) &= g_2(\rho),\\ \rho^{-1}(x_2\partial_{x_2} + x_3\partial_{x_3})U_1(x_1, x_2, x_3) &= 0 \quad \text{on } \rho = 1. \end{aligned} \tag{B-21}$$

where

$$\begin{aligned} b_i(x_1,x_2,x_3) &= (\check{1}\cot(\check{\rho})-\rho^{-1})x_{i+1}/\rho \quad \text{for } i=1,2,\\ F_1(x_1,x_2,x_3) &= \partial_{x_1} f_1(x_1,\rho),\\ F_2(x_1,x_2,x_3) &= \rho^{-1}(x_2\partial_{x_2}+x_3\partial_{x_3})f_2(x_1,\rho)\\ &\qquad\qquad\qquad + \check{1}\cot(\rho/X_0)f_2(x_1,\rho). \end{aligned} \tag{B-22}$$

We turn to the existence and uniqueness of the solution to the problem (B-21). According to the theory on second order elliptic equations with cornered boundaries and mixed type boundary conditions (see [Azzam 1980; 1981; Gilbarg and Hörmander 1980; Gilbarg and Trudinger 1983; Lieberman 1986; Vekua 1952]), we need to analyze the regularity of $b_i(x_1,x_2,x_3)$ and $F_i(x_1,x_2,x_3)$ for $i=1,2$.

First, it follows from Lemma B.1 that $b_i(x_1,x_2,x_3)$ satisfies

$$\|b_i(x_1,x_2,x_3)\|_{C^\alpha(E_2)} \le C. \tag{B-23}$$

In addition, $F_2(x_1,x_2,x_3)$ can be rewritten as

$$F_2(x_1,x_2,x_3) = \sum_{i=2}^{3}\partial_{x_i}\Big(\frac{x_i}{\rho}f_2(x_1,\rho)\Big) + (\check{1}\cot(\hat{\rho})-\rho^{-1})f_2(x_1,\rho). \tag{B-24}$$

Since $f_2(z_1,0)=0$, it follows from Lemma B.4 that

$$\sum_{i=2}^{3}\Big\|\frac{x_i}{\rho}f_2(x_1,\rho)\Big\|_{C^{k,\alpha}(E_2)} \le C\|f_2\|_{C^{k,\alpha}(E_1)} \qquad \text{for } k=0,1. \tag{B-25}$$

On the other hand, by Remark B.2, we have

$$\|(\check{1}\cot(\hat{\rho})-\rho^{-1})f_2(x_1,\rho)\|_{C^\alpha(E_2)} \le C\|f_2\|_{C^\alpha(E_1)}. \tag{B-26}$$

Because $g_i'(0)=g_i'(1)=0$ for $i=1,2$ and $\partial_{z_2}f_1(z_1,0)=0$, the compatible conditions at the corners for the problem (B-21) are satisfied. Moreover, by using (B-5) and (B-6) in Lemma B.3, we have

$$\begin{aligned} &\|g_i\|_{C^{j,\alpha}(E_4)} \le C\|g_i\|_{C^{j,\alpha}([0,1])} \quad \text{for } i=1,2 \text{ and } j=1,2,\\ &\|f_1\|_{C^{l,\alpha}(E_2)} \le C\|f_1\|_{C^{l,\alpha}(E_1)} \quad \text{for } l=0,1. \end{aligned} \tag{B-27}$$

Then by the results in [Lieberman 1986], the problem (B-21), which has the divergence form of a seconder order elliptic equation and the regularities of (B-23)–(B-26), has a unique solution $U_1(x_1,x_2,x_3)$ such that

$$\|U_1(x)\|_{C^{1,\alpha}(E_2)} \le C\sum_{i=1}^{2}\big(\|f_i(z)\|_{C^\alpha(E_1)} + \|g_i\|_{C^{1,\alpha}[0,1]}\big). \tag{B-28}$$

Furthermore, for $f_i(z) \in C^{1,\alpha}(E_1)$ and $g_i \in C^{2,\alpha}[0,1]$, due to the compatibility conditions at the corners, it follows from [Xin et al. 2009, Lemma A] that $U_1(x_1, x_2, x_3)$ is in $C^{2,\alpha}(E_2)$ and satisfies the estimate

$$\|U_1(x)\|_{C^{2,\alpha}(E_2)} \le C \sum_{i=1}^{2} \big(\|f_i(z)\|_{C^{1,\alpha}(E_1)} + \|g_i\|_{C^{2,\alpha}[0,1]}\big). \tag{B-29}$$

Next, we prove that the solution $U_1(x_1, x_2, x_3)$ in (B-21) is cylindrically symmetric. We use the transformation

$$\bar{x}_1 = x_1, \quad \bar{x}_2 = x_2 \cos\gamma_0 + x_3 \sin\gamma_0, \quad \bar{x}_3 = -x_2 \sin\gamma_0 + x_3 \cos\gamma_0,$$

with $\gamma_0 \in [0, 2\pi]$ being any fixed constant.

It is easy to verify that $U_1(\bar{x})$ also solves the problem (B-21). Thus, by the arbitrariness of γ_0, $U_1(x)$ is cylindrically symmetric with respect to (x_2, x_3), that is, $U_1(x)$ has the form $U_1(x) = U_1(x_1, \rho)$.

In addition, using the coordinate transformation

$$x_1 = z_1, \quad x_2 = z_2 \cos z_3, \quad x_3 = z_2 \sin z_3, \tag{B-30}$$

$U_1(x)$ can be expressed as $U_1 = U_1(z_1, z_2)$. Finally, it follows from (B-28)–(B-29) and Lemma B.3 that

$$\begin{aligned} \|U_1(z_1,z_2)\|_{C^{k,\alpha}(E_1)} &\le \|U(x_1,\rho)\|_{C^{k,\alpha}(E_2)} \\ &\le C \sum_{i=1}^{2} \big(\|f_i(z)\|_{C^{k-1,\alpha}(E_1)} + \|g_i\|_{C^{k,\alpha}[0,1]}\big) \qquad \text{for } k=1 \text{ or } k=2. \quad \square \end{aligned} \tag{B-31}$$

Lemma B.6. *If $F(z) \in C^{\alpha}(E_1)$, then the function*

$$U(z) = \frac{1}{\sin \check{z}_2} \int_0^{z_2} \sin \check{s} F(z_1, s) ds$$

satisfies

$$\|\partial_{z_2} U(z)\|_{C^{\alpha}(E_1)} \le C \|F(z)\|_{C^{\alpha}(E_1)}. \tag{B-32}$$

Further, if $F(z) \in C^{1,\alpha}(E_1)$ and $\partial_{z_2} F(z_1, 0) = 0$, then $U(z)$ satisfies

$$\|\partial_{z_2}^2 U(z)\|_{C^{\alpha}(E_1)} \le C \|F(z)\|_{C^{1,\alpha}(E_1)}. \tag{B-33}$$

Proof. First, $U(z)$ can be rewritten as

$$U(z) = \frac{1}{\sin \check{z}_2} \int_0^{z_2} \sin \check{s}(F(z_1,s) - F(z_1,0)) ds + X_0 \tan(\tfrac{1}{2}\check{z}_2) F(z_1, 0). \tag{B-34}$$

A direct computation yields

$$\|\partial_{z_2}^k \big(X_0 \tan(\tfrac{1}{2}\check{z}_2) F(z_1,0)\big)\|_{C^{\alpha}(E_1)} \le C \|F\|_{C^{\alpha}(E_1)} \quad \text{for } k=1 \text{ or } k=2. \tag{B-35}$$

Based on (B-34)–(B-35), in order to show Lemma B.6, it suffices to consider the case of $F(z_1, 0) = 0$ in (B-32) and $F(z_1, 0) = \partial_{z_2} F(z_1, 0) = 0$ in (B-33).

First, we prove (B-32) with $F(z_1, 0) = 0$.

It follows from a direct computation that

$$\begin{aligned}\partial_{z_2} U(z_1, z_2) &= F(z_1, z_2) - \frac{\cos \check{z}_2}{X_0 \sin^2 \check{z}_2} \int_0^{z_2} \sin(\check{s}) F(z_1, s) ds \\ &= \frac{1}{X_0 \sin^2 \check{z}_2} \int_0^{z_2} \big(\sin \check{z}_2 \cos \check{s}\, F(z_1, z_2) - \sin \check{s} \cos \check{z}_2\, F(z_1, s)\big) ds.\end{aligned}$$

This easily implies

$$\|\partial_{z_2} U(z)\|_{L^\infty(E_1)} \le C \|F(z)\|_{L^\infty(E_1)}. \tag{B-36}$$

We now estimate $[\partial_{z_2} U(z)]_\alpha$ in E_1.

For any two different points (z_{11}, z_{21}) and (z_{12}, z_{22}) in E_1, we may assume without loss of generality that $z_{21} \ge z_{22}$. Then

$$\partial_{z_2} U(z_{11}, z_{21}) - \partial_{z_2} U(z_{12}, z_{22}) = K_1 + K_2 + K_3, \tag{B-37}$$

where

$$\begin{aligned}K_1 &= \frac{1}{X_0 \sin^2 \check{z}_{21}} \int_0^{z_{21}} \big(\cos(\check{s})(\sin \check{z}_{21} F(z_{11}, z_{21}) - \sin \check{z}_{22} F(z_{12}, z_{22})) \\ &\qquad\qquad\qquad - \sin \check{s}\big(\cos(\check{z}_{21}) F(z_{11}, s) \\ &\qquad\qquad\qquad\qquad - \cos(\check{z}_{22}) F(z_{12}, s)\big)\big) ds, \\ K_2 &= \frac{1}{X_0 \sin \check{z}_{21}} \int_{z_{22}}^{z_{21}} \big(\sin \check{z}_{22} \cos(\check{s}) F(z_{12}, z_{22}) - \sin \check{s} \cos(\check{z}_{22}) F(z_{12}, s)\big) ds, \\ K_3 &= \frac{(\sin \check{z}_{22} - \sin \check{z}_{21})\big(\sin \check{z}_{22} + \sin \check{z}_{21}\big)}{X_0 (\sin \check{z}_{21} \sin \check{z}_{22})^2} \\ &\qquad \times \int_0^{z_{22}} \big(\sin \check{z}_{22} \cos(\check{s}) F(z_{12}, z_{22}) - \sin \check{s} \cos(\check{z}_{22}) F(z_{12}, s)\big) ds.\end{aligned}$$

It follows from $F(z_{11}, 0) = 0$ and a direct computation that

$$\begin{aligned}|K_1| &\le \frac{1}{X_0 \sin^2 \check{z}_{21}} \big(|\sin \check{z}_{21} - \sin \check{z}_{22}| z_{21}^{1+\alpha} [F]_\alpha \\ &\qquad\qquad + \sin(\check{z}_{22})((z_{11} - z_{12})^2 + (z_{21} - z_{22})^2)^{\alpha/2} z_{21} [F]_\alpha \\ &\qquad + \sin(\check{z}_{21}) |\cos \check{z}_{21} - \cos \check{z}_{22}| z_{21}^{1+\alpha} [F]_\alpha + \sin(\check{z}_{21}) z_{21} |z_{11} - z_{12}|^\alpha [F]_\alpha\big) \\ &\le C [F]_\alpha ((z_{11} - z_{12})^2 + (z_{21} - z_{22})^2)^{\alpha/2},\end{aligned}$$

$$\begin{aligned}
|K_2| &\le \frac{1}{\sin^2 \check{z}_{21}}(\sin(\check{z}_{22})|\sin \check{z}_{21} - \sin \check{z}_{22}| z_{22}^{\alpha}[F]_\alpha + |\cos(\check{z}_{21}) - \cos(\check{z}_{22})| z_{21}^{\alpha}[F]_\alpha)\\
&\le C[F]_\alpha((z_{11}-z_{12})^2 + (z_{21}-z_{22})^2)^{\alpha/2},\\
|K_3| &\le \frac{|\sin \check{z}_{22} - \sin \check{z}_{21}|(\sin \check{z}_{22} + \sin \check{z}_{21})}{X_0(\sin \check{z}_{21}\sin \check{z}_{22})^2}\sin(\check{z}_{22}) z_{22}^{1+\alpha}[F]_\alpha\\
&\le C[F]_\alpha\big((z_{11}-z_{12})^2 + (z_{21}-z_{22})^2\big)^{\alpha/2}.
\end{aligned}$$

This implies

$$[\partial_{z_2} U(z)]_\alpha \le C[F]_\alpha. \tag{B-38}$$

Combining (B-38) with (B-35) and (B-36) yields (B-32).

Second, we prove (B-33) in the case of $F(z_1, 0) = \partial_{z_2} F(z_1, 0) = 0$. Note that

$$\partial_{z_2}^2 U(z) = \partial_{z_2} F(z_1, z_2) - \check{1}\cot(\check{z}_2) F(z_1, z_2) + \frac{1+\cos^2 \check{z}_2}{X_0^2 \sin^3 \check{z}_2}\int_0^{z_2} \sin(\check{s}) F(z_1, s)\, ds.$$

By (B-7) of Lemma B.3, we have

$$\|\partial_{z_2} F(z_1, z_2) - \check{1}\cot(\check{z}_2) F(z_1, z_2)\|_{C^\alpha(E_1)} \le C\|F\|_{C^{1,\alpha}(E_1)}. \tag{B-39}$$

In addition, a direct computation yields

$$\begin{aligned}
\frac{1+\cos^2 \check{z}_{21}}{X_0^2 \sin^3 \check{z}_{21}}\int_0^{z_{21}} \sin(\check{s}) F(z_{11}, s)\, ds - \frac{1+\cos^2 \check{z}_{22}}{X_0^2 \sin^3 \check{z}_{22}}\int_0^{z_{22}} \sin(\check{s}) F(z_{12}, s)\, ds&\\
= K_4 + K_5 + K_6 + K_7&,
\end{aligned}$$

where

$$\begin{aligned}
K_4 &= \frac{1+\cos^2(\check{z}_{21})}{\underline{X_0^2 \sin^3 \check{z}_{21}}}\int_0^{z_{21}} \sin(\check{s}) s\Big(\int_0^1 (\partial_{z_2} F(z_{11}, \theta s) - \partial_{z_2} F(z_{12}, \theta s))\, d\theta\Big) ds,\\
K_5 &= \frac{1+\cos^2 \check{z}_{21}}{X_0^2 \sin^3 \check{z}_{21}}\int_{z_{22}}^{z_{21}} \sin(\check{s}) s\Big(\int_0^1 \partial_{z_2} F(z_{12}, \theta s)\, d\theta\Big) ds,\\
K_6 &= \frac{(\cos \check{z}_{21} - \cos \check{z}_{22})(\cos \check{z}_{21} + \cos \check{z}_{22})}{X_0^2 \sin^3 \check{z}_{21}}\\
&\qquad\qquad \times \int_0^{z_{22}} \sin(\check{s}) s\Big(\int_0^1 \partial_{z_2} F(z_{12}, \theta s)\, d\theta\Big) ds,\\
K_7 &= \frac{(1+\cos^2 \check{z}_{22})(\sin \check{z}_{22} - \sin \check{z}_{21})(\sin^2 \check{z}_{22} + \sin \check{z}_{22}\sin \check{z}_{21} + \sin^2 \check{z}_{21})}{X_0^2(\sin \check{z}_{21}\sin \check{z}_{22})^3}\\
&\qquad\qquad \times \int_0^{z_{22}} \sin(\check{s}) s\Big(\int_0^1 \partial_{z_2} F(z_{12}, \theta s)\, d\theta\Big) ds.
\end{aligned}$$

Hence, by using $F(z_1,0)=\partial_{z_2}F(z_1,0)=0$, we have

$$\begin{aligned}
|K_4| &\le \frac{\sin(\check{z}_{21})z_{21}^2}{X_0^2\sin^3\check{z}_{21}}[\partial_{z_2}F]_\alpha|z_{11}-z_{12}|^\alpha \le C[\partial_{z_2}F]_\alpha|z_{11}-z_{12}|^\alpha,\\
|K_5| &\le \frac{\sin(\check{z}_{21})z_{21}|z_{21}-z_{22}|}{X_0^2\sin^3\check{z}_{21}}[\partial_{z_2}F]_\alpha z_{12}^\alpha \le C[\partial_{z_2}F]_\alpha|z_{21}-z_{22}|^\alpha,\\
|K_6| &\le \frac{2\sin(\frac{1}{2}(\check{z}_{21}-\check{z}_{22}))}{X_0^2\sin^3\check{z}_{21}}[\partial_{z_2}F]_\alpha z_{22}^{2+\alpha}\sin(\check{z}_{22}) \le C[\partial_{z_2}F]_\alpha|z_{21}-z_{22}|^\alpha,\\
|K_7| &\le \frac{3\sin^2\check{z}_{21}|\sin\check{z}_{22}-\sin\check{z}_{21}|}{X_0^2(\sin\check{z}_{21}\sin\check{z}_{22})^3}\sin(\check{z}_{22})[\partial_{z_2}F]_\alpha z_{22}^{2+\alpha} \le C[\partial_{z_2}F]_\alpha|z_{21}-z_{22}|^\alpha.
\end{aligned}$$

This leads to

$$[\partial_{z_2}^2U(z)]_\alpha \le C[\partial_{z_2}F]_\alpha. \tag{B-40}$$

Combining (B-40) with (B-39) and (B-32), we complete the proof of (B-33). Therefore, the proof of Lemma B.6 is completed. □

Lemma B.7. *The problem*

$$\begin{aligned}
\partial_1 w &= a_1\partial_2 P + F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi) \quad \textit{in } E_1,\\
\partial_2 w + \check{1}\cot(\check{z}_2)w &= F_2(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi) - a_2\partial_1 P \quad \textit{in } E_1,\\
w(z_1,0) &= 0
\end{aligned} \tag{B-41}$$

is well posed if

$$\begin{aligned}
\big(\partial_{z_2}+\check{1}\cot(\check{z}_2)\big)&(a_1\partial_2P + F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi))\\
&- \partial_{z_1}(F_2(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi) - a_2\partial_1 P) = 0,\\
&(a_1\partial_2 P + F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi))(z_1,0) = 0.
\end{aligned}$$

Proof. Define w_i for $i=1,2$ as

$$\begin{aligned}
\partial_1 w_1 &= a_1\partial_2 P + F_1(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi) \quad \text{in } E_1,\\
w_1(1,z_2) &= \frac{1}{\sin(\check{z}_2)}\int_0^{z_2}\sin(\check{s})(F_2(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi) - a_2\partial_1 P)(1,s)ds
\end{aligned} \tag{B-42}$$

and

$$\begin{aligned}
\partial_2 w_2 + \check{1}\cot(\check{z}_2)w_2 &= F_2(\tilde{S}, P(q,\tilde{S}), V_1, V_2;\xi) - a_2\partial_1 P \quad \text{in } E_1,\\
w_2(z_1,0) &= 0.
\end{aligned} \tag{B-43}$$

Obviously, w_1 and w_2 can be determined uniquely.

From (B-43), w_2 has the expression

(B-44)
$$w_2(z_1,z_2)=\frac{1}{\sin\check{z}_2}\int_0^{z_2}\sin(\check{s})(F_2(\tilde{S},P(q,\tilde{S}),V_1,V_2;\xi)-a_2\partial_1 P)(z_1,s)ds.$$

Also it follows from (B-42) and the second equality in B.7 that $w_1(z_1,0)=0$. By (B-43) and (B-44), we arrive at

$$w_2(z_1,0)=0 \quad\text{and}\quad w_1(1,z_2)=w_2(1,z_2).$$

Next, we show $w_1=w_2$ in E_1.

Note that $(\partial_{z_2}+\check{1}\cot(\check{z}_2))$ times the first equation in (B-42) minus ∂_{z_1} applied to the first equation in (B-43) yields

$$\begin{aligned}&(\partial_{z_2}+\check{1}\cot(\check{z}_2))\partial_{z_1}(w_1-w_2)=0 \quad \text{in } E_1,\\ &\qquad w_1(z_1,0)=w_2(z_1,0)=0,\quad w_1(1,z_2)=w_2(1,z_2).\end{aligned}\tag{B-45}$$

One concludes easily that $w_1=w_2$ holds true in E_1, completing the proof. □

Lemma B.8. *Let* $(\tilde{S}_1,\tilde{P}_1,V_{11},V_{21})$ *and* $(\tilde{S}_2,\tilde{P}_2,V_{12},V_{22})$ *be in* Ξ_δ *such that*

$$\begin{aligned}T(\tilde{S}_1,\tilde{P}_1,V_{11},V_{21})&=(S_1,P_1,U_{11},U_{21}),\\ T(\tilde{S}_2,\tilde{P}_2,V_{12},V_{22})&=(S_2,P_2,U_{12},U_{22}),\end{aligned}$$

where the mapping T *is defined in* (3-37). *Denote by* $\xi_1(z_2)$ *and* $\xi_2(z_2)$ *the corresponding approximate shocks by solving* (3-7). *Define* W_i *for* $i=1,2,3,4$ *as in Section 3 with respect to* (S_1,P_1,U_{11},U_{21}) *and* (S_2,P_2,U_{12},U_{22}), *and define* $\widetilde{W}_i(i=1,2,3,4)$ *with respect to* $(\tilde{S}_1,\tilde{P}_1,V_{11},V_{21})$ *and* $(\tilde{S}_2,\tilde{P}_2,V_{12},V_{22})$.

Set

$$W_5=\frac{U_{21}}{U_{11}}-\frac{U_{22}}{U_{12}},\quad \widetilde{W}_5=\frac{V_{21}}{V_{11}}-\frac{V_{22}}{V_{12}},\quad W_6=\xi_1(z_2)-\xi_2(z_2).$$

Then under the assumptions of Theorem 3.1, we have

$$\|W_1\|_{C^{1,\alpha}(E_+)}\le C\Big(\delta\sum_{i=2}^{4}\|\widetilde{W}_i\|_{C^{1,\alpha}(E_+)}+\check{1}\|W_6\|_{C^{1,\alpha}(E_+)}\Big).\tag{B-46}$$

Proof. In the coordinate $z=(z_1,z_2)$, the characteristics $z_2^1(s;z)$ and $z_2^2(s;z)$, which go through the point (z_1,z_2) and correspond to the vector fields (V_{11},V_{21}) and (V_{12},V_{22}) in the right hand side of (2-28) respectively, can be defined as

$$\begin{aligned}&\frac{dz_2^i(s;z)}{ds}=-\frac{X_0(X_0+1-\xi_1(z_2^i))}{A_i}V_{2i}\big(\xi_1(z_2^i)+s(X_0+1-\xi_1(z_2^i)),z_2^i\big),\\ &z_2^i(z_1;z)=z_2,\quad z_2^1(0;z)=\beta,\quad z_2^2(0,z)=\tilde{\beta}\quad\text{for } i=1,2,\end{aligned}$$

where

$$A_i = (\xi_i(z_2^i) + s(X_0 + 1 - \xi_i(z_2^i)))V_{1i} + V_{2i}X_0(1-s)\xi_i'(z_2^i) \quad \text{for } i = 1, 2.$$

Set $l(s; z) = z_2^1(s; z) - z_2^2(s; z)$. Noting that $z_2^1(0; z) = \beta$ and $z_2^2(0; z) = \tilde{\beta}$, we have

$$\begin{aligned} \frac{dl}{ds} &= O(\delta)l + O(\delta)\widetilde{W}_3(s, z_2^1) + O(1)\widetilde{W}_4(s, z_2^1) \\ &\qquad\qquad + O(\delta)W_6(z_2^1) + O(\delta^2)W_6'(z_2^1), \\ l(0; z) &= \beta - \tilde{\beta}, \qquad l(z_1; z) = 0, \end{aligned} \tag{B-47}$$

where we point out that the coefficients of $l(t; z)$ are in $C^{1,\alpha}$, which will be used to derive the $C^{1,\alpha}$ estimate of $\beta - \tilde{\beta}$.

By (B-47), we can arrive at

$$\|\beta - \tilde{\beta}\|_{L^\infty} \le \|l\|_{L^\infty} \le C(\delta\|\widetilde{W}_3\|_{L^\infty} + \|\widetilde{W}_4\|_{L^\infty} + \delta\|W_6\|_{L^\infty} + \delta^2\|W_6'(z_2)\|_{L^\infty}).$$

Note that

$$z_2^1(s; z) = -\int_0^s \frac{X_0(X_0 + 1 - \xi_1(z_2^1))}{A_1} V_{21}\big(\xi_1(z_2^1) + t(X_0 + 1 - \xi_1(z_2^1)), z_2^1\big)dt + \beta,$$

which implies, in particular, that

$$z_2 = -\int_0^{z_1} \frac{X_0(X_0 + 1 - \xi_1(z_{21}))}{A_1} V_{21}\big(\xi_1(z_2^1) + t(X_0 + 1 - \xi_1(z_{21})), z_2^1\big))dt + \beta.$$

Similar expressions hold for $z_2^2(s; z)$ and z_2 with β replaced by $\tilde{\beta}$. Thus, we may obtain

$$\begin{aligned} \beta - \tilde{\beta} &= \int_0^{z_1} \big(O(\delta)\widetilde{W}_3(t, z_2^1) + O(1)\widetilde{W}_4(t, z_2^1) + O(\delta)W_6(z_2^1) \\ &\qquad\qquad + O(\delta^2)W_6'(z_2^1) + O(\delta)l(t; z)\big)dt, \\ l(s; z) &= \int_{z_1}^{s} \big(O(\delta)\widetilde{W}_3(t, z_2^1) + O(1)\widetilde{W}_4(t, z_2^1) + O(\delta)W_6(z_2^1) \\ &\qquad\qquad + O(\delta^2)W_6'(z_2^1) + O(\delta)l(t; z)\big)dt \end{aligned} \tag{B-48}$$

and

$$\|\partial_{z_1}(\beta, \tilde{\beta})\|_{C^{1,\alpha}} \le C\varepsilon, \quad \|\partial_{z_2}(\beta, \tilde{\beta})\|_{C^{1,\alpha}} \le C\varepsilon. \tag{B-49}$$

In addition, it follows from (B-47) and (B-48) that

$$\|\beta - \tilde{\beta}\|_{C^{1,\alpha}} \le C(\delta\|\widetilde{W}_3\|_{C^{1,\alpha}} + \|\widetilde{W}_4\|_{C^{1,\alpha}} + \delta\|W_6\|_{C^{2,\alpha}}). \tag{B-50}$$

This, together with (2-28) and the characteristics method, yields

$$\begin{aligned} W_1(z_1,z_2) &= W_1(0,\beta(z_1,z_2)) + O(\delta)\big(\beta(z_1,z_2)-\tilde{\beta}(z_1,z_2)\big),\\ W_1(0,z_2) &= O(\delta^2)\widetilde{W}_2(0,z_2)+O(\delta^2)\widetilde{W}_3(0,z_2)+O(\delta)\widetilde{W}_4(0,z_2)\\ &\qquad + O(\check{1})W_6(z_2), \end{aligned} \tag{B-51}$$

and

$$\begin{aligned} &\|\beta(z)-\widetilde{\beta(z)}\|_{C^{1,\alpha}(E_+)}\\ &\qquad \le C\big(\delta\|\widetilde{W}_3\|_{C^{1,\alpha}(E_+)}+\|\widetilde{W}_4\|_{C^{1,\alpha}(E_+)}+\delta\|W_6\|_{C^{2,\alpha}[0,1]}\big). \end{aligned} \tag{B-52}$$

Combining (B-52) with (B-51) yields (B-46), proving Lemma B.8. □

Remark B.9. If we choose

$$(\tilde{S}_2,\tilde{P}_2,V_{12},V_{22})=(S_2,P_2,U_{12},U_{22})=(S_0^+,\hat{P}_0^+,\hat{U}_0^+,0),$$

where $(S_0^+,\hat{P}_0^+,\hat{U}_0^+,0)$ is the background solution given in Appendix A with the exit pressure P_e, then, by the $C^{3,\alpha}$ regularity of $(S_0^+,\hat{P}_0^+,\hat{U}_0^+,0)$, we can conclude that

$$\|W_1\|_{C^{2,\alpha}(E_+)}\le C\Big(\delta\sum_{i=2}^{4}\|\widetilde{W}_i\|_{C^{2,\alpha}(E_+)}+\check{1}\|W_6\|_{C^{3,\alpha}(E_+)}\Big). \tag{B-53}$$

In fact, in this case, the corresponding coefficients of $l(s;z)$ in (B-47) and (B-48) are in $C^{2,\alpha}$. As in (B-50), we can derive that

$$\begin{aligned} &\|\beta(z)-\widetilde{\beta(z)}\|_{C^{2,\alpha}(E_+)}\\ &\qquad \le C(\delta\|\widetilde{W}_3\|_{C^{2,\alpha}(E_+)}+\|\widetilde{W}_4\|_{C^{2,\alpha}(E_+)}+\delta\|W_6\|_{C^{3,\alpha}[0,1]}). \end{aligned} \tag{B-54}$$

Subsequently, (B-53) can be shown by using (B-51) and (B-54).

Appendix C.

Here, for the problem (1-1) with (1-2)–(1-5), we give a detailed discussion of the higher order compatibility conditions on the nozzle wall and address the crucial difficulty in obtaining $C^{3,\alpha}$ regularities of solutions — that is, the appearance of the source terms in (2-8).

Due to the right hand conditions (1-2), the following expressions hold:

$$\begin{aligned} G_1(\rho,U,S)&\equiv[\rho U_1][\rho U_2^2+P]-\rho^2U_1U_2^2=0,\\ G_2(\rho,U,S)&\equiv[\rho U_1^2+P][\rho U_2^2+P]-(\rho U_1U_2)^2=0,\\ G_3(\rho,U,S)&\equiv[(\rho e+\tfrac{1}{2}\rho|U|^2+P)U_1][\rho U_2^2+P]\\ &\qquad -\rho U_1(\rho e+\tfrac{1}{2}\rho|U|^2+P)U_2^2=0. \end{aligned} \tag{C-1}$$

Since $U_2 = \partial_{z_2} P = \partial_{z_2} S = \partial_{z_2} \rho = \partial_{z_2} U_1 = 0$, at the point $M_0 := (z_1, z_2) = (0, 1)$, taking the tangential derivatives of second order and third order respectively along the shock surface yields at M_0

$$\begin{aligned}
&(\rho\partial_{z_2}^2 U_1 + U_1\partial_{z_2}^2\rho)[P] - 2\rho^2 U_1(\partial_{z_2} U_2)^2 = 0,\\
&(2\rho U_1\partial_{z_2}^2 U_1 + U_1^2\partial_{z_2}^2\rho + \partial_{z_2}^2 P)][P] - 2\rho^2 U_1^2(\partial_{z_2} U_2)^2 = 0,\\
&\bigg(\Big(\frac{\gamma}{\gamma-1}P + \tfrac{1}{2}\rho U_1^2\Big)\partial_{z_2}^2 U_1\\
&\quad + U_1\Big(\frac{\gamma}{\gamma-1}\partial_{z_2}^2 P + \tfrac{1}{2}U_1^2\partial_{z_2}^2\rho + \rho U_1\partial_{z_2}^2 U_1 + \rho(\partial_{z_2} U_2)^2\Big)\bigg)[P]\\
&\quad - 2\rho U_1\Big(\frac{\gamma}{\gamma-1}P + \tfrac{1}{2}\rho U_1^2\Big)(\partial_{z_2} U_2)^2 = 0.
\end{aligned}\tag{C-2}$$

and

$$\begin{aligned}
&(\rho\partial_{z_2}^3 U_1 + U_1\partial_{z_2}^3\rho)[P] - 6\rho^2 U_1\partial_{z_2} U_2\partial_{z_2}^2 U_2 = 0,\\
&(2\rho U_1\partial_{z_2}^3 U_1 + U_1^2\partial_{z_2}^3\rho + \partial_{z_2}^3 P)][P] - 6\rho^2 U_1^2\partial_{z_2} U_2\partial_{z_2}^2 U_2 = 0,\\
&\bigg(\Big(\frac{\gamma}{\gamma-1}P + \tfrac{1}{2}\rho U_1^2\Big)\partial_{z_2}^3 U_1\\
&\quad + U_1\Big(\frac{\gamma}{\gamma-1}\partial_{z_2}^3 P + \tfrac{1}{2}U_1^3\partial_{z_2}^3\rho + \rho U_1\partial_{z_2}^3 U_1 + 3\rho\partial_{z_2} U_2\partial_{z_2}^2 U_2\Big)\bigg)[P]\\
&\quad - 6\rho U_1\Big(\frac{\gamma}{\gamma-1}P + \tfrac{1}{2}\rho U_1^2\Big)\partial_{z_2} U_2\partial_{z_2}^2 U_2 = 0.
\end{aligned}\tag{C-3}$$

From the first two equations in (C-2) and (C-3), we have at M_0

$$\partial_{z_2}^2 P + \rho U_1\partial_{z_2}^2 U_1 = 0, \quad \partial_{z_2}^3 P + \rho U_1\partial_{z_2}^3 U_1 = 0.\tag{C-4}$$

It follows from the first and the third equations in (C-2) and (C-3) that at M_0

$$\begin{aligned}
&\Big(\frac{\gamma}{\gamma-1}P\partial_{z_2}^2 U_1 + U_1\Big(\frac{\gamma}{\gamma-1}\partial_{z_2}^2 P + \rho U_1\partial_{z_2}^2 U_1 + \rho(\partial_{z_2} U_2)^2\Big)\Big)[P]\\
&\qquad - \frac{2\gamma}{\gamma-1}\rho U_1 P(\partial_{z_2} U_2)^2 = 0,\\
&\Big(\frac{\gamma}{\gamma-1}P\partial_{z_2}^3 U_1 + U_1\Big(\frac{\gamma}{\gamma-1}\partial_{z_2}^3 P + \rho U_1\partial_{z_2}^3 U_1 + 3\rho\partial_{z_2} U_2\partial_{z_2}^2 U_2\Big)\Big)[P]\\
&\qquad - \frac{6\gamma}{\gamma-1}\rho U_1 P\partial_{z_2} U_2\partial_{z_2}^2 U_2 = 0.
\end{aligned}\tag{C-5}$$

Since $\partial_{z_2}^2 U_2 + \check{1}\cot(\check{1})\partial_{z_2} U_2 = 0$ at M_0 due to (3-5) and (3-6) and by the expression of F_2 in (2-26), this together with (C-4) and (C-5) yields

$$Q\big(\partial_{z_2}^3 P + \check{1}\cot(\check{1})\partial_{z_2}^2 P\big) = \Big(\frac{4\gamma}{\gamma-1}\rho U_1 P - 2\rho U_1[P]\Big)\partial_{z_2} U_2\partial_{z_2}^2 U_2 \quad \text{at } M_0\tag{C-6}$$

where

$$Q=\frac{\rho U_1^2-\gamma P}{(\gamma-1)\rho U_1}<0. \tag{C-7}$$

On the other hand, it follows from the first equation in (2-26), the expressions of F_1 and F_2, and (3-6) and (3-7) that $\partial_{z_2}P=\partial_{z_2}F_2=F_1=0$ at M_0 and

$$a_1(\partial_{z_2}^3P+\check{1}\cot(\check{1})\partial_{z_2}^2P)=-\partial_{z_2}^2F_1-\check{1}\partial_{z_2}F_1 \quad \text{at } M_0. \tag{C-8}$$

Also, since

$$\xi^{(3)}(1)+\check{1}\cot(\check{1})\xi^{(2)}(1)=0 \quad \text{and} \quad \partial_{z_2}^2U_2+\check{1}\cot(\check{1})\partial_{z_2}U_2=0$$

at M_0, Equation (C-8) yields

$$\frac{X_0(X_0+1-\xi)}{\xi\rho U_1^2}(\partial_{z_2}^3P+\check{1}\cot(\check{1})\partial_{z_2}^2P)=\frac{2(\partial_{z_2}U_2)^2(X_0+1-\xi)}{\xi U_1^2}\cot(\check{1})$$

at M_0, so that

$$\partial_{z_2}^3P+\check{1}\cot(\check{1})\partial_{z_2}^2P=\check{2}\rho\cot(\check{1})(\partial_{z_2}U_2)^2 \tag{C-9}$$

at M_0. Thus, it follows from (C-9) and (C-6) that

$$Q+\frac{2\gamma}{\gamma-1}U_1P-U_1[P]=0 \quad \text{or} \quad \partial_{z_2}U_2=0 \quad \text{at } M_0. \tag{C-10}$$

Meanwhile, in the general case,

$$\begin{aligned}Q+\frac{2\gamma}{\gamma-1}U_1P-U_1[P]&=\frac{1}{(\gamma-1)\rho U_1}(\rho U_1^2-\gamma P)+\frac{2\gamma}{\gamma-1}U_1P-U_1[P]\\&=\frac{1}{\gamma-1}U_1+U_1\hat{P}_0^-+\frac{\gamma+1}{\gamma-1}U_1P-\frac{\gamma}{(\gamma-1)\rho U_1}P\\&=\frac{1}{\gamma-1}U_1+U_1\hat{P}_0^-+\frac{P}{(\gamma-1)\rho U_1}((\gamma+1)\rho U_1^2-\gamma)\neq 0.\end{aligned}$$

Thus, combining this with (C-9) yields $\partial_{z_2}U_2=0$ at M_0 if the solution is in $C^{3,\alpha}$.

On the other hand, it follows from (2-26) that

$$\frac{\partial_{z_2}U_2}{U_1}+\frac{\xi}{X_0(X_0+1-\xi)}\partial_{z_1}(P-\tilde{P}_0^+)=0,$$

where $\tilde{P}_0^+$ denotes the background pressure when the shock position lies at $r=\xi(1)$. However, it seems to be rather difficult to show $\partial_{z_1}(P-\tilde{P}_0^+)=0$ at the point M_0 in general (although $\partial_{z_2}(P-\tilde{P}_0^+)=0$ there by (3-6)).

References

[Azzam 1980] A. Azzam, "On Dirichlet's problem for elliptic equations in sectionally smooth n-dimensional domains", *SIAM J. Math. Anal.* **11** (1980), 248–253. MR 82k:35032a Zbl 0439.35026

[Azzam 1981] A. Azzam, "Smoothness properties of solutions of mixed boundary value problems for elliptic equations in sectionally smooth n-dimensional domains", *Ann. Polon. Math.* **40**:1 (1981), 81–93. MR 83i:35055 Zbl 0485.35013

[Bers 1950] L. Bers, "Partial differential equations and generalized analytic functions", *Proc. Nat. Acad. Sci. U. S. A.* **36** (1950), 130–136. MR 12,173d Zbl 0036.05301

[Bers 1951] L. Bers, "Partial differential equations and generalized analytic functions, II", *Proc. Nat. Acad. Sci. U. S. A.* **37** (1951), 42–47. MR 13,352c Zbl 0042.08803

[Čanić et al. 2000] S. Čanić, B. L. Keyfitz, and G. M. Lieberman, "A proof of existence of perturbed steady transonic shocks via a free boundary problem", *Comm. Pure Appl. Math.* **53**:4 (2000), 484–511. MR 2001m:76056 Zbl 1017.76040

[Chen 2008] S. Chen, "Transonic shocks in 3-D compressible flow passing a duct with a general section for Euler systems", *Trans. Amer. Math. Soc.* **360**:10 (2008), 5265–5289. MR 2009d:35216 Zbl 1158.35064

[Chen and Feldman 2003] G.-Q. Chen and M. Feldman, "Multidimensional transonic shocks and free boundary problems for nonlinear equations of mixed type", *J. Amer. Math. Soc.* **16**:3 (2003), 461–494. MR 2004d:35182 Zbl 1015.35075

[Chen et al. 2006] G.-Q. Chen, J. Chen, and K. Song, "Transonic nozzle flows and free boundary problems for the full Euler equations", *J. Differential Equations* **229**:1 (2006), 92–120. MR 2007j:35124 Zbl 1142.35510

[Chen et al. 2007] G.-Q. Chen, J. Chen, and M. Feldman, "Transonic shocks and free boundary problems for the full Euler equations in infinite nozzles", *J. Math. Pures Appl.* (9) **88**:2 (2007), 191–218. MR 2008k:35371 Zbl 1131.35061

[Courant and Friedrichs 1948] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience, New York, 1948. MR 10,637c Zbl 0041.11302

[Embid et al. 1984] P. Embid, J. Goodman, and A. Majda, "Multiple steady states for 1-D transonic flow", *SIAM J. Sci. Statist. Comput.* **5**:1 (1984), 21–41. MR 86a:76029 Zbl 0573.76055

[Gilbarg and Hörmander 1980] D. Gilbarg and L. Hörmander, "Intermediate Schauder estimates", *Arch. Rational Mech. Anal.* **74**:4 (1980), 297–318. MR 82a:35038 Zbl 0454.35022

[Gilbarg and Trudinger 1983] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, 2nd ed., Grundlehren der Mathematischen Wissenschaften **224**, Springer, Berlin, 1983. MR 86c:35035 Zbl 0562.35001

[Glaz and Liu 1984] H. M. Glaz and T.-P. Liu, "The asymptotic analysis of wave interactions and numerical calculations of transonic nozzle flow", *Adv. in Appl. Math.* **5**:2 (1984), 111–146. MR 85j:76019 Zbl 0598.76065

[Kuz'min 2002] A. G. Kuz'min, *Boundary Value problems for Transonic Flow*, Wiley, New York, 2002.

[Li et al. 2009a] J. Li, Z. Xin, and H. Yin, "A free boundary value problem for the full Euler system and 2-D transonic shock in a large variable nozzle", *Math. Res. Lett.* **16**:5 (2009), 777–796. MR 2576697

[Li et al. 2009b] J. Li, Z. Xin, and H. Yin, “On transonic shocks in a nozzle with variable end pressures”, *Comm. Math. Phys.* **291**:1 (2009), 111–150. MR 2530157 Zbl 05655506

[Li et al. 2010a] J. Li, Z. Xin, and H. Yin, “On transonic shocks in a conic divergent nozzle with axi-symmetric exit pressures”, *J. Differential Equations* **248**:3 (2010), 423–469. MR 2557901 Zbl 05658643

[Li et al. 2010b] J. Li, Z. Xin, and H. Yin, “The uniqueness of a 3-D transonic shock solution in a finite nozzle with asixymmetric exit pressure”, preprint, IMS at Nanjing University, 2010.

[Lieberman 1986] G. M. Lieberman, “Mixed boundary value problems for elliptic and parabolic differential equations of second order”, *J. Math. Anal. Appl.* **113**:2 (1986), 422–440. MR 87h:35081 Zbl 0609.35021

[Liu 1982a] T. P. Liu, “Nonlinear stability and instability of transonic flows through a nozzle”, *Comm. Math. Phys.* **83**:2 (1982), 243–260. MR 83f:35014 Zbl 0576.76053

[Liu 1982b] T. P. Liu, “Transonic gas flow in a duct of varying area”, *Arch. Rational Mech. Anal.* **80**:1 (1982), 1–18. MR 83h:76050 Zbl 0503.76076

[Morawetz 1994] C. S. Morawetz, “Potential theory for regular and Mach reflection of a shock at a wedge”, *Comm. Pure Appl. Math.* **47**:5 (1994), 593–624. MR 95g:76030 Zbl 0807.76033

[Vekua 1952] I. N. Vekua, “Systems of differential equations of the first order of elliptic type and boundary value problems, with an application to the theory of shells”, *Mat. Sbornik N. S.* **31(73)** (1952), 217–314. In Russian. MR 15,230a

[Xin and Yin 2005] Z. Xin and H. Yin, “Transonic shock in a nozzle, I: Two-dimensional case”, *Comm. Pure Appl. Math.* **58**:8 (2005), 999–1050. MR 2006c:76079

[Xin and Yin 2008a] Z. Xin and H. Yin, “Three-dimensional transonic shocks in a nozzle”, *Pacific J. Math.* **236**:1 (2008), 139–193. MR 2009a:35170 Zbl 05366344

[Xin and Yin 2008b] Z. Xin and H. Yin, “The transonic shock in a nozzle, 2-D and 3-D complete Euler systems”, *J. Differential Equations* **245** (2008), 1014–1085. MR 2009m:35319 Zbl 1165.35031

[Xin et al. 2009] Z. Xin, W. Yan, and H. Yin, “Transonic shock problem for the Euler system in a nozzle”, *Arch. Ration. Mech. Anal.* **194**:1 (2009), 1–47. MR 2533922 Zbl 05640831

[Yuan 2006] H. Yuan, “On transonic shocks in two-dimensional variable-area ducts for steady Euler system”, *SIAM J. Math. Anal.* **38**:4 (2006), 1343–1370. MR 2008i:35162 Zbl 1121.35081

[Zheng 2003] Y. Zheng, “A global solution to a two-dimensional Riemann problem involving shocks as free boundaries”, *Acta Math. Appl. Sin. Engl. Ser.* **19**:4 (2003), 559–572. MR 2004m:35182 Zbl 1079.35068

[Zheng 2006] Y. Zheng, “Two-dimensional regular shock reflection for the pressure gradient system of conservation laws”, *Acta Math. Appl. Sin. Engl. Ser.* **22**:2 (2006), 177–210. MR 2007b:35229 Zbl 1106.35034

Received June 8, 2009.

JUN LI
DEPARTMENT OF MATHEMATICS AND IMS
NANJING UNIVERSITY
NANJING 210093
CHINA
lijun@nju.edu.cn

ZHOUPING XIN
DEPARTMENT OF MATHEMATICS AND IMS
CHINESE UNIVERSITY OF HONG KONG
SHATIN, N.T.
HONG KONG
zpxin@ims.cuhk.edu.hk

HUICHENG YIN
DEPARTMENT OF MATHEMATICS AND IMS
NANJING UNIVERSITY
NANJING 210093
CHINA
huicheng@nju.edu.cn

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

BI-HAMILTONIAN FLOWS AND THEIR REALIZATIONS AS CURVES IN REAL SEMISIMPLE HOMOGENEOUS MANIFOLDS

GLORIA MARÍ BEFFA

We describe a reduction process that allows us to define Hamiltonian structures on the manifold of differential invariants of parametrized curves for any homogeneous manifold of the form G/H, with G semisimple. We also prove that equations that are Hamiltonian with respect to the first of these reduced brackets automatically have a geometric realization as an invariant flow of curves in G/H. This result applies to some well-known completely integrable systems. We study in detail the Hamiltonian structures associated to the sphere $\mathrm{SO}(n+1)/\mathrm{SO}(n)$.

1. Introduction

Completely integrable systems are PDEs for which one can find an infinite family of preserved functionals in involution. Most of these systems are bi-Hamiltonian, that is, they are Hamiltonian with respect to two different but compatible Hamiltonian structures (compatible means that their sum is also a Hamiltonian structure). The Hamiltonian structures are used to generate a recursion operator, an operator that when reiteratively applied to one initial preserved functional generates the entire family — or hierarchy — see [Magri 1978]. In recent years a large number of publications have shown that many completely integrable systems appear linked to the geometric background of curves and surfaces; see for example [Anco 2006; Doliwa and Santini 1994; Ferapontov 1995; Gay-Balmaz and Ratiu 2007; Chou and Qu 2002; 2003; Langer and Perline 1991; 2000; Sanders and Wang 2003; Terng and Thorbergsson 2001; Terng and Uhlenbeck 2006; 2000; Yasui and Sasaki 1998; Marí 2008a; 2006; 2008b; 2007; 2005; 2009; Marí et al. 2002] and references within. Some of this work relates the integrable systems to invariant flows of (in general parametrized) curves in different types of manifolds through geometric realizations, that is, evolutions of curves inducing the integrable system on its curvatures, or differential invariants in general. Perhaps the best known example of

MSC2000: primary 37K10, 37K25, 37K30; secondary 53A55.

Keywords: invariant curve evolutions, Poisson structures, differential invariants, moving frames, homogeneous spaces.

such a geometric realization is that of the nonlinear Schrödinger equation (NLS) by the vortex filament flow (VF). Hasimoto [1972] showed that VF, viewed as a flow of spacial Euclidean curves, induces the NLS on its curvature and torsion via what it became known as the Hasimoto transformation. The Hasimoto transformation was proved to be a Poisson map between two equivalent bi-Poisson manifolds, that of the standard curvature and torsion and the manifold of natural curvatures; see [Langer and Perline 1991; Marí et al. 2002].

The author of this paper has linked the bi-Hamiltonian structures of many of these integrable systems to a process that allows us to reduce well-known compatible Poisson brackets on the manifold of loops in the dual of a Lie algebra, which we will call $\mathscr{L}\mathfrak{g}^*$, to the manifold of differential invariants. This reduction process was described in [Marí 2008a] for homogeneous manifolds of the form G/H where $\mathfrak{g}$, the Lie algebra associated to G, is $|1|$-graded. These include $\mathbb{RP}^{n+1}$, the conformal Möbius sphere, the Grassmannian, the Lagrangian Grassmannian and others. The reduction process was also described in [Marí 2006] for the case of affine geometries, that is, homogeneous manifolds of the form $G \ltimes \mathbb{R}^n/G$ with G semisimple. In both cases a well-known Poisson structure (we will refer to it as our *first bracket*) on $\mathscr{L}\mathfrak{g}^*$ can be reduced to the space of differential invariants to produce some of the best known Hamiltonian structures used in the integration of PDEs. This structure is also linked to geometric realizations in the sense that under minimal conditions one can find geometric realizations for any Hamiltonian evolution, and hence for bi-Hamiltonian integrable systems. The reduction of a *second* compatible bracket is not guaranteed, and neither is the existence of an associated integrable system. Indeed it was shown in [Marí 2005] that in the Lagrangian 2-Grassmannian manifold (or Grassmannian of Lagrangian planes in $\mathbb{R}^4$), the second bracket in $\mathscr{L}\mathfrak{sp}(2)^*$ never reduces. No completely integrable systems induced by Lagrangian flows on the differential invariants have been found. On the other hand, the reduction of the second bracket, whenever possible, points at the existence of an associated completely integrable system, or at least it is so in all known examples. Coming from a different direction, Terng, with Thorbergsson in [2001] and with Uhlenbeck in [2000; 2006], started by constructing classical completely integrable systems that are Hamiltonian with respect to the reduction of the second bracket, the bracket defined on coadjoint orbits, and after finding the existence of these systems they link them to our first bracket. These two different approaches have not been clearly bridged yet.

Even in the cases where the second bracket does not reduce, one can at times find integrable systems as level sets of Hamiltonian evolutions: the second bracket might not reduce to the complete manifold of differential invariants, but it might reduce to a submanifold of it defined by some chosen invariants. The geometric realization might exist if initial conditions are restricted to the types of curves for which the undesired invariants are constant. For example, in the case of the Lagrangian n-Grassmannian the second Poisson bracket does not reduce in general, but it does always reduce to the submanifold defined by the eigenvalues of the so-called *Lagrangian Schwarzian* derivatives, whenever the other invariants vanish. In fact, it has been conjectured (and studies are supportive of this) that the type of Poisson structures/integrable systems and the character of the chosen invariants are closely related. For example, one can usually reduce the second bracket to a submanifold of differential invariants of projective type (as was done in [Marí 2008a; 2008b; 2007]) to obtain Poisson structures and integrable systems of KdV type (for example, the KdV equation or systems of decoupled KdV in [2008a; 2008b; 2007], complexly coupled KdV equations in [2008b], and Adler–Gel'fand–Dikii evolutions in [2008a]). Similarly, one can reduce to a submanifold of curvatures of *natural-type* to obtain modified KdV vector equations and NLS systems, as in [Anco 2006; Marí et al. 2002; Sanders and Wang 2003; Terng and Thorbergsson 2001; Terng and Uhlenbeck 2006].

A last relevant feature of these brackets is the following. Some of the Poisson structures obtained when reducing our first bracket are not truly structures associated to parametrized curves, but trivial extensions of Poisson brackets associated to unparametrized curves and extended trivially to the differential invariant of arc length type, as defined in [Marí 2009]. Except for the case $G = \mathrm{GL}(n, \mathbb{R})$, all classical affine geometries $G \ltimes \mathbb{R}^n/G$ have first reductions for which Hamiltonian evolutions will always preserve the invariant of arc length type [Marí 2009]. On the other hand, all known examples for semisimple parabolic cases (G/P, with P parabolic) have reductions of the first bracket that do not preserve parameters of arc length type. Indeed, geometric realizations of equations of KdV type do not preserve any invariant of arc length type. Thus, having first reductions on parametrized or unparametrized curves seems to be linked to the type of geometry that the manifold has.

In this paper we describe the reduction process for the general case of a homogeneous manifold G/H with G semisimple. Semisimplicity can be trivially assumed for the definition of the bracket; otherwise the bracket will only be defined on the semisimple component of the algebra. The reduction process here is, in fact, a simplification of the process in [Marí 2008a]. We prove in Theorem 4.3 that any system that is Hamiltonian with respect to the first reduced bracket possesses a geometric realization by an invariant flow on G/H. Our running example is that of $\mathrm{SO}(2, 2)/P$ for an appropriate choice of parabolic subgroup P. This manifold is geometrically equivalent to $\mathbb{RP}^1 \times \mathbb{RP}^1$ and we show that both brackets reduce to produce a decoupled system of KdV bi-Hamiltonian structures. The manifold $\mathrm{SO}(3, 1)/P$ (the conformal plane) is known [Marí 2005] to produce a system of two complexly coupled KdV equations. Thus, we show that the exchange

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

in the bilinear form defining the group effectively decouples the KdV system.

Finally, the Hamiltonian structures of the sphere $\mathrm{SO}(n+1)/\mathrm{SO}(n)$ is studied in Section 5. In [Terng and Thorbergsson 2001; Anco 2006], the authors found a geometric realization on this sphere for a vector system of modified KdV equations. The authors do not study the generation of the mKdV bi-Hamiltonian structures, or their possible definition by reduction (Anco provides a recursion operator that is said to be encoded by the geometry, but he provides no explanation on how the encoding takes place). The case of $\mathrm{SO}(n+1)/\mathrm{SO}(n)$ is interesting because, being a semisimple case (albeit not a parabolic one), one would think that the arc length does not need to be preserved; said differently, the first reduced Poisson structure should be expected to be a structure on parametrized curves. On the other hand, the mKdV systems found by Anco and Terng and Thorbergsson associated to this geometric background (and found also in the Euclidean case, an affine manifold) are arc length preserving. In Section 5, we show that the first reduced bracket does not preserve arc length, so that the bracket is defined on parametrized curves, in accordance with the manifold being homogeneous and semisimple. But here it is the *second bracket* that always preserves arc length and, hence, forces any bi-Hamiltonian system to be arc length preserving, in accordance with vector mKdV being the associated integrable system. The system of vector mKdVs is shown to be a bi-Hamiltonian system with respect to both reductions.

The reduction method we use is strongly rooted in the use of group-based moving frames. The method is relatively new so we include a description in our first section, together with other background definitions.

2. Background definitions

2a. ***Moving frames, differential invariants, Serret–Frenet equations and geometric realizations.*** The classical concept of moving frame was developed by Élie Cartan [1935; 1937]. A classical moving frame along a curve in a manifold M is a curve in the frame bundle of the manifold over the curve, invariant under the action of a transformation group under consideration. This method is a very powerful tool, but its explicit application relied on intuitive choices that were not clear in a general setting. Ideas in Cartan's work and later work of Griffiths [1974], Green [1978] and others laid the foundation for the concept of a group-based moving frame, that is, an equivariant map between the jet space of curves in the manifold and the group of transformations. Recent work by Fels and Olver [1999] finally gave the precise definition of the group-based moving frame and extended its application beyond its original geometric picture. In this section we will describe Fels and Olver's moving frame and its role in our study. From now on we will assume $M = G/H$ with G semisimple and acting on M via left multiplication on representatives of a class. We will also assume that curves in M are parametrized, and that therefore the group G does not act on the parameter.

Definition 2.1. Let $J^k(\mathbb{R}, M)$ be the space of k-jets of curves, that is, the set of equivalence classes of curves in M up to k-th order of contact. If we denote by $u(x)$ a curve in M and by u_r the r-th derivative of u with respect to the parameter x, that is, $u_r = d^r u/dx^r$, then the jet space has local coordinates that can be represented by $u^{(k)} = (x, u, u_1, u_2, \dots, u_k)$. The group G acts naturally on parametrized curves; therefore it acts naturally on the jet space via the formula

$$g \cdot u^{(k)} = (x, g \cdot u, (g \cdot u)_1, (g \cdot u)_2, \dots),$$

where by $(g \cdot u)_k$ we mean the formula obtained when one differentiates $g \cdot u$ and then writes the result in terms of g, u, u_1, and so on. This is usually called the *prolonged* action of G on $J^k(\mathbb{R}, M)$.

Definition 2.2. A function $I : J^k(\mathbb{R}, M) \to \mathbb{R}$ is called a k-th order *differential invariant* if it is invariant with respect to the prolonged action of G.

Definition 2.3. A map $\rho : J^k(\mathbb{R}, M) \to G$ is called a left (respectively right) *moving frame* if it is equivariant with respect to the prolonged action of G on $J^k(\mathbb{R}, M)$ and the left (respectively right) action of G on itself.

The group-based moving frame appears in a familiar method for calculating the curvature of a curve $u(s)$ in the Euclidean plane. In this method one uses a translation to take $u(s)$ to the origin and then a rotation to make one of the axes tangent to the curve. The curvature can classically be found as the coefficient of the second order term in the expansion of the curve around $u(s)$. The crucial observation made by Fels and Olver is that the element of the group carrying out the translation and rotation depends on u and its derivatives and so defines a map from the jet space to the group. This map is a right moving frame, and it carries all the geometric information of the curve. In fact, Fels and Olver developed a similar normalization process to find right moving frames.

Theorem 2.4 [Fels and Olver 1999]. *Let $\cdot$ denote the prolonged action of the group on $u^{(k)}$ and assume we have* normalization equations *of the form $g \cdot u^{(k)} = c_k$, where at least some of the entries of c_k are constants* (*they are called normalization constants*). *Assume we have enough normalization equations to determine g as a function of $u, u_1, \dots$. Then $g = \rho$ is a right invariant moving frame.*

Next is the description of the equivalent to the classical Serret–Frenet equations. We are denoting by L^*_g (respectively R^*_g), the map induced on TG by L_g, the left multiplication by g (respectively R_g, the right multiplication). From now on, we will also assume the (local) connection on M is flat, although some modifications can be introduced to assume constant curvature.

Definition 2.5. Consider $K\,dx$ to be the horizontal component of the pullback of the left- (respectively right)-invariant Maurer–Cartan form of the group G via a left (respectively right) moving frame ρ. That is,

$$K = L^*_{\rho^{-1}}\rho_x \in \mathfrak{g} \qquad \text{(respectively} \quad K = R^*_{\rho^{-1}}\rho_x\text{)}$$

We call K the *left* (respectively *right*) *Serret–Frenet equations* for the moving frame ρ.

Notice that, if ρ is a left moving frame, then ρ^{-1} is a right moving frame and their Serret–Frenet equations are the negatives of each other. A complete set of generating differential invariants can always be found among the coordinates of group-based Serret–Frenet equations, a crucial difference with the classical picture. The next theorem is a direct consequence of the results in [Fels and Olver 1999]. A more general result can be found in [Hubert 2007].

Theorem 2.6. *Let ρ be a* (*left or right*) *moving frame along a curve u. Let us fix a basis for $\mathfrak{g}$. Then*, *the coordinates of the* (*left or right*) *Serret–Frenet equations for ρ contain a basis for the space of differential invariants of the curve. That is*, *any other differential invariant for the curve is a function of the coordinates of K and their derivatives with respect to x.*

If we can find a moving frame using a set of normalization equations as in Theorem 2.4, we can also find algebraically the explicit form of the Serret–Frenet equations of the frame, following a parallel set of *recurrence equations*. Let $K \cdot u$ represent the infinitesimal action of the algebra $\mathfrak{g}$, likewise with $K \cdot u^{(k)}$, which represents the infinitesimal prolonged action. The following theorem is a rewriting of results appearing in [Fels and Olver 1999].

Theorem 2.7. *Let $K = L^*_{\rho^{-1}}\rho_x$ be the left Serret–Frenet equation associated to the left moving frame ρ. Let ρ be determined by normalization equations of the form $\rho^{-1} \cdot u_k = c_k$. Then*, *$K$ satisfies the equations*

$$K \cdot u_k|_{\mathcal{I}} = c_{k+1} - (c_k)_x,$$

where $K \cdot u_k|_{\mathcal{I}}$ denotes what is usually called the invariantization of $K \cdot u_k$, *that is*, *the expression $K \cdot u_k$ with all u_r substituted by c_r.*

Definition 2.8 (geometric realization of an evolution of invariants). Let $\boldsymbol{k}$ denote a vector whose entries form an independent and generating system of differential invariants for curves. That is, $\boldsymbol{k}$ is a vector whose entries are differential invariants for the curve; the entries of $\boldsymbol{k}$ and their derivatives are functionally independent (no entry can be written as a function of the other entries and their derivatives); and any other differential invariant is a function of the entries of $\boldsymbol{k}$ and their derivatives.

Let

$$\boldsymbol{k}_t = F(\boldsymbol{k}, \boldsymbol{k}_x, \boldsymbol{k}_{xx}, \dots) \tag{1}$$

be an evolution of $\boldsymbol{k}$. We say that

$$u_t = Q(u, u_x, u_{xx}, u_{xxx}, \dots) \tag{2}$$

is *a geometric realization of* (1) *on* G/H whenever $u(t, x) \in G/H$, (2) is invariant under the action on G (that is, G takes solutions to solutions) and the evolution induced on $\boldsymbol{k}$ by (2) is (1). Equivalently, we say that (1) is the *invariantization* of (2).

2b. ***Poisson brackets on*** $\mathscr{L}\mathfrak{g}^*$. Consider the group of loops $\mathscr{L}G = C^\infty(S^1, G)$ and its Lie algebra $\mathscr{L}\mathfrak{g} = C^\infty(S^1, \mathfrak{g})$. Let $\hat{B} : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$ be an ad-invariant nondegenerate bilinear form of the algebra. We can use $\hat{B}$ to identify $\mathfrak{g}^*$ with $\mathfrak{g}$ so that $X^* = \hat{B}(X, \cdot) \in \mathfrak{g}^*$. For example, if $\mathfrak{g} \subset \mathfrak{gl}(n, \mathbb{R})$, then $\hat{B}$ can be the trace of the matrix product. With this bilinear form, the dual to E_{ij} (the matrix having 1 in place (i, j) and 0 elsewhere) is given by E_{ji}. The bilinear form

$$B(X, Y) = \int_{S^1} \hat{B}(X, Y) dx \tag{3}$$

will give us the analogous form defined on $\mathscr{L}\mathfrak{g}$, and we can identify $\mathscr{L}\mathfrak{g}^*$ (the regular part of $(\mathscr{L}\mathfrak{g})^*$) with $\mathscr{L}\mathfrak{g}$ using B.

One can define two natural Poisson brackets on $\mathscr{L}\mathfrak{g}^*$; for more information see [Pressley and Segal 1989]. If $\mathscr{H}, \mathscr{G} : \mathscr{L}\mathfrak{g}^* \to \mathbb{R}$ are two functionals defined on $\mathscr{L}\mathfrak{g}^*$, then $\delta\mathscr{H}/\delta L$ denotes the variational derivative of $\mathscr{H}$ at L and it can be identified, using (3), with an element of $\mathscr{L}\mathfrak{g}$ so that

$$\frac{d}{d\epsilon}\Big|_{\epsilon=0} \mathscr{H}(L + \epsilon V) = \int_{S^1} \hat{B}\Big(\frac{\delta\mathscr{H}}{\delta L}, V\Big) dx. \tag{4}$$

Likewise with $\mathscr{G}$. If $L \in \mathscr{L}\mathfrak{g}^*$, we define

$$\{\mathscr{H}, \mathscr{G}\}(L) = \int_{S^1} \Big\langle \Big(\frac{\delta\mathscr{H}}{\delta L}\Big)_x + \mathrm{ad}^*\Big(\frac{\delta\mathscr{H}}{\delta L}\Big)(L), \frac{\delta\mathscr{G}}{\delta L} \Big\rangle dx \tag{5}$$

where $\langle \cdot, \cdot \rangle$ is the natural coupling between $\mathfrak{g}^*$ and $\mathfrak{g}$ and where we identify $(\delta\mathscr{H}/\delta L)_x$ with its dual counterpart. If we identify L with its dual, then we have $\mathrm{ad}^*(\delta\mathscr{H}/\delta L)(L) = -\,\mathrm{ad}(\delta\mathscr{H}/\delta L)(L)$.

One also has a compatible family of second brackets, namely

$$\{\mathcal{H},\mathcal{G}\}(L)=\int_{S^1}\Big\langle\Big(\mathrm{ad}^*\Big(\frac{\delta\mathcal{H}}{\delta L}\Big)(L_0),\frac{\delta\mathcal{G}}{\delta L}\Big\rangle dx, \tag{6}$$

where $L_0\in\mathfrak{g}^*$ is any constant element.

In the next section we will show how (5) can be always reduced to the space of differential invariants of curves. The compatible bracket (6) can only be reduced sometimes. Recall that the appearance of compatible pairs of Poisson brackets often indicates the existence of completely integrable systems.

3. Geometric Poisson brackets on the space of differential invariants of curves

Since $H\subset G$ is a subgroup, the algebra $\mathfrak{g}$ has a splitting of the form

$$\mathfrak{g}=\mathfrak{h}\oplus\mathfrak{m}, \tag{7}$$

where $\mathfrak{m}$ is a vector subspace complement to the subalgebra $\mathfrak{h}$, but not a subalgebra in general. From now on we will also assume that our curves on homogeneous manifolds have a *group monodromy*, that is, there exists $m\in G$ such that

$$u(t+T)=m\cdot u(t),$$

where T is the period. Under these assumptions, the Serret–Frenet equations will be periodic and will belong to $\mathcal{L}\mathfrak{g}^*$ (under proper identification). Alternatively, one could assume that u is asymptotic at ±∞, so that the invariants vanish at infinity, and describe a similar situation.

Theorem 3.1. *Let u be a generic curve on the homogeneous manifold G/H. Let ρ be a left moving frame with $\rho\cdot o=u$. Locally, we can find moving frames for curves $\hat{u}$ in a neighborhood of u (with respect to the C^∞ topology) such that $\rho\cdot o=\hat{u}$. Let $\mathcal{K}$ be the submanifold of $\mathcal{L}\mathfrak{g}$ given by the Serret–Frenet equations associated to these left moving frames, in the sense of the previous section. Then, when identified with its dual, $\mathcal{K}$ defines a section of the quotient $\mathcal{L}\mathfrak{g}^*/\mathcal{L}H$, where the subgroup $\mathcal{L}H$ acts on $\mathcal{L}\mathfrak{g}^*$ via the standard gauge action*

$$a(g)(L)=L^*_{g^{-1}}g_x+\mathrm{Ad}^*(g)(L)$$

and where again the element $L^*_{g^{-1}}g_x$ is identified with its dual.

Proof. This theorem is proved using the definition of moving frame. Assume $m\in\mathcal{L}\mathfrak{g}^*$ and identify the element with an element in the algebra. Let η be a (local) solution of the equation $L^*_{\eta^{-1}}\eta_x=m$. We call $u=\eta\cdot o$ and we denote by ρ a left moving frame associated to u, with $\rho\cdot o=u$. The frame ρ has the same monodromy as u, and u has the same monodromy as η. Hence, ρ and η have the same monodromy.

With these choices we have $\rho = \eta\eta^{-1}\rho = \eta g$ and $g \cdot o = \eta^{-1}\rho \cdot o = \eta^{-1} \cdot u = o$. Since H is the isotropy group of o (which represents the class of H in G/H), we conclude that $g(x) \in H$ for any x. Furthermore, the monodromy of both ρ and η are the same, and therefore $g \in \mathscr{L}H$. The action of $\mathscr{L}G$ on the space of solutions $\eta \to \eta g$ induces the gauge action described in the theorem on the elements of $\mathscr{L}\mathfrak{g}^*$ defining the equations satisfied by η. If identified with $\mathscr{L}\mathfrak{g}$, $\mathrm{Ad}^*(g)(L) = \mathrm{Ad}(g^{-1})(L)$ and the action on $\mathscr{L}\mathfrak{g}$ induced by the gauge action is $L^*_{g^{-1}} g_x + \mathrm{Ad}(g^{-1})L$. □

Example 3.2. Our running example will be the case $G = \mathrm{SO}(2, 2)$ for $H = P$ given by a particular parabolic choice. Assume $\mathrm{SO}(2, 2)$ is the isotropy group of the bilinear form defined by the matrix

$$J = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

that is, $\mathfrak{o}(2, 2)$ is the set of matrices that are skew-symmetric with respect to the secondary diagonal. Locally, $g \in \mathrm{SO}(2, 2)$ can be factored as

$$\begin{aligned} g &= g_1(v)g_0(\alpha, \Theta)g_{-1}(y) \\ &= \begin{pmatrix} 1 & v^1 & v^2 & -v^2v^1 \\ 0 & 1 & 0 & -v^2 \\ 0 & 0 & 1 & -v^1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \Theta & 0 \\ 0 & 0 & \alpha^{-1} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ y^1 & 1 & 0 & 0 \\ y^2 & 0 & 1 & 0 \\ -y^2y^1 & -y^2 & -y^1 & 1 \end{pmatrix}, \end{aligned}$$

with $\alpha \in \mathbb{R}$ and $\Theta \in \mathrm{SO}(1, 1)$. This factorization corresponds to the algebra gradation $\mathfrak{o}(2, 2) = \mathfrak{g}_1 \oplus \mathfrak{g}_0 \oplus \mathfrak{g}_{-1}$ as in the diagram

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

Let us choose the parabolic subgroup $H = P = G_1 \cdot G_0$, that is, the subgroup defined by elements g such that $y^1 = y^2 = 0$. Notice that $\mathrm{SO}(3, 1)$ has the exact same description, with one difference, namely $\Theta \in \mathrm{SO}(2)$ (here $-v^1v^2 = -\frac{1}{2}\|\boldsymbol{v}\|_J$ — see below — while for $\mathrm{SO}(3, 1)$ we would have $-\frac{1}{2}\|\boldsymbol{v}\| = -\frac{1}{2}\boldsymbol{v}^T\boldsymbol{v}$ instead).

With this representation, the action of $\mathrm{SO}(2, 2)$ on $\mathrm{SO}(2, 2)/H$ is determined by the relation $gg_{-1}(u) = g(g \cdot u)h$ for some $h \in H$. We will use the section $\varsigma : \mathrm{SO}(2, 2)/H \to \mathrm{SO}(2, 2)$ given by $\varsigma(u) = g_{-1}(u)$ to locally identify $\mathrm{SO}(2, 2)/H$

with G_{-1}. The subgroups G_i are the exponentials of the Lie subalgebras $\mathfrak{g}_i$. One can readily find an explicit formula for the action using this notation:

$$g \cdot u = \frac{\alpha^{-1}\Theta(\boldsymbol{u}+\boldsymbol{y}) + \frac{1}{2}\alpha^{-2}\|\boldsymbol{u}+\boldsymbol{y}\|_J^2 \boldsymbol{v}^*}{1+\alpha^{-1}\boldsymbol{v}^T\Theta(\boldsymbol{u}+\boldsymbol{y}) + \frac{1}{4}\alpha^{-2}\|\boldsymbol{v}\|_J^2\|\boldsymbol{u}+\boldsymbol{y}\|_J^2} \tag{8}$$

where $\|x\|_J = \hat{x}^T J \hat{x}$ for $\hat{x} = (0, x, 0)$ and where, if $\boldsymbol{v} = \binom{v^1}{v^2}$, then $\boldsymbol{v}^* = \binom{v^2}{v^1}$. One can check that this action decouples into two projective actions. If $\Theta \in \mathrm{SO}(1,1)$ with $\Theta = \mathrm{diag}(a, a^{-1})$, the two projective actions are given by $(y^1, a\alpha^{-1}, v^1)$ acting projectively on u^1 and $(y^2, a\alpha^{-1}, v^2)$ acting projectively on u^2. This is because the isomorphism $\mathfrak{o}(2,2) \cong \mathfrak{sl}(2,\mathbb{R}) \oplus \mathfrak{sl}(2,\mathbb{R})$ induces a splitting of $\mathrm{SO}(2,2)$ into $\mathrm{SL}(2,\mathbb{R}) \times \mathrm{SL}(2,\mathbb{R})$ and also an equivalence of $\mathrm{O}(2,2)/P$ with $\mathbb{RP}^1 \times \mathbb{RP}^1$. (A choice of Θ on the second connected component of $O(1,1)$ will simply produce an involution exchanging u^1 and u^2.)

If g is as in (8), the zero normalization equation is $g \cdot \boldsymbol{u} = 0$, which can be solved with the choice $\boldsymbol{y} = -\boldsymbol{u}$. If $\boldsymbol{u} = \boldsymbol{u}(x)$, the first normalization equation is $g \cdot \boldsymbol{u}_1 = c_1$, obtained by differentiating the action (8) with respect to x and substituting $\boldsymbol{y} = -\boldsymbol{u}$. It is given by

$$\alpha^{-1}\Theta\boldsymbol{u}_1 = c_1.$$

Since $\Theta = \mathrm{diag}(a, a^{-1}) \in O(1,1)$, we need to choose nonvanishing normalization values for each of the entries of c_3. We choose $c_1 = \binom{1}{1}$, rather than the usual $c_1 = e_1$ favored in normalizations — in this case e_1 would be a singular choice. This choice forces the values

$$\alpha = \|\boldsymbol{u}_x\|_J 2^{-1/2} \quad \text{and} \quad \Theta^{-1}\begin{pmatrix}1\\1\end{pmatrix} = \frac{\sqrt{2}\boldsymbol{u}_x}{\|\boldsymbol{u}_x\|_J}.$$

This condition completely determines $\Theta = \mathrm{diag}(\alpha(u_x^1)^{-1}, \alpha^{-1}u_x^1)$.

The second normalization equation is obtained differentiating (8) twice and substituting previously found values. It is given by

$$\alpha^{-2}\Theta\boldsymbol{u}_{xx} - \boldsymbol{v} = c_2 = 0,$$

which is readily resolved choosing $\boldsymbol{v} = \alpha^{-2}\Theta\boldsymbol{u}_{xx}$. This last equation completely determines the right moving frame. Following [Fels and Olver 1999], we have a set of independent and generating invariants given by the entries of c_3; we have two invariants of third order. The interested reader can differentiate once more and find the third normalization equations and the explicit formula for c_3. It is given by $c_3 = \binom{k_1}{k_2}$ with $k_i = S(u^i)$, where $S(f) = f_x^{-1}(f_{xxx} - \frac{3}{2}(f_{xx}/f_x)^2)$ is the Schwarzian derivative of f. The Schwarzian derivative is the generator of projective differential invariants in $\mathbb{RP}^1$.

Let's call ρ the left moving frame, that is, the inverse of the frame we just found:

$$\rho = \begin{pmatrix} 1 & -(\boldsymbol{u}^*)^T & -\frac{1}{2}\|\boldsymbol{u}\|_J^2 \\ 0 & I & \boldsymbol{u} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha^{-1} & 0 & 0 \\ 0 & \Theta^{-1} & 0 \\ 0 & 0 & \alpha \end{pmatrix} \begin{pmatrix} 1 & -\boldsymbol{v}^T & -\frac{1}{2}\|\boldsymbol{v}\|_J \\ 0 & I & \boldsymbol{v}^* \\ 0 & 0 & 1 \end{pmatrix}.$$

In parallel with the normalization equations, we can use recurrence formulas of Theorem 2.7 to determine the matrix $K = \rho^{-1}\rho_x$. If $K = K_1 + K_0 + K_{-1}$ are the gradated components of K and $K_0 = K_\alpha + K_\Theta$ are the two components of K_0, the recurrence formulas are given by

$$\begin{aligned} K \cdot u|_{\mathcal{I}} &= K_{-1} = c_1 - c_0' = c_1 = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right), \\ K \cdot u_1|_{\mathcal{I}} &= K_\Theta c_1 - K_\alpha c_1 = c_2 - c_1' = 0. \end{aligned}$$

The last equation implies $K_\Theta = 0$ and $K_\alpha = 0$. The two equations describe K as being of the form

$$K = \begin{pmatrix} 0 & k_1 & k_2 & 0 \\ 1 & 0 & 0 & -k_2 \\ 1 & 0 & 0 & -k_1 \\ 0 & -1 & -1 & 0 \end{pmatrix}. \tag{9}$$

The general theory tells us that the entries of K generate all other differential invariants for u, and hence k_1 and k_2 must be generators. If one writes the recurrence equations of Theorem 2.7 for the second prolongation, we also see that K_1 coincides with c_3. This matrix is very similar to the one obtained in the case $G = \mathrm{O}(3, 1)$ for which G/H is the conformal plane; see [Marí 2008b]. The only difference is that in the conformal case, $c_1 = e_1$ is a regular value and $K_{-1} = e_1$ was chosen instead. This small difference will create a very significant one for the reduced Poisson brackets and their associated integrable systems.

Our next theorem shows that (5) can be reduced to $\mathcal{K}$, and its proof gives an algebraic method to calculate the reduced bracket explicitly (and also the reduction of (6) whenever possible).

Theorem 3.3. *The Poisson bracket defined on* $\mathcal{L}\mathfrak{g}^*$ *by* (5) *is reducible to the submanifold* $\mathcal{K}$. *We call this the first reduced Poisson bracket associated to curves on* G/H.

Proof. Observe that $\mathcal{K}$ is given locally by the quotient $\mathcal{L}\mathfrak{g}^*/\mathcal{L}H$, where $\mathcal{L}H$ acts in $\mathcal{L}\mathfrak{g}^*$ via the gauge action. The symplectic leaves of the bracket (5) are formed by the orbits of the gauge action itself. For more information on these brackets, see [Pressley and Segal 1989]. Assume we have two functionals $\mathcal{R}$ and $\mathcal{G}$ such that

$\delta\mathcal{R}/\delta L$, $\delta\mathcal{G}/\delta L \in \mathscr{L}\mathfrak{g}$ vanish on the tangent to the $\mathscr{L}H$-leaves. That means

$$\left(\frac{\delta\mathcal{R}}{\delta L}\right)_x - \operatorname{ad}\left(\frac{\delta\mathcal{R}}{\delta L}\right)(K) \in \mathfrak{h}^0 \tag{10}$$

and likewise for $\mathcal{G}$ (we are identifying $\mathscr{L}\mathfrak{g}$ with $\mathscr{L}\mathfrak{g}^*$). Then, the bracket (5) of these two functionals will also vanish on the tangent to the leaves (equivalently, it will be constant on the leaves); one only needs to apply Jacobi's identity for (5) to see this. Hence, the bracket will represent a well-defined functional on the quotient $\mathcal{K}$.

Following the same reasoning as in [Marsden and Ratiu 1986], let $r, g : \mathcal{K} \to \mathbb{R}$ be two functionals, and let $\mathcal{R}$ and $\mathcal{G}$ be two extensions that are constant on the leaves of $\mathscr{L}H$. The bracket

$$\{r, g\}(K) = \int_{S^1} \left\langle \left(\frac{\delta\mathcal{R}}{\delta L}\right)_x - \operatorname{ad}\left(\frac{\delta\mathcal{R}}{\delta L}\right)(K), \frac{\delta\mathcal{G}}{\delta L} \right\rangle dx \tag{11}$$

describes a well-defined functional on $\mathcal{K}$. It is a Poisson bracket on $\mathcal{K}$ in which Jacobi's identity is given directly by the Jacobi identity of (5). For a complete description of this and other Poisson reductions for finite-dimensional manifolds, see [Marsden and Ratiu 1986]. Our infinite-dimensional case is a straightforward generalization of the results there. □

Although this bracket seems to be complicated to compute, in all known cases the calculation follows a purely algebraic process that can be done by hand in low dimensions. The essence of the algebraic process is the use of (10).

Example 3.4. We now go back to the case $G = \mathrm{SO}(2, 2)$. In this case $\mathfrak{h} = \mathfrak{g}_0 \oplus \mathfrak{g}_1$ and so $\mathfrak{h}^0 = \mathfrak{g}_1$. If K is given as in (9), then an extension $\mathcal{R}$ of a functional $r : \mathcal{K} \to \mathbb{R}$ to $\mathscr{L}\mathfrak{o}(2, 2)^*$ will coincide with r in the direction of k_1 and k_2. The variational derivative of $\mathcal{R}$ is defined as in (4), and so

$$\frac{\delta\mathcal{R}}{\delta L}(K) = \begin{pmatrix} \beta & a & b & 0 \\ \delta r/\delta k_1 & c & 0 & -b \\ \delta r/\delta k_2 & 0 & -c & -a \\ 0 & -\delta r/\delta k_2 & -\delta r/\delta k_1 & -\beta \end{pmatrix}. \tag{12}$$

If we substitute these values in condition (10), we get along $\mathcal{K}$ that

$$\begin{pmatrix} \beta' + k_1\delta r/\delta k_1 + k_2\delta r/\delta k_2 - a - b & a' + ck_1 - \beta k_1 & b' - ck_2 - \beta k_2 & 0 \\ (\delta r/\delta k_1)_x + \beta - c & c' + a + k_2\delta r/\delta k_2 - k_1\delta r/\delta k_1 - b & 0 & * \\ (\delta r/\delta k_2)_x + \beta + c & 0 & * & * \\ 0 & * & * & * \end{pmatrix}$$

$$= \begin{pmatrix} 0 & * & * & 0 \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

From here we obtain

$$\beta=-\frac{1}{2}\Big(\frac{\delta r}{\delta k_1}+\frac{\delta r}{\delta k_2}\Big)', \qquad c=\frac{1}{2}\Big(\frac{\delta r}{\delta k_1}-\frac{\delta r}{\delta k_2}\Big)'$$
$$a=-\frac{1}{2}\Big(\frac{\delta r}{\delta k_1}\Big)''+k_1\frac{\delta r}{\delta k_1}, \quad b=-\frac{1}{2}\Big(\frac{\delta r}{\delta k_2}\Big)''+k_2\frac{\delta r}{\delta k_2}$$

The reduced Poisson bracket is defined by (11), where $\mathcal{R}$ and $\mathcal{G}$ are appropriate extensions with variational derivatives as above. After straightforward calculations, these can be written as

$$\{r,g\}(\boldsymbol{k})=\int_{S^1}\frac{\delta g}{\delta k_1}\big(-\tfrac{1}{2}D^3+Dk_1+k_1D\big)\frac{\delta r}{\delta k_1}+\frac{\delta g}{\delta k_2}\big(-\tfrac{1}{2}D^3+Dk_1+k_1D\big)\frac{\delta r}{\delta k_2};$$

therefore the first reduced bracket is defined by two *decoupled second Poisson structures* for KdV equations, one in each k_1 and k_2. We can also check whether or not, for some choice of L_0, the bracket (6) reduces to $\mathcal{H}$ by evaluating (6) in our extensions. If we choose $L_0=E_{12}-E_{21}+E_{13}-E_{31}$ (that is, the element dual to K_{-1}), the result is

$$\{r,g\}_0(\boldsymbol{k})=\int_{S^1}\Big\langle\frac{\delta\mathcal{G}}{\delta L}(K),\Big[L_0,\frac{\delta\mathcal{R}}{\delta L}(K)\Big]\Big\rangle dx=-2\int_{S^1}\frac{\delta g}{\delta k_1}D\frac{\delta r}{\delta k_1}+\frac{\delta g}{\delta k_2}D\frac{\delta r}{\delta k_2}.$$

Thus, the second reduced bracket is given by two *decoupled first Poisson structures* for KdV equations.

This result fits well with the equivalence $\mathrm{SO}(2,2)/H\cong\mathbb{RP}^1\times\mathbb{RP}^1$. Indeed, the two reduced Poisson brackets associated to the geometry of flows in $\mathbb{RP}^1$ are known to be the two KdV Hamiltonian structures. On the other hand, $O(3,1)/P$ is the conformal plane and the two reduced Poisson brackets were given by the two Hamiltonian structures for a complexly coupled system of KdV equations [Marí 2005]. Thus the change $O(3,1)\to O(2,2)$ decouples the Hamiltonian structures.

4. Geometric realizations of Hamiltonian evolutions

Let $\Phi_g:G/H\to G/H$ be defined by the action of $g\in G$ on the quotient, that is, $\Phi_g(x)=\Phi_g([y])=[gy]=g\cdot x$. Let $\varsigma:G/H\to G$ be a section of the homogeneous quotient such that $\varsigma(o)=e$. The section is compatible with the action of G on G/H, that is,

$$g\varsigma(x)=\varsigma(\Phi_g(x))h \quad \text{for some } h\in H. \tag{13}$$

This relation in fact determines the action of the group on G/H uniquely, as we saw in our running example. As before, we consider the splitting of the Lie algebra $\mathfrak{g}=\mathfrak{h}\oplus\mathfrak{m}$, where $\mathfrak{m}$ is not in general a Lie subalgebra. Since ς is s section, $d\varsigma(o)$ is an isomorphism between $\mathfrak{m}$ and T_oM.

The following theorem was proved in [Marí 2006] and describes the most general form of invariant evolutions in terms of left moving frames.

Theorem 4.1. *Let $u(t,x) \in G/H$ be a flow, that is, the solution of an invariant evolution of the form*

$$u_t = F(u, u_x, u_{xx}, u_{xxx}, \dots).$$

Assume the evolution is invariant under the action of G, that is, G takes solutions to solutions. Let $\rho(t,x)$ be a family of left moving frames along $u(t,x)$ such that $\rho \cdot o = u$. Then, there exists an invariant family of tangent vectors $\boldsymbol{r}(t,x)$, that is, a family depending on the differential invariants of u and their derivatives, such that

$$u_t = d\Phi_\rho(o)\boldsymbol{r}.$$

An interpretation of this theorem is as follows. If we choose coordinates and $d\Phi_\rho(o)$ is considered as an element on $\mathrm{GL}(n,\mathbb{R})$, then its columns $d\Phi_\rho(o) = (T_1, \dots, T_n)$ form a classical moving frame, that is, an invariant curve in the frame bundle. If in those coordinates $\boldsymbol{r} = (r_1, \dots, r_n)^T$, then $u_t = r_1 T_1 + \cdots + r_n T_n$ for some r_i functions of the differential invariants and their derivatives. Many readers might be more familiar with this writing of an invariant evolution, and it is equivalent to ours.

Before we describe the relation between the evolutions of u and geometric Hamiltonian evolutions, it is convenient to prove this:

Lemma 4.2. *Let $u(t,x)$ be a one-parameter family of curves in G/H. Assume $u(t,x)$ evolves following an evolution invariant under the action of G. Assume the evolution is written as*

$$u_t = d\Phi_\rho(o)\boldsymbol{r}, \tag{14}$$

where ρ is a left moving frame that can be locally factored as $\varsigma(u)\rho_H$ with $\rho_H \in H$, and where $\boldsymbol{r}$ is some invariant tangent vector.

Let $N = L^*_{\rho^{-1}}\rho_t$ be the left invariant vector field defining the evolution of ρ under (14). *Let $N = N_{\mathfrak{m}} + N_{\mathfrak{h}}$ be the splitting of N in its $\mathfrak{m}$ and $\mathfrak{h}$ component. Then $N_{\mathfrak{m}} = d\varsigma(o)\boldsymbol{r}$.*

Note $\rho_H \cdot o = o$ since $\rho_H \in H$. Using (13) we have

$$\varsigma(u)\varsigma(o) = \varsigma(u) = \varsigma(\varsigma(u) \cdot o)h,$$

which is uniquely determined for some value of $h \in H$. The choices $h = e$ and $\varsigma(u) \cdot o = u$ satisfy the equation, so we can conclude that $\varsigma(u) \cdot o = u$.

Proof. Assume $\rho = \varsigma(u)\rho_H$. If we calculate N, we have

$$N = \mathrm{Ad}(\rho_H^{-1}) L^*_{\varsigma(u)^{-1}} d\varsigma(u) u_t + L^*_{\rho_H^{-1}} d\rho_H(u) u_t.$$

Since $L^*_{\rho_H^{-1}} d\rho_H(u)u_t \in \mathfrak{h}$ we need to look only for the $\mathfrak{m}$ component of

$$\mathrm{Ad}(\rho_H^{-1}) L^*_{\varsigma(u)^{-1}} d\varsigma(u)u_t.$$

On the other hand, differentiating (13) gives

$$L^*_g d\varsigma(u)u_t = d\varsigma(\Phi_g(u))d\Phi_g(u)u_t h(u,g) + \varsigma(\Phi_g(u)dh(u)u_t.$$

Evaluating this at $g = \rho^{-1}$, we get

$$L^*_{\rho^{-1}} d\varsigma(u)u_t = R^*_{h(u,\rho^{-1})} d\varsigma(o) d\Phi_{\rho^{-1}}(u)u_t + dh(u,\rho^{-1})u_t.$$

Also from (13),

$$\rho^{-1}\varsigma(u) = \rho_H^{-1} = \varsigma(\rho^{-1}\cdot u)h(u,\rho^{-1}) = \varsigma(o)h(u,\rho^{-1}) = h(u,\rho^{-1}),$$

and $(d\Phi_{\rho^{-1}}(u))^{-1} = d\Phi_\rho(\rho^{-1}\cdot u) = d\Phi_\rho(o)$. Therefore

$$R^*_{\rho_H} L^*_{\rho^{-1}} d\varsigma(u)u_t = R^*_{\rho_H} R^*_{h(u,\rho^{-1})} d\varsigma(o)(d\Phi_\rho(o))^{-1}u_t = d\varsigma(o)\boldsymbol{r}$$

whenever u evolves as in (14). This is precisely $N_{\mathfrak{m}}$. □

In what follows we will assume the manifold to be flat, so its Cartan connection is given by the Maurer–Cartan form. If, for example, the manifold has constant curvature, some modifications can be introduced to adapt the result, much as was done in [Terng and Thorbergsson 2001; Anco 2006; Marí et al. 2002].

Theorem 4.3. *Assume that $\mathcal{K}$ is described by an affine subspace of $\mathcal{L}\mathfrak{g}^*$. Assume that* (14) *is an invariant evolution of curves on G/H and there is a Hamiltonian functional $h : \mathcal{K} \to \mathbb{R}$ such that, if $\mathcal{H} : \mathcal{L}\mathfrak{g}^* \to \mathbb{R}$ is an extension of h satisfying condition* (10), *then*

$$\frac{\delta\mathcal{H}}{\delta L}(\boldsymbol{k})_{\mathfrak{m}} = d\varsigma(o)\boldsymbol{r},$$

where

$$\frac{\delta\mathcal{H}}{\delta L}(\boldsymbol{k}) = \frac{\delta\mathcal{H}}{\delta L}(\boldsymbol{k})_{\mathfrak{m}} + \frac{\delta\mathcal{H}}{\delta L}(\boldsymbol{k})_{\mathfrak{h}}$$

are the components defined by the splitting of the algebra. Then $\mathcal{K}$ induces by (14) *an evolution Hamiltonian with respect to the first reduced Poisson bracket* (11), *with Hamiltonian functional h. In particular, any Hamiltonian evolution in $\boldsymbol{k}$ with respect to the first reduced Poisson bracket* (11) *and Hamiltonian functional $h(\boldsymbol{k})$ has a geometric realization given by*

$$u_t = d\Phi_\rho(o)d\varsigma(o)^{-1}\frac{\delta\mathcal{H}}{\delta L}(\boldsymbol{k})_{\mathfrak{m}}$$

where $\mathcal{H}$ is any extension of h satisfying (10).

Proof. Assume that an evolution of u as in (14) induces a Hamiltonian evolution on $\mathcal{K}$, with Hamiltonian functional $h : \mathcal{K} \to \mathbb{R}$. If $\mathcal{K}$ is an affine subspace of $\mathcal{L}\mathfrak{g}^*$, then K_t is a linear subspace of $\mathcal{L}\mathfrak{g}^*$. Assume $r : \mathcal{K} \to \mathbb{R}$ is any other Hamiltonian functional, and let $\mathcal{R}$ be an extension satisfying (10). Then

$$\int_{S^1} \left\langle K_t, \frac{\delta \mathcal{R}}{\delta L}(K) \right\rangle dx = \{h, r\}(K).$$

On one hand, if $\mathcal{H}$ is an extension of h satisfying (10), then

$$\{h, r\}(K) = \int_{S^1} \left\langle \left(\frac{\delta \mathcal{H}}{\delta L}(K) \right)_x + \operatorname{ad}^* \left(\frac{\delta \mathcal{H}}{\delta L}(K) \right)(K), \frac{\delta \mathcal{R}}{\delta L}(K) \right\rangle. \tag{15}$$

On the other hand, if $N = L^*_{\rho^{-1}} \rho_t$, then applying the structure equation for the Maurer–Cartan form to the commuting vector fields d/dx and d/dt along ρ results in the compatibility condition $K_t = N_x + \operatorname{ad}(K)(N)$. Therefore, we obtain that

$$\begin{aligned} \left\langle K_t, \frac{\delta \mathcal{R}}{\delta L}(K) \right\rangle &= \left\langle \left(\frac{\delta \mathcal{H}}{\delta L}(K) \right)_x + \operatorname{ad}^* \left(\frac{\delta \mathcal{H}}{\delta L}(K) \right)(K), \frac{\delta \mathcal{R}}{\delta L}(K) \right\rangle \\ &= \left\langle N_x + \operatorname{ad}^*(N)(K), \frac{\delta \mathcal{R}}{\delta L}(K) \right\rangle, \end{aligned}$$

where we are again identifying K with its dual, so that $\operatorname{ad}(K)(N) = \operatorname{ad}^*(N)(K)$. Finally, from (10), the only component involved in (15) is $\delta \mathcal{R}/\delta L_{\mathfrak{m}}$. Likewise for $\delta \mathcal{H}/\delta L$ by skew-symmetry. Therefore, if $\delta \mathcal{H}/\delta L_{\mathfrak{m}} = N_{\mathfrak{m}}$, the evolution induced on $\boldsymbol{k}$ will be Hamiltonian with Hamiltonian functional h. Using the lemma, we arrive to the conclusion of the theorem. □

In general, N and $\delta \mathcal{H}/\delta L$ are different. Only their components tangent to the manifold need to coincide.

Example 4.4. Using the data we have on $\mathrm{SO}(2,2)/H$, one can easily calculate the formula for a general invariant evolution to be

$$u_t = d\Phi_\rho(o)\boldsymbol{r} = \alpha \Theta \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} u_x^1 & 0 \\ 0 & u_x^2 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}$$

which results on the decoupling $u_t^i = u_x^i r_i$ for $i = 1, 2$, where the r_i are any functions depending on k_1, k_2 and their derivatives. The evolutions are not decoupled unless the r_i are decoupled. From the data we obtained in (12) we have

$$\frac{\delta \mathcal{H}}{\delta L}(K)_{\mathfrak{m}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \dfrac{\delta h}{\delta k_1} & 1 & 0 & 0 \\ \dfrac{\delta h}{\delta k_2} & 0 & 1 & 0 \\ 0 & -\dfrac{\delta h}{\delta k_2} & -\dfrac{\delta h}{\delta k_1} & 0 \end{pmatrix} \quad \text{and} \quad d\varsigma(o)\boldsymbol{r} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ bfr_1 & 1 & 0 & 0 \\ r_2 & 0 & 1 & 0 \\ 0 & -r_2 & -r_1 & 0 \end{pmatrix},$$

so that the condition for a geometric realization to exist is $\delta h/\delta k_i = r_i$ for $i = 1, 2$. In particular, a pair of decoupled KdV equations is obtained when

$$h(k_1, k_2) = \frac{1}{2}\int_{S^1}(k_1^2 + k_2^2)dx$$

for which $r_i = k_i$ produces a geometric realization. In the conformal case for which $G = \mathrm{O}(3, 1)$, these same choices produced a geometric realization for a complexly coupled system of KdV equations. That is, changing from $\mathrm{SO}(3, 1)$ to $\mathrm{SO}(2, 2)$ effectively decouples the system of coupled KdV equations.

5. The sphere $\mathrm{SO}(n+1)/\mathrm{SO}(n)$

In this case $G = \mathrm{SO}(n + 1)$ and $H = \mathrm{SO}(n)$ is not a parabolic subgroup. We consider the splitting $\mathfrak{o}(n + 1) = \mathfrak{m} \oplus \mathfrak{h}$ of the Lie algebra into subspaces (unlike the previous example, only $\mathfrak{h}$ is a Lie subalgebra here) with

$$\begin{pmatrix} 0 & y \\ y^T & 0 \end{pmatrix} \in \mathfrak{m} \quad \text{and} \quad \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \in \mathfrak{h}, \tag{16}$$

where $y \in \mathbb{R}^n$ and $A \in \mathfrak{o}(n)$. Associated to this splitting we have a local factorization in the group into factors belonging to $H = \mathrm{SO}(n)$ and $\exp(\mathfrak{m})$. This factorization is given by

$$g = \begin{pmatrix} \Theta & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} I + \cos_y yy^T & \sin_y y \\ -\sin_y y^T & \cos\|y\| \end{pmatrix} \tag{17}$$

where $$\cos_y = \frac{\cos\|y\| - 1}{\|y\|^2}, \quad \sin_y = \frac{\sin\|y\|}{\|y\|}, \quad \|y\|^2 = y^T y.$$

The factorization exists locally.

Let $\varsigma : M \to G$ be the section defined by the exponential, that is,

$$\varsigma(u) = \begin{pmatrix} I + \cos_u uu^T & \sin_u u \\ -\sin_u u^T & \cos\|u\| \end{pmatrix}.$$

Clearly $d\varsigma(o) : T_oM \to \mathfrak{m}$ is an isomorphism given by $d\varsigma(o)y = \left(\begin{smallmatrix} 0 & y \\ y^T & 0 \end{smallmatrix}\right)$. The action of $\mathrm{SO}(n+1)$ on the sphere — let's denote it by $g \cdot u$ — is determined by the relation $g\varsigma(u) = \varsigma(g \cdot u)h$ for some $h \in \mathrm{SO}(n)$ that is also determined by this relation. Let g be as in (17). Straightforward calculations show that if $\eta = g \cdot u$, then

$$\sin_\eta \eta = \sin_u \Theta u + (\cos_y \sin_u y^T u + \sin_y \cos\|u\|)\Theta y \tag{18}$$

$$\cos\|\eta\| = \cos\|y\| \cos\|u\| - \sin_u \sin_y y^T u. \tag{19}$$

5a. ***Left moving frames, Serret–Frenet equations and geometric Hamiltonian structures for generic curves on the sphere.***

Moving frames. With the factorization above in mind we can use normalization procedures to calculate a right moving frame along a generic curve u. Indeed, if g is as in (17), then the first normalization equation is $g \cdot u = o$, which is resolved by choosing $y = -u$. Notice that, if ς is our section, $\varsigma(u)^{-1} = \varsigma(-u)$, and $\mathrm{SO}(n)$ preserves the origin o.

The first normalization equation is given in terms of the prolonged action of the group. The action is an action on parametrized curves. Therefore, its explicit expression is found, as before, by differentiating $g \cdot u$ with respect to the parameter. If we do that and substitute $y = -u$, the first normalization equation is then

$$\sin_u \Theta u_1 + (1 - \sin_u) \frac{\|u\|_1}{\|u\|} \Theta u = s e_1$$

where $s = (\sin_u^2 \|u_x\|^2 + (1 - \sin_u^2)\|u\|_x^2)^{1/2}$ is the spherical *arc length invariant*. The vector e_1 is an arbitrary choice; any other unit vector can be chosen instead. We will not, in general, consider unparametrized curves, so this invariant is not, a priori, constant.

Subsequent normalization equations (up to order n) will determine $\Theta^{-1} e_i$ for $i = 2, \dots, n$ and with it Θ. The r-th normalization equation will be of the form $\Theta f_r(u^{(r)}) = c_r$ for some function f_r depending on u and its derivatives. The fact that $\Theta \in \mathfrak{o}(n)$ implies that the vector c_r is a function of r differential invariants of order r. Among these r differential invariants, $r - 1$ of them will be functions of lower order differential invariants and their derivatives. Hence, at each step we get a new invariant of order r that is functionally independent from those of lower order. Thus, we have n invariants or increasingly high order, the order increasing by one at each step. According to the theory developed in [Fels and Olver 1999], these would be generators of all differential invariants of the curve u. For the purpose of this example, no more details are needed.

Serret–Frenet equations and natural moving frames. First of all, the $\mathfrak{m}$ component of $\rho(\rho^{-1})_x = \widehat{K}$ is equal to $d\varsigma(0)(e_1)$, as we proved in our previous section when studying the general case.

Indeed, after some straightforward calculations,

$$\begin{aligned}
\rho(\rho^{-1})_x &= \begin{pmatrix} \Theta & 0 \\ 0 & 1 \end{pmatrix} s(u)(s(-u))_x \begin{pmatrix} \Theta^{-1} & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} \Theta(\Theta^{-1})_x & 0 \\ 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} \Theta(\Theta^{-1})_x + \Theta(\cos_u u_1 u^T - \cos_u u u_1^T)\Theta^{-1} & \sin_u \Theta u_1 + (1 - \sin_u) \frac{\|u\|_1}{\|u\|} \Theta u \\ - \sin_u u_1^T \Theta^T - (1 - \sin_u) \frac{\|u\|_1}{\|u\|} u^T \Theta^T & 0 \end{pmatrix} \\
&= \begin{pmatrix} K_0 & s e_1 \\ -s e_1^T & 0 \end{pmatrix} = \widehat{K}.
\end{aligned}$$

Theorem 5.1. *There exists a left moving frame ρ whose associated Serret–Frenet equations are given by*

$$K = \begin{pmatrix} 0 & -\boldsymbol{v}^T & s \\ \boldsymbol{v} & 0 & 0 \\ -s & 0 & 0 \end{pmatrix}, \tag{20}$$

where s is the arc length invariant and $\boldsymbol{v} = (v_i)$ are the natural curvatures. The moving frame will in general be nonlocal, and it is known as the natural moving frame (see [Bishop 1975] for the original definition).

Proof. Let ρ be our previous moving frame. Any other left moving frame will be of the form ρg, where $g \in \mathcal{L}\mathrm{SO}(n+1)$ is an invariant element of the group, that is, a matrix in $\mathrm{SO}(n+1)$ depending on the differential invariants and their derivatives. Since we do not want to change the $\mathfrak{m}$ component of the equation, we will choose $g \in \mathcal{L}H$. If the natural frame (let us call it ρ_n) exists, then $\rho_n = \rho g$ for some invariant g and $K = (\rho g)^{-1}(\rho g)_x = g^{-1}\widehat{K}g + g^{-1}g_x$. If $g = \left(\begin{smallmatrix} \theta & 0 \\ 0 & 1 \end{smallmatrix}\right)$, this relation becomes

$$K = \begin{pmatrix} \theta^T\theta_x + \theta^T K_0 \theta & s\theta^T e_1 \\ -se_1^T\theta & 0 \end{pmatrix}.$$

We want the $\mathfrak{m}$ component to remain the same, and so θ should leave e_1 invariant. That is,

$$\theta = \begin{pmatrix} 1 & 0 \\ 0 & \eta \end{pmatrix} \quad \text{for } \eta \in \mathrm{SO}(n-1).$$

Furthermore, we need

$$\theta^T\theta_x + \theta^T K_0 \theta = \begin{pmatrix} 0 & v^T \\ v & 0 \end{pmatrix}, \quad \text{that is,} \quad \eta^T\eta_x + \eta^T K_1 \eta = 0 \quad \text{for } K_0 = \begin{pmatrix} 0 & * \\ * & K_1 \end{pmatrix}.$$

In general, the solution of $\eta_x = -K_1\eta$ will be nonlocal. Also, the solution will in general have a monodromy, and it does not need to be periodic. Hence, the calculations that follow are, in that sense, formal. This situation was discussed in [Marí 2006]. □

When choose a natural moving frame, rather than a classical Riemannian one, the familiar reduced Hamiltonian structures and integrable systems emerge. Any other choice of frame gives an equivalent system, but it will not look familiar to us in general.

5a1. *Geometric Hamiltonian structures.* Finally, we will look into the reduced Poisson bracket defined on the affine subspace $\mathcal{K} \subset \mathcal{L}\mathfrak{o}(n+1)^*$ consisting of matrices of the form (20). For this example we will use as bilinear form the usual $\langle M, N\rangle = \frac{1}{2}\operatorname{tr}(MN)$. As explained in the previous section, we start by considering a Hamiltonian functional $h : \mathcal{K} \to \mathbb{R}$ and extend it to $\mathcal{H} : \mathcal{L}\mathfrak{o}(n+1)^* \to \mathbb{R}$ so that its variational derivative satisfies

$$\Big(\frac{\delta\mathcal{H}}{\delta L}(K)\Big)_x + \Big[K, \frac{\delta\mathcal{H}}{\delta L}(K)\Big] \in \mathfrak{o}(n)^0. \tag{21}$$

If we write

$$\frac{\delta\mathcal{H}}{\delta L}(K) = \begin{pmatrix} 0 & \frac{\delta h}{\delta \upsilon}^T & -\frac{\delta h}{\delta s} \\ -\frac{\delta h}{\delta \upsilon} & H_0 & \boldsymbol{v} \\ \frac{\delta h}{\delta s} & -\boldsymbol{v}^T & 0 \end{pmatrix} \tag{22}$$

for some $H_0(s, \upsilon) \in \mathfrak{o}(n-1)$ and $\boldsymbol{v}(s, \upsilon) \in \mathbb{R}^n$, then condition (21) becomes

$$\begin{pmatrix} 0 & \Big(\frac{\delta h}{\delta \upsilon}^T\Big)_x - \boldsymbol{v}^T H_0 - s\boldsymbol{v}^T & -\Big(\frac{\delta h}{\delta s}\Big)_x - \boldsymbol{v}^T\boldsymbol{v} \\ -\Big(\frac{\delta h}{\delta \upsilon}\Big)_x - H_0\boldsymbol{v} + s\boldsymbol{v} & (H_0)_x + \boldsymbol{v}\Big(\frac{\delta h}{\delta \upsilon}\Big)^T - \frac{\delta h}{\delta \upsilon}\boldsymbol{v}^T & \boldsymbol{v}_x - \frac{\delta h}{\delta s}\boldsymbol{v} + s\frac{\delta h}{\delta \upsilon} \\ \Big(\frac{\delta h}{\delta s}\Big)_x + \boldsymbol{v}^T\boldsymbol{v} & -\boldsymbol{v}_x^T - s\frac{\delta h}{\delta \upsilon}^T + \frac{\delta h}{\delta s}\boldsymbol{v}^T & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & * \\ 0 & 0 & * \\ * & * & 0 \end{pmatrix}.$$

This results in

$$\boldsymbol{v} = \frac{1}{s}\Big(\Big(\frac{\delta h}{\delta \upsilon}\Big)_x + H_0\boldsymbol{v}\Big) \quad \text{and} \quad H_0 = D^{-1}\Big(\frac{\delta h}{\delta \upsilon}\boldsymbol{v}^T - \boldsymbol{v}\Big(\frac{\delta h}{\delta \upsilon}\Big)^T\Big).$$

If $h, g : \mathcal{K} \to \mathbb{R}$ are two such functionals and the notation is as above, then the reduced bracket defined on $\mathcal{K}$ is given by

$$\begin{aligned} \{h, g\}_R(s, \upsilon) &= \int_{S^1} \Big\langle \Big(\frac{\delta\mathcal{H}}{\delta L}(K)\Big)_x + \Big[K, \frac{\delta\mathcal{H}}{\delta L}(K)\Big], \frac{\delta\mathcal{G}}{\delta L}(K) \Big\rangle\, dx \\ &= \int_{S^1} \Bigg\langle \begin{pmatrix} 0 & 0 & -\big(\frac{\delta h}{\delta s}\big)_x - \boldsymbol{v}^T\boldsymbol{v} \\ 0 & 0 & \boldsymbol{v}_x - \frac{\delta h}{\delta s}\boldsymbol{v} + s\frac{\delta h}{\delta \upsilon} \\ \big(\frac{\delta h}{\delta s}\big)_x + \boldsymbol{v}^T\boldsymbol{v} & -\boldsymbol{v}_x^T - s\frac{\delta h}{\delta \upsilon}^T + \frac{\delta h}{\delta s}\boldsymbol{v}^T & 0 \end{pmatrix}, \begin{pmatrix} 0 & \frac{\delta g}{\delta \upsilon}^T & -\frac{\delta g}{\delta s} \\ -\frac{\delta g}{\delta \upsilon} & G_0 & \boldsymbol{v}_g \\ \frac{\delta g}{\delta s} & -\boldsymbol{v}_g^T & 0 \end{pmatrix} \Bigg\rangle dx \\ &= -\int_{S^1} \frac{\delta g}{\delta s}\Big(\Big(\frac{\delta h}{\delta s}\Big)_x + \boldsymbol{v}^T\boldsymbol{v}\Big) + \boldsymbol{v}_g^T\Big(\boldsymbol{v}_x + \frac{\delta h}{\delta s}\boldsymbol{v} - s\frac{\delta h}{\delta \upsilon}\Big) dx. \end{aligned}$$

Substituting the known values for $\boldsymbol{v}$ and $\boldsymbol{v}_g$ we get an explicit expression of the first reduced Hamiltonian structure on the sphere. Let

$$\mathcal{Q}\Big(\frac{\delta h}{\delta \upsilon}\Big) = H_0\upsilon = D^{-1}\Big(\frac{\delta h}{\delta \upsilon}\upsilon^T - \upsilon\frac{\delta h}{\delta \upsilon}^T\Big)\upsilon.$$

It is known (see [Anco 2006; Terng and Thorbergsson 2001] for example) that $D + \mathcal{Q}$ defines a Poisson bracket. In terms of this operator, the reduced bracket is written as

$$\{h,g\}(s,\boldsymbol{v}) = -\int_{S^1}\begin{pmatrix}\frac{\delta g}{\delta s} & \frac{\delta g}{\delta v}\end{pmatrix}\mathcal{P}\begin{pmatrix}\delta h/\delta s\\ \delta h/\delta v\end{pmatrix}dx \tag{23}$$

where $\mathcal{P}$ is the matrix of differential operators given by

$$\mathcal{P} = \begin{pmatrix} D & \frac{1}{s}v^T D + \frac{1}{s}v^T\mathfrak{Q} \\ Dv\frac{1}{s} + \mathfrak{Q}\frac{1}{s}v & -D\frac{1}{s}D\frac{1}{s}D - D\frac{1}{s}D\frac{1}{s}\mathfrak{Q} - \mathfrak{Q}\frac{1}{s}D\frac{1}{s}D - \mathfrak{Q}\frac{1}{s}D\frac{1}{s}\mathfrak{Q} - D + \mathfrak{Q}\end{pmatrix}.$$

This bracket does not preserve arc length. In that sense it is a true bracket on parametrized curves. We will come back to this point later.

The companion bracket (6) also reduces to $\mathcal{K}$ for the value $L_0 = E_{1,n+1} - E_{n+1,1}$. Indeed

$$\begin{aligned}\{h,g\}_0(s,\boldsymbol{v}) &= \int_{S^1}\Big\langle\frac{\delta\mathcal{G}}{\delta L},\Big[L_0,\frac{\delta\mathcal{H}}{\delta L}\Big]\Big\rangle dx\\ &= \int_{S^1}\boldsymbol{v}^T\frac{\delta g}{\delta v} - \boldsymbol{v}_g^T\frac{\delta h}{\delta v} = \int_{S^1}\frac{\delta g}{\delta v}^T\mathcal{P}_0\frac{\delta h}{\delta v}dx,\end{aligned} \tag{24}$$

where the Poisson operator $\mathcal{P}_0$ is given by

$$\begin{pmatrix}0 & 0\\ 0 & \frac{1}{s}D + D\frac{1}{s} + 2\mathfrak{Q}\end{pmatrix}. \tag{25}$$

This operator, in turn, leaves the arc length parameter invariant and hence is a Poisson brackets defined on invariants of unparametrized curves. A discussion about this difference follows in the next subsection. Our last theorem has now been proved:

Theorem 5.2. *The space $\mathcal{K}$ of differential invariants of the Riemannian sphere* SO$(n+1)/$SO(n) *is a bi-Poisson manifold with compatible geometric Poisson brackets given by* (23) *and* (24).

5b. ***Geometric realizations of Hamiltonian k-evolutions, a geometric realization for a vector modified KdV evolution.*** In our final section we will describe the general formula for an invariant evolution of curves u and determine which ones are Hamiltonian with respect to (23).

Theorem 5.3. *Let $u_t = F(u, u_x, u_{xx}, \dots)$ be an invariant evolution of curves on the sphere* SO$(n+1)/$SO(n). *Let Θ be given by our right moving frame under the factorization in* (17). *Then*

$$u_t = \Big(\sin_u^{-1}\Big(I - \frac{uu^T}{\|u\|^2}\Big) + \frac{uu^T}{\|u\|^2}\Big)\Theta^{-1}\boldsymbol{r} \tag{26}$$

for some invariant vector $\boldsymbol{r}$ depending on s, v and their derivatives.

Proof. First, the action of H on the manifold is linear $(\alpha, \Theta) \cdot u = \alpha^{-1}\Theta u$. On the other hand, $\rho_H = (1, \Theta^{-1})$ and so $d\Phi_{\rho_H}(o)u = \Theta^{-1}u$. The action of $\varsigma(u)$ is slightly more complicated; we can calculate directly that

$$d\Phi_{\varsigma(u)}(o) = \sin_u^{-1}\Big(I - \frac{uu^T}{\|u\|^2}\Big) + \frac{uu^T}{\|u\|^2}$$

Following Theorem 4.1 we can straightforwardly calculate the most general form for an invariant evolution to be given by

$$u_t = \Big(\sin_u^{-1}\Big(I - \frac{uu^T}{\|u\|^2}\Big) + \frac{uu^T}{\|u\|^2}\Big)\Theta^{-1}\boldsymbol{r}$$

for some invariant vector $\boldsymbol{r}$ depending on υ, s and their derivatives. □

Theorem 5.4. *If* $u(t, x)$ *evolves following* (26), *then the differential invariants* (s, υ) *evolve following the equations*

$$\begin{aligned} s_t &= (r_1)_x - \upsilon^T\hat{\boldsymbol{r}} \\ \upsilon_t &= \frac{1}{s}\Big(\hat{\boldsymbol{r}}_{xx} + (r_1\upsilon)_x - D^{-1}\frac{1}{s}(\upsilon\hat{\boldsymbol{r}}_x^T - \hat{\boldsymbol{r}}_x\upsilon^T)\Big), \quad \textit{where } \boldsymbol{r} = \begin{pmatrix} r_1 \\ \hat{\boldsymbol{r}} \end{pmatrix}. \end{aligned}$$

Proof. We want to calculate $N = \rho^{-1}\rho_t$ whenever $\rho(x, t)$ is the natural left moving frame along the flow $u(x, t)$. Lemma 4.2 tells us that N is of the form

$$N = \rho^{-1}\rho_t = \begin{pmatrix} N_0 & \boldsymbol{r} \\ -\boldsymbol{r} & 0 \end{pmatrix}.$$

Evaluating the Maurer–Cartan structure equations along $\frac{d}{dx}$ and $\frac{d}{dt}$ implies

$$K_t = N_x + [K, N],$$

that is, with $\Upsilon = \begin{pmatrix} 0 & -\upsilon^T \\ \upsilon & 0 \end{pmatrix}$,

$$\begin{pmatrix} \Upsilon & se_1 \\ -se_1^T & 0 \end{pmatrix}_t = \begin{pmatrix} N_0 & \boldsymbol{r} \\ -\boldsymbol{r}^T & 0 \end{pmatrix}_x + \begin{pmatrix} [\Upsilon, N_0] - s(e_1\boldsymbol{r}^T - \boldsymbol{r}e_1^T) & \Upsilon\boldsymbol{r} - sN_0e_1 \\ -se_1^TN_0 + r^T\Upsilon & 0 \end{pmatrix}.$$

The $\mathfrak{m}$ component of the equation gives $s_te_1 = \boldsymbol{r}_x + \Upsilon\boldsymbol{r} - sN_0e_1$ and implies

$$N_0e_1 = \frac{1}{s}\begin{pmatrix} 0 \\ \hat{\boldsymbol{r}} + r_1\upsilon \end{pmatrix}$$

where $\boldsymbol{r} = (r_i)$ and $\hat{\boldsymbol{r}} = (r_2, r_3, \ldots, r_{n-1})^T$, and

$$s_t = (r_1)_x - \upsilon^T\hat{\boldsymbol{r}}. \tag{27}$$

The evolution $\Upsilon_t = (N_0)_x + [\Upsilon, N_0] - s(e_1\boldsymbol{r}^T - \boldsymbol{r}e_1^T)$ in the $\mathfrak{o}(n)$ block imposes conditions on N_0. Namely, if

$$N_0 = \begin{pmatrix} 0 & -\hat{\boldsymbol{r}}^T - r_1 v^T \\ \hat{\boldsymbol{r}} + r_1 v & \widehat{N}_0 \end{pmatrix}, \quad \text{then } \widehat{N}_0 = D^{-1}\frac{1}{s}(v\hat{\boldsymbol{r}}_x^T - \hat{\boldsymbol{r}}_x v^T).$$

We also get directly the evolution of v:

$$v_t = \frac{1}{s}\Big(\hat{\boldsymbol{r}}_{xx} + (r_1 v)_x - D^{-1}\frac{1}{s}(v\hat{\boldsymbol{r}}_x^T - \hat{\boldsymbol{r}}_x v^T)v\Big). \tag{28}$$ □

Finally, our last theorem is the direct translation of Theorem 4.3, having in mind the description in (22).

Theorem 5.5. *Let* (26) *be an invariant evolution such that*

$$\boldsymbol{r} = \begin{pmatrix} r_1 \\ \hat{\boldsymbol{r}} \end{pmatrix} = \begin{pmatrix} -\dfrac{\delta h}{\delta s} \\ \dfrac{1}{s}(D + \mathcal{Q})\Big(\dfrac{\delta h}{\delta v}\Big) \end{pmatrix}$$

for some Hamiltonian functional $h(s, v)$. Then (26) *induces an evolution on* (s, v) *that is Hamiltonian with respect to* (23), *with Hamiltonian functional h.*

As was pointed out in [Terng and Thorbergsson 2001] and [Anco 2006], the choice of invariant vector $r_1 = \frac{1}{2}\|v\|^2$ and $\hat{\boldsymbol{r}} = v_x$ results in an arc length preserving evolution ($s_t = 0$, we will assume $s = 1$) given by

$$v_t = v_{xxx} + \tfrac{3}{2}\|v\|^2 v_x$$

that is, the vector modified KdV equation.

The final question is whether or not the modified KdV equation is bi-Hamiltonian with respect to the two compatible Poisson brackets we found. Our previous general theorem 4.3 states that the condition for the evolution to be Hamiltonian is the existence of a Hamiltonian $h : \mathcal{K} \to \mathbb{R}$ and an extension $\mathcal{H} : \mathcal{L}\mathfrak{g}^* \to \mathbb{R}$ such that $\delta\mathcal{H}/\delta L_{\mathfrak{m}} = d\varsigma(o)\boldsymbol{r}$. Using (22), this condition is equivalent to

$$-\frac{\delta h}{\delta s} = r_1 = \tfrac{1}{2}\|v\|^2 \quad \text{and} \quad \boldsymbol{v} = \Big(\frac{\delta h}{\delta v}\Big)_x + \mathcal{Q}\Big(\frac{\delta h}{\delta v}\Big) = \hat{\boldsymbol{r}} = v_x.$$

Notice that the second relation is satisfied by $\delta h/\delta v = v$.

Consider the Hamiltonian functional

$$h(s, v) = \int_{S^1} -\tfrac{1}{2}\|v\|^2 s + \|v\|^2.$$

Clearly,

$$\frac{\delta h}{\delta s} = -\tfrac{1}{2}\|v\|^2 \quad \text{and} \quad \frac{\delta h}{\delta v} = (2 - s)v.$$

On the preserved level set $s = 1$, the Hamiltonian has the desired properties.

Finally, the vector modified KdV equation is also Hamiltonian with respect to our second reduced Poisson bracket. If we consider as Hamiltonian the operator $h_0 : \mathcal{K} \to \mathbb{R}$ given by

$$h_0(v) = \tfrac{1}{2} \int_{S^1} -\|v_x\|^2 + \tfrac{1}{4}\|v\|^4,$$

then

$$v_t = v_{xxx} + \tfrac{3}{2}\|v\|^2 = \mathcal{P}_0(v_{xx} + \tfrac{1}{2}\|v\|^2 v) = \mathcal{P}_0 \frac{\delta h_0}{\delta v}.$$

Therefore, the modified KdV vector equation is bi-Hamiltonian with respect to both brackets as long as we assume the parameter to be the spherical arc length. This condition is forced upon the equations if we want the equations to be Hamiltonian with respect to the second reduced bracket (24). The second bracket, but not the first, appeared already in [Terng and Thorbergsson 2001; Anco 2006].

The role of invariants of arc length type was studied in [Marí 2007] in the case of affine geometries, which are manifolds of the form $G \ltimes \mathbb{R}^n / G$. Among the classical Lie groups, all manifolds except $G = \mathrm{GL}(n)$ have a common feature: their first geometric Poisson bracket (11) always preserves an invariant of arc length type — they are brackets associated to unparametrized curves. Therefore, any Hamiltonian evolution will have geometric realizations by evolutions that preserve arc length type parameters. This is not a choice, but is imposed by the background geometry. On the other hand, homogeneous manifolds of the form G/H in general do not have this property. All known examples have a geometric Poisson bracket defined as in (11) that does not preserve a parameter of arc length type as defined in [Marí 2009]. On the other hand, the modified KdV equation is usually associated to Riemannian manifolds in general, and to natural frames in particular; it is always the invariantization of a curve evolution parametrized by arc length. Thus, it seemed contradictory that it appears on manifolds of the form G/H, with G semisimple; at the very least, it seemed counterintuitive. As we saw in our example, the imposition of arc length preservation does not come from the first geometric bracket, but from the second. The first bracket does not preserve arc length, in agreement with all other examples of the type G/H, but the second does, in agreement with modified KdV being an evolution associated to evolutions that do so.

References

[Anco 2006] S. C. Anco, "Hamiltonian flows of curves in $G/\mathrm{SO}(N)$ and vector soliton equations of mKdV and sine-Gordon type", *SIGMA Symmetry Integrability Geom. Methods Appl.* **2** (2006), Paper 044. MR 2007a:37073 Zbl 1102.37042

[Bishop 1975] R. L. Bishop, "There is more than one way to frame a curve", *Amer. Math. Monthly* **82** (1975), 246–251. MR 51 #6604 Zbl 0298.53001

[Cartan 1935] E. Cartan, *La méthode du repère mobile, la théorie des groupes continus, et les espaces généralisés*, Exposés de Géométrie **5**, Hermann, Paris, 1935. JFM 61.0785.01

[Cartan 1937] E. Cartan, *La théorie des groupes finis et continus et la géométrie différentielle traitées par la méthode du repère mobile*, Cahiers Scientifiques **18**, Gauthier-Villars, Paris, 1937. Zbl 0018.29804 JFM 63.1227.02

[Chou and Qu 2002] K.-S. Chou and C. Qu, "Integrable equations arising from motions of plane curves", *Phys. D* **162**:1-2 (2002), 9–33. MR 2003c:37106 Zbl 0987.35139

[Chou and Qu 2003] K.-S. Chou and C.-Z. Qu, "Integrable equations arising from motions of plane curves, II", *J. Nonlinear Sci.* **13**:5 (2003), 487–517. MR 2005d:37155 Zbl 1045.35063

[Doliwa and Santini 1994] A. Doliwa and P. M. Santini, "An elementary geometric characterization of the integrable motions of a curve", *Phys. Lett. A* **185**:4 (1994), 373–384. MR 95b:58075 Zbl 0941.37532

[Fels and Olver 1999] M. Fels and P. J. Olver, "Moving coframes, II: Regularization and theoretical foundations", *Acta Appl. Math.* **55**:2 (1999), 127–208. MR 2000h:58024 Zbl 0937.53013

[Ferapontov 1995] E. V. Ferapontov, "Isoparametric hypersurfaces in spheres, integrable nondiagonalizable systems of hydrodynamic type, and N-wave systems", *Differential Geom. Appl.* **5**:4 (1995), 335–369. MR 97c:58047 Zbl 0872.53005

[Gay-Balmaz and Ratiu 2007] F. Gay-Balmaz and T. S. Ratiu, "Group actions on chains of Banach manifolds and applications to fluid dynamics", *Ann. Global Anal. Geom.* **31**:3 (2007), 287–328. MR 2008g:58012 Zbl 1122.58006

[Green 1978] M. L. Green, "The moving frame, differential invariants and rigidity theorems for curves in homogeneous spaces", *Duke Math. J.* **45**:4 (1978), 735–779. MR 80a:53011 Zbl 0414.53039

[Griffiths 1974] P. Griffiths, "On Cartan's method of Lie groups and moving frames as applied to uniqueness and existence questions in differential geometry", *Duke Math. J.* **41** (1974), 775–814. MR 53 #14355 Zbl 0294.53034

[Hasimoto 1972] R. Hasimoto, "A soliton on a vortex filament", *J. Fluid Mechanics* **51** (1972), 477–485.

[Hubert 2007] E. Hubert, "Generation properties of Maurer–Cartan invariants", preprint 00194528, INRIA, 2007, available at http://hal.inria.fr/inria-00194528/en.

[Langer and Perline 1991] J. Langer and R. Perline, "Poisson geometry of the filament equation", *J. Nonlinear Sci.* **1**:1 (1991), 71–93. MR 92k:58118 Zbl 0795.35115

[Langer and Perline 2000] J. Langer and R. Perline, "Geometric realizations of Fordy–Kulish nonlinear Schrödinger systems", *Pacific J. Math.* **195**:1 (2000), 157–178. MR 2001i:37114 Zbl 1115.37353

[Magri 1978] F. Magri, "A simple model of the integrable Hamiltonian equation", *J. Math. Phys.* **19**:5 (1978), 1156–1162. MR 80a:35112 Zbl 0383.35065

[Marí 2005] G. Marí Beffa, "Poisson brackets associated to the conformal geometry of curves", *Trans. Amer. Math. Soc.* **357**:7 (2005), 2799–2827. MR 2006a:37064 Zbl 1081.37042

[Marí 2006] G. Marí Beffa, "Poisson geometry of differential invariants of curves in some nonsemisimple homogeneous spaces", *Proc. Amer. Math. Soc.* **134** (2006), 779–791. MR 2006f:53122 Zbl 1083.37053

[Marí 2007] G. Marí Beffa, "On completely integrable geometric evolutions of curves of Lagrangian planes", *Proc. Roy. Soc. Edinburgh Sect. A* **137**:1 (2007), 111–131. MR 2008i:37144 Zbl 1130.37032

[Marí 2008a] G. Marí Beffa, "Geometric Hamiltonian structures on flat semisimple homogeneous manifolds", *Asian J. Math.* **12**:1 (2008), 1–33. MR 2009f:37077 Zbl 1173.37054

[Marí 2008b] G. Marí Beffa, "Projective-type differential invariants and geometric curve evolutions of KdV-type in flat homogeneous manifolds", *Ann. Inst. Fourier* (*Grenoble*) **58**:4 (2008), 1295–1335. MR 2009d:37126 Zbl 05303676

[Marí 2009] G. Marí Beffa, "Hamiltonian evolution of curves in classical affine geometries", *Phys. D* **238**:1 (2009), 100–115. MR 2571970 Zbl 1163.37023

[Marí et al. 2002] G. Marí Beffa, J. A. Sanders, and J. P. Wang, "Integrable systems in three-dimensional Riemannian geometry", *J. Nonlinear Sci.* **12**:2 (2002), 143–167. MR 2003f:37137 Zbl 1140.37361

[Marsden and Ratiu 1986] J. E. Marsden and T. Ratiu, "Reduction of Poisson manifolds", *Lett. Math. Phys.* **11**:2 (1986), 161–169. MR 87h:58067 Zbl 0602.58016

[Pressley and Segal 1989] A. Pressley and G. Segal, *Loop groups*, 2nd ed., Oxford University Press, New York, 1989. MR 88i:22049

[Sanders and Wang 2003] J. A. Sanders and J. P. Wang, "Integrable systems in n-dimensional Riemannian geometry", *Mosc. Math. J.* **3**:4 (2003), 1369–1393. MR 2004m:37142 Zbl 1050.37035

[Terng and Thorbergsson 2001] C.-L. Terng and G. Thorbergsson, "Completely integrable curve flows on adjoint orbits", *Results Math.* **40**:1-4 (2001), 286–309. MR 2002k:37141 Zbl 1023.37041

[Terng and Uhlenbeck 2000] C.-L. Terng and K. Uhlenbeck, "Bäcklund transformations and loop group actions", *Comm. Pure Appl. Math.* **53**:1 (2000), 1–75. MR 2000k:37116 Zbl 1031.37064

[Terng and Uhlenbeck 2006] C.-L. Terng and K. Uhlenbeck, "Schrödinger flows on Grassmannians", pp. 235–256 in *Integrable systems, geometry, and topology*, edited by C.-L. Terng, AMS/IP Stud. Adv. Math. **36**, Amer. Math. Soc., Providence, RI, 2006. MR 2007a:37079 Zbl 1110.37056

[Yasui and Sasaki 1998] Y. Yasui and N. Sasaki, "Differential geometry of the vortex filament equation", *J. Geom. Phys.* **28**:1-2 (1998), 195–207. MR 99m:58119 Zbl 0982.37072

Received December 3, 2008.

GLORIA MARÍ BEFFA
MATHEMATICS DEPARTMENT
UNIVERSITY OF WISCONSIN
MADISON, WI 53706
UNITED STATES

maribeff@math.wisc.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

CLOSED ORBITS OF A CHARGE IN A WEAKLY EXACT MAGNETIC FIELD

WILL J. MERRY

We prove that for a weakly exact magnetic system on a closed connected Riemannian manifold, almost all energy levels contain a closed orbit. More precisely, we prove the following stronger statements. Let (M, g) denote a closed connected Riemannian manifold and $\sigma \in \Omega^2(M)$ a weakly exact 2-form. Let $\phi_t : TM \to TM$ denote the magnetic flow determined by σ, and let $c(g, \sigma) \in \mathbb{R} \cup \{\infty\}$ denote the Mañé critical value of the pair (g, σ). We prove that if $k > c(g, \sigma)$, then for every nontrivial free homotopy class of loops on M there exists a closed orbit of ϕ_t with energy k whose projection to M belongs to that free homotopy class. We also prove that for almost all $k < c(g, \sigma)$ there exists a closed orbit of ϕ_t with energy k whose projection to M is contractible. In particular, when $c(g, \sigma) = \infty$ this implies that almost every energy level has a contractible closed orbit. As a corollary we deduce that a weakly exact magnetic flow with $[\sigma] \neq 0$ on a manifold with amenable fundamental group (which implies $c(g, \sigma) = \infty$) has contractible closed orbits on almost every energy level.

1. Introduction

Let (M, g) denote a closed connected d-dimensional Riemannian manifold, with tangent bundle $\pi : TM \to M$ and universal cover $\widetilde{M}$. We will assume M is not simply connected, as otherwise $\widetilde{M} = M$ and all results proved in this paper reduce to special cases of the results in [Contreras 2006]. Let $\sigma \in \Omega^2(M)$ denote a closed 2-form, and let $\tilde{\sigma} \in \Omega^2(\widetilde{M})$ denote its pullback to the universal cover. In this paper we consider the case where σ is weakly exact, that is, when $\tilde{\sigma}$ is exact (this is equivalent to requiring that $\sigma|_{\pi_2(M)} = 0$); however we do not assume that $\tilde{\sigma}$ necessarily admits a bounded primitive.

Let ω_g denote the standard symplectic form on TM obtained by pulling back the canonical symplectic form $dq \wedge dp$ on T^*M via the Riemannian metric. Let $\omega := \omega_g + \pi^*\sigma$ denote the *twisted symplectic form* determined by the pair (g, σ). Let $E : TM \to \mathbb{R}$ denote the energy Hamiltonian $E(q, v) = \frac{1}{2}|v|^2$. Let $\phi_t : TM \to TM$

MSC2000: 37J45, 70H12.

Keywords: magnetic flow, twisted geodesic flow, periodic orbits, Mañé critical value.

denote the flow of the symplectic gradient of E with respect to ω; such ϕ_t is known as a *twisted geodesic flow* or a *magnetic flow*. The reason for the latter terminology is that this flow can be thought of as modeling the motion of a particle of unit mass and unit charge under the effect of a magnetic field represented by the 2-form σ. Given $k \in \mathbb{R}^+ := \{t \in \mathbb{R} : t > 0\}$, let $\Sigma_k := E^{-1}(k) \subseteq TM$.

There exists a number $c = c(g, \sigma) \in \mathbb{R} \cup \{\infty\}$, the *Mañé critical value* (see [Mañé 1996; Contreras et al. 1997; Contreras and Iturriaga 1999; Burns and Paternain 2002] or Section 2 for the precise definition), such that the dynamics of the hypersurface Σ_k differs dramatically depending on whether $k < c$, $k = c$ or $k > c$. Moreover $c < \infty$ if and only if $\tilde{\sigma}$ admits a bounded primitive.

In this paper we study the old problem of the existence of closed orbits on prescribed energy levels. In the case when σ is exact, this has been essentially solved by Contreras [2006]; see Theorem D therein in particular, which gives contractible closed orbits in almost every energy level below the Mañé critical value, and closed orbits in every free homotopy class for any energy level above the critical value. In the case of surfaces a stronger result is known to hold: Contreras, Macarini and Paternain have proved in [Contreras et al. 2004, Theorem 1.1] that in this case *every* energy level admits a closed orbit. However the case of a magnetic monopole (that is, when σ is not exact) remains open, although much progress has been made. Let us describe now some of these results. A more comprehensive survey can be found in the introduction to [Contreras et al. 2004]; see also [Ginzburg 1996] for a introductory account of the problem.

Macarini [2004], extending an earlier result of Polterovich [1998], proved that if $[\sigma] \neq 0$ there exist nontrivial contractible closed orbits of the magnetic flow in a sequence of arbitrarily small energy levels. Kerman [2000] proved the same result for magnetic fields given by symplectic forms. This was then sharpened by Ginzburg and Gürel [2009] and finally by Usher [2009], where it is proved that when σ is symplectic, contractible closed orbits exist for all low energy levels. See also [Lu 2006] for another interesting approach to the problem in the case of symplectic σ. Perhaps the most general result so far is due to Schlenk [2006], who showed that for any closed 2-form (not necessarily weakly exact), almost every sufficiently small energy level contains a contractible closed orbit.

This paper extends [Contreras 2006, Theorem D] to the weakly exact case.

Theorem 1.1. *Let (M, g) denote a closed connected Riemannian manifold, and let $\sigma \in \Omega^2(M)$ denote a closed 2-form whose pullback to the universal cover $\widetilde{M}$ is exact. Let $c = c(g, \sigma) \in \mathbb{R} \cup \{\infty\}$ denote the Mañé critical value, and let ϕ_t denote the magnetic flow defined by σ.*

(1) *If $c < \infty$, then for all $k > c$ and for each nontrivial homotopy class $\nu \in [\mathbb{T}, M]$, there is a closed orbit of ϕ_t with energy k whose projection to M belongs to ν.*

(2) *For almost all $k \in (0, c)$, where possibly $c = \infty$, there exists a contractible closed orbit of ϕ_t with energy k.*

Theorem 1.1(1) has, under a mild technical assumption on $\pi_1(M)$, been proved by Paternain [2006]. We use a completely different method of proof however, which bypasses the need for this additional assumption. For $c(g, \sigma) < \infty$, Theorem 1.1(2) is due to Osuna [2005]; we believe the main contribution of this paper is the case $c(g, \sigma) = \infty$.

Remark. We will actually prove a slightly stronger statement than the one stated above; see Proposition 5.8 below for details.

When $\pi_1(M)$ is amenable and σ is not exact, we always have $c(g, \sigma) = \infty$; see for instance [Paternain 2006, Corollary 5.4]. Thus the following corollary is immediate.

Corollary 1.2. *Let (M, g) denote a closed connected Riemannian manifold, and let $\sigma \in \Omega^2(M)$ denote a closed nonexact 2-form whose pullback to the universal cover $\widetilde{M}$ is exact. Suppose $\pi_1(M)$ is amenable. Then almost every energy level contains a contractible closed orbit of the magnetic flow defined by σ.*

Let us now give a brief outline of our method of attack. Fix a primitive θ of $\tilde{\sigma}$, and consider the Lagrangian $L : T\widetilde{M} \to \mathbb{R}$ defined by

$$L(q, v) := \tfrac{1}{2}|v|^2 - \theta_q(v).$$

The Euler–Lagrange flow of L is precisely the lifted flow $\tilde{\phi}_t : T\widetilde{M} \to T\widetilde{M}$ of the magnetic flow $\phi_t : TM \to TM$; see for example [Contreras and Iturriaga 1999]. Recall that the *action* $A(y)$ of the Lagrangian L over an absolutely continuous curve $y : [0, T] \to \widetilde{M}$ is given by

$$A(y) := \int_0^T L(y(t), \dot{y}(t))dt = \int_0^T \tfrac{1}{2}|\dot{y}(t)|^2 dt - \int_y \theta.$$

Set

$$A_k(y) := \int_0^T (L(y(t), \dot{y}(t)) + k)dt = A(y) + kT.$$

A closed orbit of $\tilde{\phi}_t$ with energy k can be realized as a critical point of the functional $y \mapsto A_k(y)$. More precisely, let $\Lambda_{\widetilde{M}}$ denote the Hilbert manifold of absolutely continuous square integrable curves $x : \mathbb{T} \to \widetilde{M}$ and consider $\tilde{S}_k : \Lambda_{\widetilde{M}} \times \mathbb{R}^+ \to \mathbb{R}$ defined by

$$\tilde{S}_k(x, T) := \int_0^1 T \cdot L(x(t), \dot{x}(t)/T)dt + kT = \int_0^1 \frac{1}{2T}|\dot{x}(t)|^2 dt + kT - \int_x \theta.$$

Then the pair (x, T) is a critical point $\tilde{S}_k$ if and only if $y(t) := x(t/T)$ is the projection to $\widetilde{M}$ of a closed orbit of $\tilde{\phi}_t$ with energy k; see [Contreras et al. 2000].

If σ was actually exact then we could define L on TM, instead of just on $T\widetilde{M}$. In this case it has been shown in [Contreras et al. 2000] that $\tilde{S}_k$ for $k > c(g, \sigma)$ satisfies the Palais–Smale condition and is bounded below. Standard results from Morse theory [Contreras et al. 2000, Corollary 23] then tell us that $\tilde{S}_k$ admits a global minimum, and this gives us our desired closed orbit. In [Contreras 2006] this was extended to give contractible orbits on almost every energy level below the critical value. Crucially however, these results use compactness of M and hence are not applicable directly in the weakly exact case, since then L is defined only on $T\widetilde{M}$.

In the weakly exact case, whilst $\tilde{S}_k$ is not well defined on TM, its differential is. This leads to our key observation that we can still work directly on Λ_M. More precisely, we define a functional $S_k : \Lambda_M \times \mathbb{R}^+ \to \mathbb{R}$ with[1] the property that (x, T) is a critical point of S_k if and only if a lift $\tilde{y}$ to $\widetilde{M}$ of the curve $y(t) := x(t/T)$ is the projection to $\widetilde{M}$ of a flow line of $\tilde{\phi}_t$ with energy k. The functional S_k is given by

$$S_k(x, T) := \int_0^1 \frac{1}{2T}|\dot{x}(t)|^2 dt + kT - \int_{C(x)} \sigma,$$

where $C(x)$ is any cylinder with boundary $x(\mathbb{T}) \cup x_\nu(\mathbb{T})$, where $x_\nu \in \Lambda_M$ is some fixed reference loop in the free homotopy class $\nu \in [\mathbb{T}, M]$ that x belongs to. If $c(g, \sigma) < \infty$, then since σ is weakly exact, the value $\int_{C(x)} \sigma$ is independent of the choice of cylinder $C(x)$ for any curve $x \in \Lambda_M$. In the case $c(g, \sigma) = \infty$, the value $\int_{C(x)} \sigma$ is independent of the choice of cylinder only when x is a contractible loop.

The functional S_k allows one to extend other results previously known only for the exact case to the weakly exact case. For instance, in [Merry 2010] we will use S_k to establish the short exact sequence [Cieliebak et al. 2010; Abbondandolo and Schwarz 2009a] between the Rabinowitz Floer homology of a weakly exact twisted cotangent bundle and the singular (co)homology of the free loop space.

2. Preliminaries

The setup. It will be convenient to view M and $\widetilde{M}$ as being embedded isometrically in some $\mathbb{R}^N$ (which is possible by Nash's theorem). We will be interested in various spaces of absolutely continuous curves.

Given $q_0, q_1 \in M$ and $T \geq 0$, let $C^{\mathrm{ac}}_M(q_0, q_1; T)$ denote the set of absolutely continuous curves $y : [0, T] \to M$ with $y(0) = q_0$ and $y(T) = q_1$. Let

$$C^{\mathrm{ac}}_M(q_0, q_1) := \bigcup_{T \geq 0} C^{\mathrm{ac}}_M(q_0, q_1; T).$$

[1] If $\tilde{\sigma}$ does not admit any bounded primitives, S_k is only defined on $\Lambda_0 \times \mathbb{R}^+$, where $\Lambda_0 \subseteq \Lambda_M$ is the subset of contractible loops.

We can repeat the construction on $\widetilde{M}$ to obtain for $q_0, q_1 \in \widetilde{M}$ sets $C^{\mathrm{ac}}_{\widetilde{M}}(q_0, q_1; T)$ and $C^{\mathrm{ac}}_{\widetilde{M}}(q_0, q_1)$ of curves on $\widetilde{M}$.

Next, consider the space $W^{1,2}(\mathbb{R}^N)$ of absolutely continuous maps $x : I \to \mathbb{R}^N$ such that $\int_0^1 |\dot{x}(t)|^2 dt < \infty$ and the space

$$W^{1,2}(M) := \{x \in W^{1,2}(\mathbb{R}^N) : x(I) \subseteq M\},$$

with $W^{1,2}(\widetilde{M})$ defined similarly. Here and throughout, $I := [0, 1]$.

Let $\Lambda_{\mathbb{R}^N} \subseteq W^{1,2}(\mathbb{R}^N)$ denote the set of closed loops of class $W^{1,2}$ on $\mathbb{R}^N$, and let $\Lambda_M := W^{1,2}(M) \cap \Lambda_{\mathbb{R}^N}$. We will think of maps $x \in \Lambda_M$ as maps $x : \mathbb{T} \to M$ (here $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, which we shall often identify with S^1). Given a free homotopy class $\nu \in [\mathbb{T}, M]$, let $\Lambda_\nu \subseteq \Lambda_M$ denote the connected component of Λ_M consisting of the loops belonging to ν.

The tangent space to $\Lambda_{\mathbb{R}^N}$ at $x \in \Lambda_{\mathbb{R}^N}$ is given by

$$T_x \Lambda_{\mathbb{R}^N} = \{\xi \in W^{1,2}(\mathbb{R}^N) : \xi(0) = \xi(1)\}.$$

Given $(x, T) \in \Lambda_M \times \mathbb{R}^+$, we thus have

$$T_{(x,T)}(\Lambda_M \times \mathbb{R}^+) = \{(\xi, \psi) \in W^{1,2}(\mathbb{R}^N) \times \mathbb{R} : \xi(0) = \xi(1)\}.$$

Let $\langle \cdot , \cdot \rangle$ denote the standard Euclidean metric. The metric on $W^{1,2}(\mathbb{R}^N)$ we will work with is

$$\langle \xi, \zeta \rangle_{1,2} := \langle \xi(0), \zeta(0) \rangle + \int_0^1 \langle \dot{\xi}(t), \dot{\zeta}(t) \rangle \, dt.$$

This defines a metric that we shall denote simply by $\langle \cdot , \cdot \rangle$ on $W^{1,2}(\mathbb{R}^N) \times \mathbb{R}^+$ by

$$\langle (\xi, \psi), (\zeta, \chi) \rangle := \langle \xi, \zeta \rangle_{1,2} + \psi \chi. \tag{2-1}$$

Mañé's critical value. We now recall the definition of $c(g, \sigma)$, the *critical value* introduced in [Mañé 1996], which plays a decisive role in all that follows.

Let us fix a primitive θ of $\tilde{\sigma}$. Given $k \in \mathbb{R}^+$, we define A_k as follows. Let $q_0, q_1 \in \widetilde{M}$. Define $A_k : C^{\mathrm{ac}}_{\widetilde{M}}(q_0, q_1) \to \mathbb{R}$ by

$$A_k(y) := \int_0^T \tfrac{1}{2} |\dot{y}(t)|^2 + kT - \int_y \theta.$$

We define *Mañé's action potential* $m_k : \widetilde{M} \times \widetilde{M} \to \mathbb{R} \cup \{-\infty\}$ by

$$m_k(q_0, q_1) := \inf_{T > 0} \inf_{y \in C^{\mathrm{ac}}_{\widetilde{M}}(q_0, q_1; T)} A_k(y).$$

Then we have the following result; for a proof see [Contreras and Iturriaga 1999, Proposition 2-1.1] for the first five statements, and [Burns and Paternain 2002, Appendix A] for a proof of the last statement.

Lemma 2.1. *Properties of m_k:*

(1) *If $k \le k'$, then $m_k(q_0, q_1) \le m_{k'}(q_0, q_1)$ for all $q_0, q_1 \in \widetilde{M}$.*

(2) *For all $k \in \mathbb{R}$ and all $q_0, q_1, q_2 \in \widetilde{M}$, we have*

$$m_k(q_0, q_2) \le m_k(q_0, q_1) + m_k(q_1, q_2).$$

(3) *Fix $k \in \mathbb{R}$. Then either $m_k(q_0, q_1) = -\infty$ for all $q_0, q_1 \in \widetilde{M}$, or $m_k(q_0, q_1) \in \mathbb{R}$ for all $q_0, q_1 \in \widetilde{M}$ and $m_k(q, q) = 0$ for all $q \in \widetilde{M}$.*

(4) *If*

$$c(g, \sigma) := \inf\{k \in \mathbb{R} : m_k(q_0, q_1) \in \mathbb{R} \text{ for all } q_0, q_1 \in \widetilde{M}\},$$

then $m_{c(g,\sigma)}$ is finite everywhere.

(5) *We can alternatively define $c(g, \sigma)$ by*

$$c(g, \sigma) = \inf_{u \in C^\infty(\widetilde{M})} \sup_{q \in \widetilde{M}} \tfrac{1}{2}|d_q u + \theta_q|^2. \tag{2-2}$$

We call the number $c(g, \sigma)$ the *Mañé critical value*. Using (2-2) it is clear that $c(g, \sigma) < \infty$ if and only if θ is bounded, that is, if

$$\|\theta\|_\infty := \sup_{q \in \widetilde{M}} |\theta_q| < \infty. \tag{2-3}$$

The functional S_k. We will now define a second functional S_k, which will be our main object of study. In the case $c(g, \sigma) < \infty$, it is defined on $\Lambda_M \times \mathbb{R}^+$. For $c(g, \sigma) = \infty$, it is only defined on $\Lambda_0 \times \mathbb{R}^+$. The following lemma is the key observation required to define S_k. In the statement, $\mathbb{T}^2$ denotes the 2-torus.

Lemma 2.2. *If $c(g, \sigma) < \infty$, then $f^*\sigma$ is exact for any smooth map $f : \mathbb{T}^2 \to M$.*

Proof. Consider $G := f_*(\pi_1(\mathbb{T}^2)) \le \pi_1(M)$. Then G is amenable since $\pi_1(\mathbb{T}^2) = \mathbb{Z}^2$, which is amenable. Then [Paternain 2006, Lemma 5.3] says that since $\|\theta\|_\infty < \infty$, we can replace θ by a G-invariant primitive θ' of $\tilde{\sigma}$, which descends to define a primitive $\theta'' \in \Omega^1(\mathbb{T}^2)$ of $f^*\sigma$. □

For each free homotopy class $\nu \in [\mathbb{T}, M]$, pick a reference loop $x_\nu \in \Lambda_\nu$. Given any $x \in \Lambda_\nu$, let $C(x)$ denote a cylinder with boundary $x(\mathbb{T}) \cup x_\nu(\mathbb{T})$.

Define $S_k : \Lambda_\nu \times \mathbb{R}^+ \to \mathbb{R}$ by

$$S_k(x, T) := \int_0^1 \frac{1}{2T}|\dot{x}(t)|^2 dt + kT - \int_{C(x)} \sigma,$$

This is well defined because $\int_{C(x)} \sigma$ is independent of the choice of cylinder: If $C'(x)$ is another cylinder with the same boundary, then $\mathbb{T}^2(x) := C(x) \cup \overline{C'(x)}$ is a torus (where $\overline{C'(x)}$ denotes the cylinder $C'(x)$ taken with the opposite orientation), and $\int_{\mathbb{T}^2(x)} \sigma = 0$ since $\sigma|_{\mathbb{T}^2(x)}$ is exact by the previous lemma.

If $c(g,\sigma)=\infty$, we cannot define S_k on all of $\Lambda_M \times \mathbb{R}^+$, since in this case Lemma 2.2 fails. It is however well defined on $\Lambda_0 \times \mathbb{R}^+$. To see why, consider the case of contractible loops when $c(g,\sigma)<\infty$ again. If $x:\mathbb{T}\to M$ is contractible and $\boldsymbol{x}:D^2\to M$ denotes a capping disc, so that $\boldsymbol{x}|_{\partial D^2}=x$, it is easy to see that

$$\int_{C(x)} \sigma = \int_{D^2} \boldsymbol{x}^*\sigma; \tag{2-4}$$

note that the right side is (as it should be) independent of the choice of capping disc $\boldsymbol{x}$, and depends only on x and σ, since $\sigma|_{\pi_2(M)}=0$. Moreover the right side is well defined and depends only on x and σ even when $c(g,\sigma)=\infty$. Thus it makes sense to *define* $S_k|_{\Lambda_0\times\mathbb{R}^+}$ by

$$S_k(x,T)=\int_0^1 \frac{1}{2T}|\dot{x}(t)|^2dt+kT-\int_{D^2}\boldsymbol{x}^*\sigma;$$

this is consistent with the previous definition of $S_k|_{\Lambda_0\times\mathbb{R}^+}$ when $c(g,\sigma)<\infty$.

Next we will explicitly calculate the derivative of S_k. Let (x_s,T_s) be a variation of (x,T), with $\xi(t):=\frac{\partial}{\partial s}\big|_{s=0}x_s(t)$ and $\psi:=\frac{\partial}{\partial s}\big|_{s=0}T_s$. Write E_q and E_v for $\frac{\partial E}{\partial q}$ and $\frac{\partial E}{\partial v}$ respectively. Then an easy calculation in local coordinates shows that the first variation (that is, the Gateaux derivative) of S_k at (ξ,ψ), that is, $\frac{\partial}{\partial s}\big|_{s=0}S_k(x_s,T_s)$ is given by

$$\begin{aligned}\frac{\partial}{\partial s}\bigg|_{s=0} & S_k(x_s,T_s) \\ &=\psi\int_0^1(k-E(x(t),\dot{x}(t)/T))dt+\int_0^1\sigma_{x(t)}(\xi(t),\dot{x}(t))dt \\ &\qquad+\int_0^1\big(T\cdot E_q(x(t),\dot{x}(t)/T)\cdot\xi(t)+E_v(x(t),\dot{x}(t)/T)\cdot\dot{\xi}(t)\big)dt.\end{aligned}\tag{2-5}$$

We claim now that S_k is differentiable with respect to the canonical Hilbert manifold structure of $\Lambda_\nu\times\mathbb{R}^+$ (that is, S_k is Fréchet differentiable). In fact, S_k is of class C^2. For this we quote the fact that

$$(x,T)\mapsto\int_0^1\frac{1}{2T}|\dot{x}(t)|^2dt+kT$$

is of class C^2 (see for instance [Abbondandolo and Schwarz 2009b]) and thus is remains to check that $x\mapsto\int_{C(x)}\sigma$ is differentiable. This can be checked directly. It thus follows that the first variation $\frac{\partial}{\partial s}\big|_{s=0}S_k(x_s,T_s)$ is actually equal to the (Fréchet) derivative $d_{(x,T)}S_k(\xi,\psi)$.

Finally, let us note that

$$\frac{\partial}{\partial T}S_k(x,T)=\frac{1}{T}\int_0^T(k-E(y,\dot{y}))dt,\qquad\text{where } y(t):=x(t/T).\tag{2-6}$$

Relating S_k and A_k. Next, if (x, T) is a critical point of S_k, then $y(t) := x(t/T)$ satisfies

$$\int_0^T \Big(E_q(y, \dot y) - \frac{d}{dt}E_v(y, \dot y)\Big)\zeta dt - \frac{1}{T}\int_0^T \sigma_y(\zeta, \dot y)dt = 0,$$

where $\zeta(t) = \xi(t/T)$. Since this holds for all variations ζ, this implies that if $\tilde y : [0, T] \to \widetilde M$ is a lift of y, then $\tilde y$ satisfies the Euler–Lagrange equations for L, that is,

$$L_q(\tilde y, \dot{\tilde y}) - \frac{d}{dt}L_v(\tilde y, \dot{\tilde y}) = 0.$$

Thus $\tilde y$ is the lift to $\widetilde M$ of the projection to M of an orbit of ϕ_t, and we have the following result.

Corollary 2.3. *Let $x \in \Lambda_M$ and $\tilde x$ denote a lift of x to $\widetilde M$. Let $T \in \mathbb{R}^+$. Define $\tilde y(t) := \tilde x(t/T)$. Then the following are equivalent*:

(1) *The pair (x, T) is a critical point of S_k.*

(2) *$\tilde y$ is a critical point of A_k.*

Thus the pair $(x, T) \in \Lambda_M \times \mathbb{R}^+$ is a critical point of S_k if and only if $t \mapsto x(t/T)$ is the projection to M of a closed orbit of ϕ_t.

To specify the lifts we work with, let us fix a lift $\tilde x_\nu : I \to \widetilde M$ of x_ν for each $\nu \in [\mathbb{T}, M]$. Throughout the paper, given any two paths y and y' such that the end point of y is the start point of y', the path $y * y'$ is the path obtained by first going along y and then going along y'. Similarly the path y^{-1} is the path obtained by going along y backwards.

Suppose now that $c(g, \sigma) < \infty$. Fix a free homotopy class $\nu \in [\mathbb{T}, M]$ (which could be the trivial free homotopy class). Let $x \in \Lambda_\nu$, and let x_s denote a free homotopy from $x_0 = x$ to $x_1 = x_\nu$. Let $z(s) := x_s(0)$. Let $\tilde x_s$ denote the unique homotopy of curves on $\widetilde M$ that projects down onto x_s and satisfies $\tilde x_1(t) = \tilde x_\nu(t)$. Let $\tilde x(t) := \tilde x_0(t)$, $\tilde z_0(s) := \tilde x_s(0)$ and $\tilde z_1(s) := \tilde x_s(1)$.

Now observe that if $R \subseteq \widetilde M$ denotes the rectangle $R = \operatorname{im} \tilde x_s$, then we have

$$\int_{C(x)} \sigma = \int_R \tilde\sigma = \int_R d\theta = \int_{\partial R} \theta = \int_{\tilde x * \tilde z_1 * \tilde x_\nu^{-1} * \tilde z_0^{-1}} \theta.$$

Let $\varphi \in \pi_1(M)$ denote the unique covering transformation taking $\tilde z_0$ to $\tilde z_1$. Since $\langle\varphi\rangle \le \pi_1(M)$ is an amenable subgroup, [Paternain 2006, Lemma 5.3] allows us to assume without loss of generality that θ is φ-invariant. Thus $\int_{\tilde z_0^{-1}} \theta + \int_{\tilde z_1} \theta = 0$. It thus follows that $\int_{C(x)} \sigma = \int_{\tilde x} \theta + \int_{\tilde x_\nu^{-1}} \theta$.

Set $a_\nu := \int_{\tilde x_\nu^{-1}} \theta$. We conclude that $\int_{C(x)} \sigma = \int_{\tilde x} \theta + a_\nu$. This computation shows if $c(g, \sigma) < \infty$, then for any $(x, T) \in \Lambda_\nu \times \mathbb{R}^+$, if $\tilde x$ is any lift of x and

$\tilde{y}(t) := \tilde{x}(t/T)$, then

$$S_k(x, T) = A_k(\tilde{y}) + a_\nu. \tag{2-7}$$

For the case $\nu = 0 \in [\mathbb{T}, M]$ the trivial free homotopy class, we may choose the curve x_0 above to be a constant map, from which it is easy to see that $a_0 = 0$. In particular, if $(x, T) \in \Lambda_0 \times \mathbb{R}^+$ and $\tilde{y}$ is defined as before, then

$$S_k(x, T) = A_k(\tilde{y}). \tag{2-8}$$

Finally, if $c(g, \sigma) = \infty$, S_k is only defined on $\Lambda_0 \times \mathbb{R}^+$, and it is clear that (2-8) still holds.

3. The Palais–Smale condition

Let $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ be a Riemannian Hilbert manifold, and suppose $S : \mathcal{M} \to \mathbb{R}$ is C^1.

Definition 3.1. We say S satisfies the *Palais–Smale condition* if every sequence $(x_n) \subseteq \mathcal{M}$ such that $\|d_{x_n} S\| \to 0$ as $n \to \infty$ and $\sup_n |S(x_n)| < \infty$ admits a convergent subsequence. We say S satisfies the *Palais–Smale condition at the level* $\mu \in \mathbb{R}$ if every sequence $(x_n) \subseteq \mathcal{M}$ with $\|d_{x_n} S\| \to 0$ as $n \to \infty$ and $S(x_n) \to \mu$ admits a convergent subsequence.

The following result, concerning S_k satisfying the Palais–Smale condition, is adapted from [Contreras 2006, Propositions 3.8 and 3.12]. We will first consider only the case where $c(g, \sigma) < \infty$; see Proposition 3.7 for the case $c(g, \sigma) = \infty$. In the statement of the theorem, $\| \cdot \|$ denotes the operator norm with respect to the metric $\langle \cdot, \cdot \rangle$.

Theorem 3.2. *Suppose $c(g, \sigma) < \infty$. Let $A, B, k \in \mathbb{R}^+$, and suppose $(x_n, T_n) \subseteq \Lambda_M \times \mathbb{R}^+$ satisfies*

$$\sup_n |S_k(x_n, T_n)| \le A, \quad \sup_n T_n \le B, \quad \|d_{(x_n, T_n)} S_k\| < 1/n.$$

(1) *If* $\liminf T_n > 0$, *then, passing to a subsequence if necessary, the sequence (x_n, T_n) is convergent in the $W^{1,2}$-topology.*

(2) *If* $\liminf T_n = 0$ *and the x_n are all contractible, then passing to a subsequence if necessary, $S_k(x_n, T_n) \to 0$.*

Before proving the theorem, let us now fix some notation that we will use throughout this section, as well as implicitly in the rest of the paper. Given a sequence $(x_n, T_n) \subseteq \Lambda_M \times \mathbb{R}^+$, let $y_n : [0, T_n] \to M$ be defined by $y_n(t) := x_n(t/T_n)$. Define

$$l_n := \int_0^1 |\dot{x}_n(t)| dt \quad \text{and} \quad e_n := \int_0^1 \frac{1}{2T_n} |\dot{x}_n(t)|^2 dt.$$

Note that l_n is the length of y_n and e_n is the energy of y_n. The Cauchy–Schwarz inequality implies

$$l_n^2 \leq 2T_n e_n. \tag{3-1}$$

Suppose now $c(g,\sigma) < \infty$. Since $\|\theta\|_\infty < \infty$, there exist constants $b_1, b_2 \in \mathbb{R}^+$ such that

$$L(q,v) \geq b_1|v|^2 - b_2 \quad \text{for all } (q,v) \in T\widetilde{M}. \tag{3-2}$$

Given $A, B, k \in \mathbb{R}^+$ and a free homotopy class $\nu \in [\mathbb{T}, M]$, we denote by $\mathbb{D}(A,B,k,\nu) \subseteq \Lambda_M \times \mathbb{R}^+$ set of pairs (x,T) such that $x \in \Lambda_\nu$, $S_k(x,T) \leq A$ and $T \leq B$.

Proof of Theorem 3.2. We begin with three preparatory lemmas.

Lemma 3.3. *Suppose $c(g,\sigma) < \infty$. Let $(x_n, T_n) \subseteq \mathbb{D}(A,B,k,\nu)$. Then if*

$$b(A,B,\nu) := \frac{A + b_2 B + |a_\nu|}{2b_1}$$

then $e_n \leq b(A,B,\nu)$ for all $n \in \mathbb{N}$.

Proof. We have by (2-7) and (3-2) that

$$A \geq S_k(x_n,T_n) = A_k(\tilde{y}_n) - a_\nu \geq 2b_1 e_n - b_2 T_n + kT_n + a_\nu,$$

and thus

$$e_n \leq \frac{A + b_2 T_n - kT_n + |a_\nu|}{2b_1} \leq \frac{A + b_2 B + |a_\nu|}{2b_1}. \qquad \square$$

Lemma 3.4. *Suppose $c(g,\sigma) < \infty$, and suppose $(x_n) \subseteq \Lambda_0$ are such that $l_n \to 0$. Then $\int_{C(x_n)} \sigma \to 0$.*

Proof. Let $\boldsymbol{x}_n : D^2 \to M$ denote a capping disc for x_n, so (as in (2-4))

$$\boldsymbol{x}_n|_{\partial D^2} = x_n \quad \text{and} \quad \int_{C(x_n)} \sigma = \int_{D^2} \boldsymbol{x}_n^* \sigma.$$

Let $\tilde{\boldsymbol{x}}_n : D^2 \to \widetilde{M}$ denote a lift of $\boldsymbol{x}_n$ to $\widetilde{M}$. Then

$$\left|\int_{D^2} \boldsymbol{x}_n^* \sigma\right| = \left|\int_{D^2} \tilde{\boldsymbol{x}}_n^*(d\theta)\right| = \left|\int_{\tilde{x}_n} \theta\right| \leq \|\theta\|_\infty l_n \to 0. \qquad \square$$

We now reduce Theorem 3.2(1) to a simpler situation:

Lemma 3.5. *Suppose $c(g,\sigma) < \infty$ and $(x_n,T_n) \in \mathbb{D}(A,B,k,\nu)$ with $\liminf T_n > 0$. Passing to a subsequence we may assume that there exists $x \in \Lambda_\nu$ such that the x_n converge to x in the C^0-topology.*

Proof. First by compactness of M, passing to a subsequence if necessary we may assume there exists $q \in M$ and $T \in \mathbb{R}^+$ such that $\lim_{n\to\infty} x_n(0) = x_n(1) = q$ and $\lim_{n\to\infty} T_n = T$. Consider g-geodesics $c_n : I \to M$ such that $c_n(0) = q$ and $c_n(1) = x_n(0)$. By passing to a subsequence we may assume that $\operatorname{dist}_g(x_n(0), q) < 1$, and thus we have $|\dot{c}_n| \le 1$. Now consider the curves

$$w_n : [0, T_n+2] \to M, \quad t \mapsto c_n * y_n * c_n^{-1} \quad \text{and} \quad z_n : \mathbb{T} \to M, \quad t \mapsto w_n(t/T_n+2).$$

Thus $z_n(0) = z_n(1) = q$, and $(z_n) \subseteq \Lambda_\nu$.

Given $0 \le t_1 < t_2 < T_n + 2$,

$$\operatorname{dist}_g(w_n(t_1), w_n(t_2)) \le \int_{t_1}^{t_2} |\dot{w}_n(t)|\,dt \le \sqrt{2}|t_2 - t_1|^{1/2}\Big(\int_0^{T_n+2} \tfrac{1}{2}|\dot{w}_n(t)|^2 dt\Big)^{1/2}.$$

By Lemma 3.3 we have

$$\int_0^{T_n+2} \frac{1}{2}|\dot{w}_n(t)|^2 dt = \int_0^1 \tfrac{1}{2}|\dot{c}_n(t)|^2 dt + e_n + \int_0^1 \tfrac{1}{2}|\dot{c}_n^{-1}(t)|^2 dt \le 1 + b(A, B, \nu),$$

and thus $\operatorname{dist}_g(w_n(t_1), w_n(t_2)) \le \sqrt{2}|t_2 - t_1|^{1/2}(1 + b(A, B, \nu))^{1/2}$. Hence the family (w_n) is equicontinuous. The Arzelà–Ascoli theorem then completes the proof. □

Proof of Theorem 3.2. We begin by proving Theorem 3.2(1). This part of the proof is very similar to the proof of [Contreras et al. 2000, Theorem B]. Suppose $(x_n, T_n) \subseteq \mathbb{D}(A, B, k, \nu)$ with $\liminf T_n > 0$. By the previous lemma, after passing to a subsequence if necessary, we may assume that (x_n, T_n) converges in the C^0 topology to some (x, T), where $T > 0$.

Without loss of generality, let us assume that the limit curve x is contained in a single chart U (otherwise simply repeat these arguments finitely many times). Then after passing possibly to another subsequence, we may assume that the x_n are all contained in U as well. There exists a constant $b_3 \in \mathbb{R}^+$ such that in the coordinates on U,

$$b_3 := \sup_{q \in U, v \in T_q M} \frac{|E_q(q, v)|}{1 + |v|^2} < \infty. \tag{3-3}$$

Write $z_n(t) := T_n^{-1} x_n(t)$. By Lemma 3.3 we can find a constant $R > 0$ such that

$$|x_n|_{1,2} \le R \quad \text{and} \quad |z_n|_{1,2} \le R.$$

Now since $\|d_{(x_n,T_n)} S_k\| \to 0$ as $n \to \infty$, given $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for every (ξ, ψ) satisfying $|(\xi, \psi)| \le 2R$ and $n, m \ge N$, we have

$$|d_{(x_n,T_n)} S_k(\xi, \psi) - d_{(x_m,T_m)} S_k(\xi, \psi)| < \varepsilon.$$

Take $\xi = x_n - x_m$ and $\psi = 0$ and use (2-5) to discover that

$$\begin{aligned} \text{(3-4)} \quad \Bigg| \int_0^1 \big(T_n \cdot E_q(x_n, \dot{z}_n) - T_m \cdot E_q(x_m, \dot{z}_m)\big)(x_n - x_m)dt \\ + \int_0^1 (E_v(x_n, \dot{z}_n) - E_v(x_m, \dot{z}_m))\,(\dot{x}_n - \dot{x}_m)dt \\ + \int_0^1 \sigma_{x_n}(\dot{x}_n, \dot{x}_m) - \sigma_{x_m}(\dot{x}_n, \dot{x}_m)dt \Bigg| < \varepsilon. \end{aligned}$$

Here we are using the canonical parallel transport available to us on Euclidean spaces to view $\dot{x}_n - \dot{x}_m$ as a tangent vector in any tangent space of our choosing. Using (3-3) we can bound the first integral as follows:

$$\begin{aligned} \Bigg| \int_0^1 \big(T_n \cdot E_q(x_n, \dot{z}_n) - T_m \cdot E_q(x_m, \dot{z}_m)\big)(x_n - x_m)dt \Bigg| \\ \leq (2Bb_3 + 2b_3 b(A, B, \nu)) \|x_n - x_m\|_\infty. \end{aligned}$$

Let us write $\sigma|_U$ in local coordinates as $\sigma = \sigma_{ij} dq^i \wedge dq^j$, where $\sigma_{ij} \in C^\infty(U, \mathbb{R})$. Then since

$$|\sigma_{ij}(x_n(t)) - \sigma_{ij}(x_m(t))| \to 0 \quad \text{as } n, m \to \infty, \text{ uniformly in } t,$$

and

$$\int_0^1 |\dot{x}_n||\dot{x}_m| dt \leq 2\sqrt{T_n T_m e_n e_m}$$

is bounded, it follows that for n and m large the third integral is small. Thus the second integral must also be small for large n and m. Since

$$|v - v'|^2 = (E_v(q, v) - E_v(q', v')) \cdot (v - v'),$$

we have

$$\int_0^1 |\dot{z}_n - \dot{z}_m|^2 dt \leq \int_0^1 (E_v(x_n, \dot{z}_n) - E_v(x_m, \dot{z}_m))(\dot{z}_n - \dot{z}_m)dt,$$

and hence the fact that the second integral in (3-4) is small for large n and m implies that the sequence (z_n), and hence the sequence (x_n), converges in the $W^{1,2}$ topology. This completes the proof of Theorem 3.2(1).

We now prove Theorem 3.2(2). This part of the proof follows the proof of [Contreras 2006, Theorem 3.8] very closely. Assume $(x_n, T_n) \subseteq \mathbb{D}(A, B, k, 0)$ (where $0 \in [\mathbb{T}, M]$ denotes the trivial free homotopy class) and that $\liminf T_n = 0$. Passing to a subsequence, we may assume that $T_n \to 0$. It suffices to show that passing to a subsequence we have $e_n \to 0$. Then by Lemma 3.4, $S_k(x_n, T_n) = e_n + kT_n - \int_{C(x_n)} \sigma \to 0$.

We know that e_n remains bounded by Lemma 3.3. Since $T_n \to 0$, (3-1) implies that $l_n \to 0$. Thus as before we may assume that all the curves x_n take their image in the domain of some chart U on M. Thus for the remainder of the proof we work in coordinates as if $M = \mathbb{R}^d$. Let $\xi_n(t) := x_n(t) - x_n(0)$, so that $\xi_n(0) = \xi_n(1) = 0$. Then $(\xi_n, 0) \in T_{(x_n,T_n)}(\Lambda_{\mathbb{R}^d} \times \mathbb{R}^+)$. Let also $\zeta_n(t) := \xi_n(t/T_n)$, so that $\dot{\zeta}_n(t) = \dot{y}_n(t)$. Then

$$|d_{(x_n,T_n)} S_k(\xi_n, 0)| \le \frac{1}{n}\Big(T_n \int_0^{T_n} |\dot{\zeta}_n(t)|^2 dt\Big)^{1/2} \le \frac{1}{n}\sqrt{2T_n e_n}.$$

Using (2-5) we have

$$d_{(x_n,T_n)} S_k(\xi_n, 0) = \int_0^{T_n} \big(E_q(y_n, \dot{y}_n) \cdot \zeta_n + E_v(y_n, \dot{y}_n) \cdot \dot{\zeta}_n(t)\big) dt \\ + \int_0^1 \sigma_{x_n(t)}(\xi_n(t), \dot{x}_n(t)) dt.$$

There exists $b_4 \in \mathbb{R}^+$ such that

$$\left| \int_0^1 \sigma_{x_n(t)}(\xi_n(t), \dot{x}_n(t)) dt \right| \le b_4 \int_0^1 |\xi_n(t)| |\dot{x}_n(t)| dt \le b_4 l_n^2.$$

Thus using (3-3) and the fact that $E_v(q, v) \cdot \xi = \langle v, \xi \rangle$, we have

$$\begin{aligned} d_{(x_n,T_n)} S_k(\xi_n, 0) &\ge -b_3 \int_0^{T_n} (1 + |\dot{y}_n(t)|^2) |y_n(t) - y_n(0)| dt + 2e_n - b_4 l_n^2 \\ &\ge -b_3 l_n (T_n + 2e_n) + 2e_n - b_4 l_n^2. \end{aligned}$$

Putting this together and dividing through by $\sqrt{T_n}$, we have

$$-b_3 l_n \sqrt{T_n} - 2b_3 \frac{e_n l_n}{\sqrt{T_n}} + 2\frac{e_n}{\sqrt{T_n}} - b_4 \frac{l_n^2}{\sqrt{T_n}} \le \frac{1}{n}\sqrt{2e_n}.$$

By (3-1), we have

$$\lim_{n\to\infty} \frac{l_n^2}{\sqrt{T_n}} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{l_n}{\sqrt{T_n}} \text{ is bounded;}$$

thus $\lim_{n\to\infty} e_n/\sqrt{T_n}$ must also be bounded, and this can happen if and only if $e_n \to 0$. This completes the proof of Theorem 3.2(2). □

We now wish to study the case where $c(g, \sigma) = \infty$. Recall in this case S_k is only defined on $\Lambda_0 \times \mathbb{R}^+$. For a result similar to Theorem 3.2 to hold in the unbounded setting, we must restrict to a subset of $\Lambda_0 \times \mathbb{R}^+$.

Definition 3.6. Suppose $K \subseteq \widetilde{M}$ is compact. Define $\Lambda_0^K \subseteq \Lambda_0$ to be the set of loops $x \in \Lambda_0$ such that there exists a lift $\tilde{x} : \mathbb{T} \to \widetilde{M}$ of x such that $\tilde{x}(\mathbb{T}) \subseteq K$.

Here is the extension of Theorem 3.2 to the case $c(g, \sigma) = \infty$.

Proposition 3.7. *Suppose that $c(g,\sigma)=\infty$. Let $A, B, k \in \mathbb{R}^+$ and take $K \subseteq \widetilde{M}$ to be compact. Suppose $(x_n, T_n) \subseteq \Lambda_0^K \times \mathbb{R}^+$ satisfy*

$$\sup_n |S_k(x_n, T_n)| \le A, \quad \sup_n T_n \le B, \quad \|d_{(x_n,T_n)} S_k\| < 1/n.$$

(1) *If* $\liminf T_n > 0$, *then passing to a subsequence if necessary the sequence* (x_n, T_n) *is convergent in the* $W^{1,2}$*-topology.*

(2) *If* $\liminf T_n = 0$, *then passing to a subsequence if necessary it holds that* $S_k(x_n, T_n) \to 0$.

Proof. The proof proceeds exactly as before, since any primitive θ of $\tilde{\sigma}$ is bounded on K. □

4. Supercritical energy levels: The case $k > c(g,\sigma)$

We now assume $c(g,\sigma) < \infty$, and study supercritical energies $k > c(g,\sigma)$. We aim to prove Theorem 1.1(1). The key fact we will use is the following result. As before, let $(\mathcal{M}, \langle \cdot, \cdot \rangle)$ be a Riemannian Hilbert manifold, and let $S : \mathcal{M} \to \mathbb{R}$ be of class C^1.

Proposition 4.1. *Suppose S is bounded from below and satisfies the Palais–Smale condition, and suppose for every $A \in \mathbb{R}^+$ the set $\{x \in \mathcal{M} : S(x) \le A\}$ is complete. Then S has a global minimum.*

A proof may be found in [Contreras et al. 2000, Corollary 23]. Fix a nontrivial free homotopy class $\nu \in [\mathbb{T}, M]$. The aim of this section is to verify for $k > c(g,\sigma)$ that the functional S_k on the Hilbert manifold $\Lambda_\nu \times \mathbb{R}^+$ satisfies the hypotheses of Proposition 4.1, for then the global minimum whose existence Proposition 4.1 guarantees is our desired closed orbit of energy k.

The first step then is the following lemma, whose proof only requires $k \ge c(g,\sigma)$, and works for any free homotopy class $\nu \in [\mathbb{T}, M]$.

Lemma 4.2. *Let $k \ge c(g,\sigma)$. Then $S_k|_{\Lambda_\nu \times \mathbb{R}^+}$ is bounded below.*

Proof. The argument begins by replicating an argument seen earlier in Section 2. Fix a free homotopy class $\nu \in [\mathbb{T}, M]$ (which could be the trivial free homotopy class). Let $(x, T) \in \Lambda_\nu \times \mathbb{R}^+$, and let x_s denote a free homotopy from $x_0 = x$ to $x_1 = x_\nu$. Let $z(s) := x_s(0)$. Lift x_s to a homotopy $\tilde{x}_s$ in $\widetilde{M}$ with $\tilde{x}_1(t) = \tilde{x}_\nu(t)$, and let $\tilde{x}(t) := \tilde{x}_0(t)$, $\tilde{z}_0(s) = \tilde{x}_s(0)$ and $\tilde{z}_1(s) = \tilde{x}_s(1)$.

Now observe that if $R \subseteq \widetilde{M}$ denotes the rectangle $R = \operatorname{im} \tilde{x}_s$, then we have

$$\int_{C(x)} \sigma = \int_R \tilde{\sigma} = \int_R d\theta = \int_{\partial R} \theta = \int_{\tilde{x} * \tilde{z}_1 * \tilde{x}_\nu^{-1} * \tilde{z}_0^{-1}} \theta.$$

Suppose $\varphi \in \pi_1(M)$ denotes the unique covering transformation taking $\tilde{z}_0$ to $\tilde{z}_1$. Since $\langle \varphi \rangle \le \pi_1(M)$ is an amenable subgroup, [Paternain 2006, Lemma 5.3] allows us to assume that without loss of generality, θ is φ-invariant. Thus $\int_{\tilde z_0^{-1}}\theta+\int_{\tilde z_1}\theta=0$. It thus follows that

$$\int_{C(x)}\sigma=\int_{\tilde x}\theta+\int_{\tilde x_\nu^{-1}}\theta. \tag{4-1}$$

Let $\tilde x_n:=\varphi^n\tilde x$, and use similar notations for $\tilde z_n$ and $\tilde x_{\nu,n}$. Let $\tilde y_n:=\tilde x_n(t/T)$, so $\tilde y_n:[0,T]\to\widetilde M$. Then for any $n\in\mathbb{N}$ we consider the closed loop $u_n:[0,T_n]\to\widetilde M$ defined by

$$u_n=\tilde y_0*\tilde y_1*\cdots*\tilde y_n*\tilde z_{n+1}*\tilde x_{\nu,n}^{-1}*\cdots*\tilde x_{\nu,1}^{-1}*\tilde x_\nu^{-1}*\tilde z_0^{-1},$$

where $T_n:=(n+1)T+1+(n+1)+1$. We have

$$\begin{aligned}A_k(u_n)=(n+1)\Big(\int_0^T\tfrac12|\dot{\tilde y}(t)|^2dt+\int_0^1\tfrac12|\dot{\tilde x}_\nu^{-1}|^2dt-\int_{\tilde y_0}\theta-\int_{\tilde x_\nu^{-1}}\theta\Big)&\\+\int_0^1\tfrac12|\dot{\tilde z}_1(t)|^2dt+\int_0^1\tfrac12|\dot{\tilde z}_0^{-1}(t)|^2dt+kT_n.&\end{aligned}$$

Now if $k\ge c(g,\sigma)$, then by definition of $c(g,\sigma)$ we have $A_k(u_n)\ge0$. Thus,

$$\begin{aligned}0\le\int_0^T\tfrac12|\dot{\tilde y}_0(t)|^2dt+\int_0^1\tfrac12|\dot{\tilde x}_\nu^{-1}|^2dt-\int_{\tilde y_0}\theta-\int_{\tilde x_\nu^{-1}}\theta+\frac{kT_n}{n+1}&\\+\frac{1}{n+1}\Big(\int_0^1|\dot{\tilde z}_1(t)|^2dt+\int_0^1|\dot{\tilde z}_0^{-1}(t)|^2dt\Big).&\end{aligned}$$

Letting $n\to\infty$ and substituting for the terms with $\tilde y_0$ we obtain

$$\int_0^1\frac{1}{2T}|\dot x(t)|^2dt+\int_0^1\tfrac12|\dot{\tilde x}_\nu^{-1}|^2dt-\int_{\tilde x}\theta-a_\nu+k(T+1)\ge0. \tag{4-2}$$

Now

$$\begin{aligned}S_k(x,T)&=\int_0^1\frac{1}{2T}|\dot x(t)|^2dt+kT-\int_{C(x)}\sigma\\&=\int_0^1\frac{1}{2T}|\dot x(t)|^2dt+kT-\int_{\tilde x}\theta-a_\nu,\end{aligned}$$

and hence by (4-1) and (4-2),

$$S_k(x,T)+\int_0^1\tfrac12|\dot{\tilde x}_\nu^{-1}(t)|^2dt+k\ge0,$$

that is,

$$S_k(x,T)\ge-\int_0^1\tfrac12|\dot{\tilde x}_\nu(t)^{-1}|^2dt-k>-\infty. \qquad\square$$

Let us set$i_{k,\nu} := \inf_{(x,T)\in\Lambda_\nu\times\mathbb{R}^+} S_k(x,T)$, so that the lemma tells us $i_{k,\nu} > -\infty$ for $k \geq c(g,\sigma)$.

The next lemma implies that $\{S_k|_{\Lambda_\nu\times\mathbb{R}^+} \leq A\}$ is complete for any $A \geq 0$.

Lemma 4.3. *Suppose $c(g,\sigma) < \infty$. Let $\nu \in [\mathbb{T}, M]$ be a nontrivial free homotopy class and let $A \in \mathbb{R}^+$. There exists $T_0 = T_0(A,k,\nu) \in \mathbb{R}^+$ such that $T \geq T_0$ if $(x,T) \in \mathbb{D}(A,\infty,k,\nu)$.*

Proof. Let $\tilde{x}$ denote a lift of x and let $\tilde{y} : [0,T] \to \widetilde{M}$ be the curve $t \mapsto \tilde{x}(t/T)$. Using (2-7) and (3-2), we compute that

$$A \geq S_k(x,T) = A_k(\tilde{y}) + a_\nu \geq \frac{b_1}{T}\int_0^1 |\dot{\tilde{x}}|^2 dt - (k-b_2)T + a_\nu \geq \frac{b_1}{T} l(\nu) - (k-b_2)T + a_\nu,$$

where $l(\nu) := \inf\{\int_0^1 |\dot{x}(t)|dt : x \in \Lambda_\nu\}$. Since M is closed and ν is a nontrivial free homotopy class, we have $l(\nu) > 0$, which implies the lemma. □

Proof of Theorem 1.1(1). Take $k > c(g,\sigma)$, and fix a nontrivial free homotopy class $\nu \in [\mathbb{T}, M]$. Let $(x_n, T_n) \subseteq \mathbb{D}(A,\infty,k,\nu)$. We want to show that (x_n,T_n) admits a convergent subsequence in the $W^{1,2}$-topology. In view of Theorem 3.2, it suffices to show that there exists $B > 0$ such that $(x_n,T_n) \subseteq \mathbb{D}(A,B,k,\nu)$ and that $\liminf T_n > 0$.

Lemma 4.4. *The sequence (T_n) is bounded above and bounded away from zero.*

Proof. First we claim that (T_n) is bounded. Indeed, if $c = c(g,\sigma)$,

$$A \geq S_k(x_n,T_n) = S_c(x_n,T_n) + (k-c)T_n \geq i_{c,\nu} + (k-c)T_n,$$

and thus (T_n) is bounded. Say $T_n \leq B$ for all n, where $B \in \mathbb{R}^+$. Passing to a subsequence we may assume that if $T := \liminf T_n$, then $T_n \to T$. It remains to check $T > 0$. From (3-1) and Lemma 3.3 if $T = 0$, then $l_n \to 0$. This is a contradiction since $l_n > l(\nu) > 0$ (see the proof of the previous lemma). □

5. Subcritical energy levels: The case $k < c(g,\sigma)$

In this section we drop the assumption that $c(g,\sigma) < \infty$, and study subcritical energies $k < c(g,\sigma)$.

Mountain pass geometry. Again let $(\mathcal{M}, \langle\cdot,\cdot\rangle)$ be a Riemannian Hilbert manifold and $S : \mathcal{M} \to \mathbb{R}$ a function of class C^2. Let Φ_s denote the (local) flow of $-\nabla S$. Define $\alpha : \mathcal{M} \to \mathbb{R}^+ \cup \{\infty\}$ by

$$\alpha(x) := \sup\{r > 0 : s \mapsto \Phi_s(x) \text{ is defined on } [0,r]\}.$$

An *admissible time* is a differentiable function $\tau : \mathcal{M} \to \mathbb{R}$ such that $0 \leq \tau(x) < \alpha(x)$ for all $x \in \mathcal{M}$.

Let $\mathcal{F}$ denote a family of subsets of $\mathcal{M}$, and define $\mu := \inf_{F\in\mathcal{F}} \sup_{x\in F} S(x)$. Suppose that $\mu \in \mathbb{R}$. We say that $\mathcal{F}$ is *S-forward invariant* if the following holds: if τ is an admissible time such that $\tau(x) = 0$ if $S(x) \le \mu - \delta$ for some $\delta > 0$, then for all $F \in \mathcal{F}$ the set $F_\tau := \{\Phi_{\tau(x)}(x) : x \in F\}$ is also a member of $\mathcal{F}$.

For convenience, given a subset $\mathcal{V} \subseteq \mathcal{M}$ and $a \in \mathbb{R}$, let $K_{a,\mathcal{V}} := \operatorname{crit} S \cap S^{-1}(a) \cap \mathcal{V}$ denote the set of critical points of S in $\mathcal{V}$ at the level a.

Our main tool will be the following mountain pass theorem, whose statement is similar to that of [Contreras 2006, Proposition 6.3]. In what follows, a *strict local minimizer* of a function $S : \mathcal{M} \to \mathbb{R}$ is a point $x \in \mathcal{M}$ such that there exists a neighborhood $\mathcal{N}$ of x such that $S(y) > S(x)$ for all $y \in \mathcal{N} \setminus \{x\}$.

Theorem 5.1. *Let $\mathcal{M}$ be a Riemannian Hilbert manifold and $S : \mathcal{M} \to \mathbb{R}$ a function of class C^2. Suppose we are given a sequence $(\mathcal{F}_n)$ of families of subsets of $\mathcal{M}$ with $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for all $n \in \mathbb{N}$. Set $\mathcal{F}_\infty := \bigcup_n \mathcal{F}_n$. Set $\mu_\infty := \inf_{F\in\mathcal{F}_\infty} \sup_{x\in F} S(x)$. Suppose in addition that*

1. *$\mathcal{F}_\infty$ is S-forward invariant, and the sets $F \in \mathcal{F}_\infty$ are connected*;
2. *$\mu_\infty \in \mathbb{R}$*;
3. *the flow Φ_s of $-\nabla S$ is relatively complete on $\{\mu_\infty - \eta \le S \le \mu_\infty + \eta\}$ for some $\eta > 0$*;
4. *there are closed subsets $(\mathcal{U}_n)$ of $\mathcal{M}$ such that for all $\varepsilon > 0$, there exists $n(\varepsilon) \in \mathbb{N}$ such that for all $n \ge n(\varepsilon)$ there exists $F \in \mathcal{F}_n$ and $0 < \varepsilon_1(n) < \varepsilon$ such that*
$$F \subseteq \{S \le \mu_\infty - \varepsilon_1(n)\} \cup (\mathcal{U}_n \cap \{S \le \mu_\infty + \varepsilon\}); \quad \textit{and}$$
5. *there are closed subsets $(\mathcal{V}_n)$ and a sequence $(r_n) \subseteq \mathbb{R}^+$ such that*
$$\mathcal{B}_{r_n}(\mathcal{U}_n) := \{x \in \mathcal{M} : \operatorname{dist}(x, \mathcal{U}_n) < r_n\} \subseteq \mathcal{V}_n,$$
and such that $S|_{\mathcal{V}_n}$ satisfies the Palais–Smale condition at the level μ_∞.

Then if $\mathcal{V}_\infty := \bigcup_{n\in\mathbb{N}} \mathcal{V}_n$, then S has a critical point $x \in \mathcal{V}_\infty$ with $S(x) = \mu_\infty$, that is, $K_{\mu_\infty,\mathcal{V}_\infty} \ne \varnothing$. Moreover, if

$$\sup_{F\in\mathcal{F}_\infty} \inf_{x\in F} S(x) < \mu_\infty, \tag{5-1}$$

then there is a point in $K_{\mu_\infty,\mathcal{V}_\infty}$ that is not a strict local minimizer of S.

The proof is an easy application of the following result, which can be found as [Contreras 2006, Lemma 6.2].

Lemma 5.2. *Let $\mathcal{M}$ be a Riemannian Hilbert manifold and let $\mathcal{U} \subseteq \mathcal{V} \subseteq \mathcal{M}$ be closed subsets such that $\mathcal{B}_r(\mathcal{U}) \subseteq \mathcal{V}$ for some $r > 0$. Let $S : \mathcal{M} \to \mathbb{R}$ be a C^2 function, and let $\mu \in \mathbb{R}$ be such that $S|_{\mathcal{V}}$ satisfies the Palais–Smale condition at the level μ.*

Suppose also that the flow Φ_s *of* $-\nabla S$ *is relatively complete on* $\{|S-\mu|\leq\eta\}$ *for some* $\eta>0$.

Then if $\mathcal{N}$ *is any neighborhood of* $K_{\mu,\mathcal{V}}$ *relative to* $\mathcal{V}$, *then for any* $\lambda>0$ *there exists* ε *and* δ *with* $0<\varepsilon<\delta<\lambda$ *such that for any* $0<\varepsilon_1<\varepsilon$ *there exists an admissible time* τ *such that* $\tau(x)=0$ *for all* $x\in\{|S-\mu|\geq\delta\}$, *and such that if*

$$F:=\{S\leq\mu-\varepsilon_1\}\cup(\mathcal{U}\cap\{S\leq\mu+\varepsilon\}),$$

then $F_\tau\subseteq\mathcal{N}\cup\{S\leq\mu-\varepsilon_1\}$.

Proof of Theorem 5.1. We will show that $K_{\mu_\infty,\mathcal{V}_n}\neq\varnothing$ for n large enough. Fix $0<\varepsilon<\delta<\lambda:=1$ as in the statement of Lemma 5.2. By hypothesis there exists $n(\varepsilon)\in\mathbb{N}$ such that for all $n\geq n(\varepsilon)$ there exists $0<\varepsilon_1(n)<\varepsilon$ and $F\in\mathcal{F}_n$ such that

$$F\subseteq\{S\leq\mu_\infty-\varepsilon_1(n)\}\cup(\mathcal{U}_n\cap\{S\leq\mu_\infty+\varepsilon\}).$$

For such n, we have $K_{\mu_\infty,\mathcal{V}_n}\neq\varnothing$. Indeed, if $K_{\mu_\infty,\mathcal{V}_n}=\varnothing$, by Lemma 5.2, there exists an admissible time τ such that $\tau\equiv0$ on $\{S\leq\mu_\infty-\delta\}$, and such that F_τ satisfies $F_\tau\subseteq\{S\leq\mu_\infty-\varepsilon_1(n)\}$ (for we may take $\mathcal{N}=\varnothing$ in Lemma 5.2). Since $\mathcal{F}_\infty$ is forward invariant, $F_\tau\in\mathcal{F}_\infty$. This contradicts the definition of μ_∞.

To prove the last statement, suppose that $K_{\mu_\infty,\mathcal{V}_\infty}$ consists entirely of strict local minimizers, and that (5-1) holds. Choose $\lambda_0>0$ such that

$$\sup_{F\in\mathcal{F}_\infty}\inf_{x\in F}S(x)<\mu_\infty-2\lambda_0.$$

For each $x\in K_{\mu_\infty,\mathcal{V}_\infty}$, let $\mathcal{N}(x)$ denote a neighborhood of x such that $S(y)>S(x)$ for all $y\in\mathcal{N}(x)\setminus\{x\}$, and let

$$\mathcal{N}_0:=\bigcup_{x\in K_{\mu_\infty,\mathcal{V}_\infty}}\mathcal{N}(x)\quad\text{and}\quad\mathcal{N}_n:=\mathcal{N}_0\cap\mathcal{V}_n\text{ for each }n\in\mathbb{N}.$$

Let $0<\varepsilon<\delta<\lambda_0$ be given by Lemma 5.2 for $\mathcal{N}_0$. By hypothesis there exists $n(\varepsilon)\in\mathbb{N}$ such that for all $n\geq n(\varepsilon)$ there exists $0<\varepsilon_1(n)<\varepsilon$ and $F\in\mathcal{F}_n$ such that

$$F\subseteq\{S\leq\mu_\infty-\varepsilon_1(n)\}\cup(\mathcal{U}_n\cap\{S\leq\mu_\infty+\varepsilon\}).$$

By Lemma 5.2, there exists an admissible time τ such that $\tau\equiv0$ on $\{S\leq\mu_\infty-\delta\}$ and such that $F_\tau\subseteq\mathcal{N}_n\cup\{S\leq\mu_\infty-\varepsilon_1(n)\}\subseteq\mathcal{N}_0\cup\{S\leq\mu_\infty-\varepsilon_1(n)\}$. By definition of $\mathcal{N}_0$, the sets $\mathcal{N}_0$ and $\{S\leq\mu_\infty-\varepsilon_1(n)\}$ are disjoint, so $\mathcal{N}_0\cup\{S\leq\mu_\infty-\varepsilon_1(n)\}$ is disconnected. Since F_τ is connected by hypothesis, we either have $F_\tau\subseteq\mathcal{N}_0$ and $F_\tau\cap\{S\leq\mu_\infty-\varepsilon_1(n)\}=\varnothing$, or $F_\tau\subseteq\{S\leq\mu_\infty-\varepsilon_1(n)\}$. The former fails since $\varepsilon_1(n)<\varepsilon<\lambda_0$, and the value of S decreases under Φ_s, and the latter contradicts the definition of μ_∞. □

Proof of the second statement of Theorem 1.1(2). The main tool we will use will be Theorem 5.1. The first step however is the following result, whose statement and proof closely parallel [Contreras 2006, Proposition C].

Proposition 5.3. *Let $k \in \mathbb{R}^+$. Then there exists a constant $\mu_0 > 0$ such that if $f : I \to \Lambda_0 \times \mathbb{R}^+$ is any path such that, with $f(0) = (x_0, T_0)$ and $f(1) = (x_1, T_1)$, we have*

(1) *$S_k(x_0, T_0) < 0$, and*

(2) *x_1 is the constant curve $x_1(t) \equiv x_0(0)$,*

then $\sup_{s\in I} S_k(f(s)) > \mu_0 > 0$.

Remark. The constant μ_0 *does not* depend on T_1.

In the statement of the following, as before, we put $l(x) := \int_0^1 |\dot{x}(t)|dt$.

Lemma 5.4 [Contreras 2006, Lemma 5.1]. *Let $\theta \in \Omega^1(\widetilde{M})$. Given any $q \in \widetilde{M}$ and any open neighborhood $V \subseteq \widetilde{M}$ of q, there exists an open neighborhood $W \subseteq V$ of q and a constant $\beta > 0$ such that $|\int_x \theta| \le \beta l(x)^2$ for any closed curve $x : I \to W$.*

Proof of Proposition 5.3. Compactness of M and the previous lemma imply that there exists $\beta, \rho_0 > 0$ such that if $x : I \to M$ is any closed contractible curve with $x(I)$ contained in a ball of radius ρ_0 then $|\int_{C(x)} \sigma| \le \beta l(x)^2$. Let $q := x_0(0)$ and let W denote the ball of radius ρ_0 about q. Pick $\rho \in \mathbb{R}^+$ such that

$$0 < \rho < \min\{\rho_0, \sqrt{k/(2\beta)^2}\}$$

Write $f(s) = (x_s, T_s)$, so $x_s \in \Lambda_0$ for all s. We claim that there exists $s_0 \in (0, 1)$ such that $l(x_{s_0}) = \rho$. Since the functional $s \mapsto l(x_s)$ is continuous and $l(x_0) = 0$, it suffices to show that there exists $s_1 \in [0, 1)$ such that $l(x_{s_1}) > \rho$.

If there exists $s_1 \in [0, 1)$ such that $x_{s_1}(I) \subsetneq W$, then we are done, since then $l(x_{s_1}) \ge \rho_0 > \rho$. The other possibility is that $x_s(I) \subseteq W$ for all $s \in I$. In this case we claim that we may take $s_1 = 0$, that is, $l(x_0) > \rho$. By assumption if $y_0(t) = x_0(t/T_0)$, we have

$$\begin{aligned} 0 > S_k(x_0, T_0) &= \int_0^1 \frac{1}{2T_0}|\dot{x}_0(t)|^2 dt + kT_0 - \int_{C(x_0)} \sigma \\ &\ge \frac{1}{2T_0} l(x_0)^2 + kT_0 - \left|\int_{x_0} \theta\right| \\ &\ge \Big(\frac{1}{2T_0} - \beta\Big) l(x_0)^2 + kT_0, \end{aligned} \tag{5-2}$$

where the second inequality came from (3-1) and the third from Lemma 5.4. From this it follows that $T_0 > 1/(2\beta)$, and thus

$$l(x_0)^2 > \frac{kT_0}{\beta - 1/(2T_0)} > \frac{k}{2\beta^2} > \rho^2.$$

and we are done as before.

We claim finally that $S_k(f(s_0)) > 0$. Since $x_{s_0} \in C^{\mathrm{ac}}_M(q,q)$ and $l(x_{s_0}) < \rho_0$, we have $x_{s_0}(I) \subseteq W$. In particular, (5-2) holds for x_{s_0}, and so we have

$$S_k(f(s_0)) \geq \Big(\frac{1}{2T_{s_0}} - \beta\Big)\ell^2 + kT_{s_0} = P(T_{s_0}) \geq \min_{t\in\mathbb{R}^+} P(t),$$

where $P(t) := (1/(2t) - \beta)\rho^2 + kt$. It is elementary to see that

$$\min_{t\in\mathbb{R}^+} P(t) = \sqrt{\rho^2/(2k)} =: \mu_0 > 0,$$

and this completes the proof. □

The next lemma will be needed to prove relative completeness of the flow of $-\nabla S_k$ on any interval not containing zero.

Lemma 5.5. *There exists a constant* $C > 0$ *such that for any* $(x_0, T_0) \in \Lambda_M \times \mathbb{R}^+$ *and any* $r > 0$, *if* $(x_1, T_1) \in \Lambda_M \times \mathbb{R}^+$ *satisfies* $\mathrm{dist}((x_0, T_0), (x_1, T_1)) < r$, *then*

$$|T_0 - T_1| < r \quad \textit{and} \quad \mathrm{dist}_{\mathrm{HD}}(x_0, x_1) < Cr.$$

This result is essentially proved by Contreras [2006, Lemma 2.3]; Contreras used a different metric on $\Lambda_M \times \mathbb{R}^+$, which meant that an additional condition was imposed in the statement of the lemma. Since we are working with the standard metric (2-1) on $\Lambda_M \times \mathbb{R}^+$ this additional condition is not needed, and the proof in [Contreras 2006] goes through without any changes.

Corollary 5.6. *Let* $K \subseteq \widetilde{M}$ *and* $B > 0$. *Let* $\mathcal{U} := \{(x, T) \in \Lambda_0^K \times \mathbb{R}^+ : T \leq B\}$. *Let* C *be as in the statement of Lemma 5.5. Then if* $L \subseteq \widetilde{M}$ *satisfies*

$$\{q \in \widetilde{M} : \mathrm{dist}_{\tilde{g}}(q, q') \leq Cr \textit{ for some } q' \in K\} \subseteq L$$

and we set $\mathcal{V} := \{(x, T) \in \Lambda_0^L \times \mathbb{R}^+ : T \leq B + r\}$, *then* $\mathcal{B}_r(\mathcal{U}) \subseteq \mathcal{V}$.

Proof. Suppose $(x_1, T_1) \in \mathcal{B}_r(\mathcal{U})$. Then there exists $(x_0, T_0) \in \mathcal{U}$ with

$$\mathrm{dist}((x_0, T_0), (x_1, T_1)) < r.$$

By Lemma 5.5, $\mathrm{dist}_{\mathrm{HD}}(x_0, x_1) < Cr$ and $|T_0 - T_1| < r$. Thus $(x_1, T_1) \in \mathcal{V}$. □

Next, we prove relative completeness of the flow of $-\nabla S_k$ on any interval that doesn't contain zero. This proof is very similar to [Contreras 2006, Lemma 6.9].

Lemma 5.7. *For all* $k \in \mathbb{R}^+$, *if* $[a, b] \subseteq \mathbb{R}$ *is an interval such that* $0 \notin [a, b]$, *then the local flow of* $-\nabla S_k$ *is relatively complete on* $(\Lambda_0 \times \mathbb{R}^+) \cap \{a \leq S_k \leq b\}$.

Proof. Let $\Phi_s : \Lambda_M \times \mathbb{R}^+ \to \Lambda_M \times \mathbb{R}^+$ denote the local flow of the vector field $-\nabla S_k$. Then for any $(x, T) \in \Lambda_M \times \mathbb{R}^+$,

$$S_k(\Phi_{s_1}(x, T)) - S_k(\Phi_{s_2}(x, T)) = \int_{s_1}^{s_2} |\nabla S_k(\Phi_s(x, T))|^2 ds.$$

By the Cauchy–Schwarz inequality we see that

$$\begin{aligned}\operatorname{dist}(\Phi_{s_1}(x,T),\Phi_{s_2}(x,T))^2 &\le \Big(\int_{s_1}^{s_2}|\nabla S_k(\Phi_s(x,T))|ds\Big)^2\\ &\le |s_1-s_2|\int_{s_1}^{s_2}|\nabla S_k(\Phi_s(x,T))|^2ds,\end{aligned}$$

and hence

$$\operatorname{dist}(\Phi_{s_1}(x,T),\Phi_{s_2}(x,T))^2 \le |s_1-s_2|\,|S_k(\Phi_{s_1}(x,T))-S_k(\Phi_{s_2}(x,T))|. \tag{5-3}$$

Now suppose we are given a pair $(x,T)\in\Lambda_0\times\mathbb{R}^+$, such that there exists $a,b\in\mathbb{R}$ with $0\notin[a,b]$ and

$$a\le S_k(\Phi_s(x,T))\le b\quad \text{for all } s \text{ such that } \Phi_s(x,T) \text{ is defined.}$$

Let $[0,\alpha)$ be the maximum interval of definition of $s\mapsto\Phi_s(x,T)$, where $\alpha>0$. To complete the proof we need to show $\alpha=\infty$. Suppose the contrary.

Write $\Phi_s(x,T)=(x_s,T_s)$. If $s_n\uparrow\alpha$, then $(\Phi_{s_n}(x,T)) =: (x_n,T_n)$ is a Cauchy sequence in $(\Lambda_0\times\mathbb{R}^+)\cap\{a\le S_k\le b\}$ by (5-3). Thus $T_\alpha:=\lim_{s\uparrow\alpha}T_s$ exists and is finite.

If $T_\alpha>0$, then $(x_\alpha,T_\alpha):=\lim_{n\to\infty}(x_n,T_n)$ exists and is equal to $\Phi_\alpha(x,T)$ since the sequence (x_n,T_n) is Cauchy. Since S_k is C^2 we can extend the solution $s\mapsto\Phi_s(x,T)$ at $s=\alpha$, contradicting the definition of α. Thus $T_\alpha=0$. Hence there exists a sequence $s_m\uparrow\alpha$ such that $\frac{d}{ds}T_{s_m}\le 0$. As before write $x_m:=x_{s_m}$ and $T_m:=T_{s_m}$. By (5-3) and Lemma 5.5, we may assume there exists a compact set $K\subseteq\widetilde{M}$ such that $(x_m,T_m)\subseteq\Lambda_0^K\times\mathbb{R}^+$ for all m. If $y_m(t):=x_m(t/T_m)$, then

$$0\ge\frac{d}{ds}T_m=-\frac{\partial}{\partial T}S_k(x_m,T_m)=\frac{1}{T_m}\int_0^{T_m}(-k+E(y_m,\dot y_m))dt=-k+\frac{e_m}{T_m},$$

where the penultimate equality uses (2-6). Since $\lim_{m\to\infty}T_m=0$, this forces $\lim_{m\to\infty}e_m=0$. As in the proof of the second part of Theorem 3.2, this implies $S_k(x_m,T_m)\to 0$, contradicting the fact that $0\notin[a,b]$. This implies that we must have originally had $\alpha=\infty$, and so completes the proof. □

We now move towards proving Theorem 1.1(2). In fact, we will prove a stronger result, which is based on [Contreras 2006, Proposition 7.1]:

Proposition 5.8. *Let $c=c(g,\sigma)\in\mathbb{R}\cup\{\infty\}$. For almost all $k\in(0,c)$ there exists a contractible closed orbit of ϕ_t with energy k. This orbit has positive S_k-action, and is not a strict local minimizer of S_k on $\Lambda_0\times\mathbb{R}^+$. This holds for a specific $k\in(0,c)$ if S_k is known to satisfy the Palais–Smale condition on the level k.*

Proof. Fix $k_0\in(0,c)$. There exists $(x_0,T_0)\in\Lambda_0\times\mathbb{R}^+$ such that $S_{k_0}(x_0,T_0)<0$. Indeed, there exists a closed curve $\tilde y:[0,T_0]\to\widetilde{M}$ such that $A_{k_0}(\tilde y)<0$. Then the projection $y : [0, T_0] \to M$ of $\tilde{y}$ to M is a closed curve, and if $x_0(t) := y(tT_0)$, then $(x_0, T_0) \in \Lambda_0 \times \mathbb{R}^+$ and $S_{k_0}(x_0, T_0) = A_{k_0}(\tilde{y}) < 0$. There exists $\varepsilon > 0$ such that for all $k \in J := [k_0, k_0 + \varepsilon]$, we have $S_k(x_0, T_0) < 0$.

Let x_1 denote the constant loop at $x_0(0)$. Given $k \in J$, let $\mu_0(k) > 0$ be the constant given by Proposition 5.3 such that any path $f \in C^0(I, \Lambda_0 \times \mathbb{R}^+)$ with $f(0) = (x_0, T_0)$ and $f(1) = (x_1, T)$ for some $T > 0$ has $\sup_{s \in I} S_k(f(s)) > \mu_0(k)$. Choose $T_1 > 0$ such that $T_1 < \inf_{k \in J} \mu_0(k)/k$. Then

$$\max\{S_k(x_0, T_0), S_k(x_1, T_1)\} = kT_1 < \mu_0(k) \quad \text{for all } k \in J.$$

Set $\Gamma := \{f \in C^0(I, \Lambda_0 \times \mathbb{R}^+) : f(0) = (x_0, T_0), f(1) = (x_1, T_1)\}$. Let $(K_n) \subseteq \widetilde{M}$ denote compact sets such that $K_n \subseteq K_{n+1}$ and $\bigcup_n K_n = \widetilde{M}$. Let

$$\Gamma_n := \Gamma \cap C^0(I, \Lambda_0^{K_n} \times \mathbb{R}^+).$$

Define $\mu_n(k) := \inf_{f \in \Gamma_n} \sup_{s \in I} S_k(f(s))$ and $\mu_\infty(k) := \inf_{f \in \Gamma} \sup_{s \in I} S_k(f(s))$ for $k \in J$. Then $\mu_n(k) \geq \mu_{n+1}(k) \geq \mu_\infty(k) \geq \mu_0(k)$ for all $n \in \mathbb{N}$ and $k \in J$, and the functions $\mu_n : J \to \mathbb{R}$ converge pointwise to μ_∞. Both μ_n and μ_∞ are nondecreasing. Since μ_∞ is nondecreasing, by Lebesgue's theorem there exists a subset $J_0 \subseteq (k_0, k_0 + \varepsilon)$ with $J \setminus J_0$ having measure zero such that $\mu_\infty|_{J_0}$ is locally Lipschitz. In other words, for all $j \in J_0$ there exist constants $M(j) > 0$ and $\delta(j) > 0$ such that

$$|\mu_\infty(j + \delta) - \mu_\infty(j)| < M(j)|\delta| \quad \text{for all } |\delta| < \delta(j).$$

Fix $j \in J_0$ and a sequence $(j_m) \subseteq J_0$ with $j_m \downarrow j$. Let $f_{n,m} \in \Gamma_n$ be paths such that

$$\max_{s \in I} S_{j_m}(f_{n,m}(s)) \leq \mu_n(j_m) + (j_m - j).$$

Next, define $\mathcal{U}_n := \{(x, T) \in \Lambda_0^{K_n} \times \mathbb{R}^+ : T \leq M(j) + 2\}$. Choose another collection $(L_n) \subseteq \widetilde{M}$ of compact sets such that $K_n \subseteq L_n$, and such that $\mathcal{B}_1(\mathcal{U}_n) \subseteq \mathcal{V}_n$ for $\mathcal{V}_n := \{(x, T) \in \Lambda_0^{L_n} \times \mathbb{R}^+ : T \leq M(j) + 3\}$. Such a collection (L_n) exists by Corollary 5.6. Since $\mu_\infty(j) \neq 0$, from Proposition 3.7 it follows that $S_j|_{\mathcal{V}_n}$ satisfies the Palais–Smale condition at the level $\mu_\infty(j)$ for all $n \in \mathbb{N}$.

Since $k \mapsto S_k(x, T)$ is increasing,

$$\max_{s \in I} S_j(f_{n,m}(s)) \leq \max_{s \in I} S_{j_m}(f_{n,m}(s)) \leq \mu_n(j_m) + (j_m - j). \tag{5-4}$$

If $s \in I$ is such that $S_j(f_{n,m}(s)) > \mu_\infty(j) - (j_m - j)$, writing $f_{n,m}(s) =: (x_s^{n,m}, T_s^{n,m})$ we have

$$\begin{aligned} T_s^{n,m} &= \frac{S_{j_m}(f_{n,m}(s)) - S_j(f_{n,m}(s))}{j_m - j} \\ &\leq \frac{\mu_\infty(j_m) - \mu_n(j)}{j_m - j} + 2 \leq \frac{\mu_\infty(j_m) - \mu_\infty(j)}{j_m - j} + 2 \leq M(j) + 2, \end{aligned}$$

for n large enough.

Given $\varepsilon > 0$, first choose m large enough that $j_m - j < \varepsilon/(2(M(j)+1))$, and then select n large enough that $\mu_n(j_m) - \mu_\infty(j_m) < \varepsilon/2$. Then

$$\begin{aligned}\mu_n(j_m) - \mu_\infty(j) + (j_m - j)& \\ &= (\mu_n(j_m) - \mu_\infty(j_m)) + (\mu_\infty(j_m) - \mu_\infty(j)) + (j_m - j) \\ &< \varepsilon/2 + M(j)(j_m - j) + (j_m - j) < \varepsilon.\end{aligned}$$

Then by (5-4),

$$f_{n,m}(I) \subseteq \{S_j \le \mu_\infty(j) - (j_m - j)\} \cap (\mathcal{U}_n \cap \{S_j \le \mu_\infty(j) + \varepsilon\}).$$

Since $\mu_\infty(j) \neq 0$, by Lemma 5.7 the gradient flow of $-S_j$ is relatively complete on $\{\mu_\infty(j) - \eta \le S_j \le \mu_\infty(j) + \eta\}$ for some $\eta > 0$. Theorem 5.1 then gives a critical point for $S_j|_{\Lambda_0 \times \mathbb{R}^+}$ that is not a strict local minimizer (here we are applying Theorem 5.1 with $\mathcal{F}_n := \{f(I) : f \in \Gamma_n\}$).

Finally, suppose that $k < c(g, \sigma) \le \infty$ is such that S_k satisfies the Palais–Smale condition. Then the theorem is immediate from Lemma 5.7 and Theorem 5.1. Indeed, by Lemma 5.7 we may simply take $\mathcal{U}_n = \mathcal{V}_n = \mathcal{V}_\infty = \Lambda_0 \times \mathbb{R}^+$, as then the hypotheses of Theorem 5.1 are trivially satisfied. This completes the proof of Theorem 1.1. □

Acknowledgment

I thank my PhD adviser Gabriel P. Paternain for many helpful discussions.

References

[Abbondandolo and Schwarz 2009a] A. Abbondandolo and M. Schwarz, "Estimates and computations in Rabinowitz–Floer homology", *J. Topol. Anal.* **1**:4 (2009), 307–405. MR 2597650

[Abbondandolo and Schwarz 2009b] A. Abbondandolo and M. Schwarz, "A smooth pseudo-gradient for the Lagrangian action functional", *Adv. Nonlinear Stud.* **9**:4 (2009), 597–623. MR 2560122 Zbl 1185.37145

[Burns and Paternain 2002] K. Burns and G. P. Paternain, "Anosov magnetic flows, critical values and topological entropy", *Nonlinearity* **15**:2 (2002), 281–314. MR 2004d:37076 Zbl 1161.37337

[Cieliebak et al. 2010] K. Cieliebak, U. Frauenfelder, and A. Oancea, "Rabinowitz Floer homology and symplectic homology", *Annales scientifiques de l'ENS* **43** (2010), fascicule 6. arXiv 0903.0768

[Contreras 2006] G. Contreras, "The Palais–Smale condition on contact type energy levels for convex Lagrangian systems", *Calc. Var. Partial Differential Equations* **27**:3 (2006), 321–395. MR 2007i:37116 Zbl 1105.37037

[Contreras and Iturriaga 1999] G. Contreras and R. Iturriaga, *Global minimizers of autonomous Lagrangians*, Colóquio Brasileiro de Matemática **22**, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1999. MR 2001j:37113 Zbl 0957.37065

[Contreras et al. 1997] G. Contreras, J. Delgado, and R. Iturriaga, "Lagrangian flows: The dynamics of globally minimizing orbits, II", *Bol. Soc. Brasil. Mat.* (*N.S.*) **28**:2 (1997), 155–196. MR 98i:58093 Zbl 0892.58065

[Contreras et al. 2000] G. Contreras, R. Iturriaga, G. P. Paternain, and M. Paternain, "The Palais–Smale condition and Mañé's critical values", *Ann. Henri Poincaré* **1**:4 (2000), 655–684. MR 2001k:37101 Zbl 0986.58005

[Contreras et al. 2004] G. Contreras, L. Macarini, and G. P. Paternain, "Periodic orbits for exact magnetic flows on surfaces", *Int. Math. Res. Not.* **2004**:8 (2004), 361–387. MR 2005a:37103 Zbl 1086.37032

[Ginzburg 1996] V. L. Ginzburg, "On closed trajectories of a charge in a magnetic field: An application of symplectic geometry", pp. 131–148 in *Contact and symplectic geometry* (Cambridge, 1994), edited by C. B. Thomas, Publ. Newton Inst. **8**, Cambridge Univ. Press, 1996. MR 97j:58128 Zbl 0873.58034

[Ginzburg and Gürel 2009] V. L. Ginzburg and B. Z. Gürel, "Periodic orbits of twisted geodesic flows and the Weinstein–Moser theorem", *Comment. Math. Helv.* **84** (2009), 865–907. MR 2534483 Zbl 1184.37046

[Kerman 2000] E. Kerman, *Symplectic geometry and the motion of a particle in a magnetic field*, thesis, Univ. California, Santa Cruz, 2000, available at http://tinyurl.com/32h3bzu. MR 2616826

[Lu 2006] G. Lu, "Finiteness of the Hofer–Zehnder capacity of neighborhoods of symplectic submanifolds", *Int. Math. Res. Not.* **2006** (2006), Art. ID 76520. MR 2007a:53159 Zbl 1132.53317

[Mañé 1996] R. Mañé, "Lagrangian flows: The dynamics of globally minimizing orbits", pp. 120–131 in *International Conference on Dynamical Systems* (Montevideo, 1995), edited by F. Ledrappier et al., Pitman Res. Notes Math. Ser. **362**, Longman, Harlow, 1996. MR 98g:58059 Zbl 0870.58026

[Macarini 2004] L. Macarini, "Hofer–Zehnder capacity and Hamiltonian circle actions", *Commun. Contemp. Math.* **6**:6 (2004), 913–945. MR 2005k:53170 Zbl 1076.53098

[Merry 2010] W. J. Merry, "On the Rabinowitz Floer homology of twisted cotangent bundles", preprint, 2010. arXiv 1002.0162

[Osuna 2005] O. Osuna, "Periodic orbits of weakly exact magnetic flows", preprint, 2005.

[Paternain 2006] G. P. Paternain, "Magnetic rigidity of horocycle flows", *Pacific J. Math.* **225**:2 (2006), 301–323. MR 2008e:37029 Zbl 1116.37020

[Polterovich 1998] L. Polterovich, "Geometry on the group of Hamiltonian diffeomorphisms", pp. 401–410 in *Proceedings of the International Congress of Mathematicians* (Berlin, 1998), vol. 2, Doc. Math. **1998**, 1998. MR 2000c:37120 Zbl 0909.58004

[Schlenk 2006] F. Schlenk, "Applications of Hofer's geometry to Hamiltonian dynamics", *Comment. Math. Helv.* **81**:1 (2006), 105–121. MR 2007f:53117 Zbl 1094.37031

[Usher 2009] M. Usher, "Floer homology in disk bundles and symplectically twisted geodesic flows", *J. Mod. Dyn.* **3**:1 (2009), 61–101. MR 2010b:53157 Zbl 1186.53099

Received June 28, 2009.

WILL J. MERRY
DEPARTMENT OF PURE MATHEMATICS AND MATHEMATICAL STATISTICS
UNIVERSITY OF CAMBRIDGE
CAMBRIDGE CB3 0WB
ENGLAND
w.merry@dpmms.cam.ac.uk

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

RINGEL–HALL ALGEBRAS AND TWO-PARAMETER QUANTIZED ENVELOPING ALGEBRAS

XIN TANG

Let $\mathfrak{g}$ be a finite-dimensional complex simple Lie algebra and Λ be the finite-dimensional hereditary algebra associated to $\mathfrak{g}$. Let $U_{r,s}^{+}(\mathfrak{g})$ (respectively $U_{r,s}^{\geq 0}(\mathfrak{g})$) denote the two-parameter quantized enveloping algebra of the positive maximal nilpotent (respectively Borel) Lie subalgebra of $\mathfrak{g}$. We study the two-parameter quantized enveloping algebras $U_{r,s}^{+}(\mathfrak{g})$ and $U_{r,s}^{\geq 0}(\mathfrak{g})$ using the approach of Ringel–Hall algebras. First of all, we show that $U_{r,s}^{+}(\mathfrak{g})$ is isomorphic to a certain two-parameter twisted Ringel–Hall algebra $H_{r,s}(\Lambda)$, which generalizes a result of Reineke. Based on detailed computations in $H_{r,s}(\Lambda)$, we show that $H_{r,s}(\Lambda)$ can be presented as an iterated skew polynomial ring. As an result, we obtain a PBW-basis for $H_{r,s}(\Lambda)$, which can be further used to construct a PBW-basis for the two-parameter quantized enveloping algebra $U_{r,s}(\mathfrak{g})$. We also show that all prime ideals of $U_{r,s}^{+}(\mathfrak{g})$ are completely prime under some mild conditions on the parameters r, s. Second, we study the two-parameter extended Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$. In particular, we define a Hopf algebra structure on $\overline{H_{r,s}(\Lambda)}$; and we prove that $U_{r,s}^{\geq 0}(\mathfrak{g})$ is isomorphic as a Hopf algebra to the two-parameter extended Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$.

Introduction

The interest in two-parameter quantum groups (or multiparameter quantum groups) arose in the early 1990s. Various definitions (or constructions) of two-parameter quantum groups (or multiparameter quantum groups) have appeared in the vast literature [Artin et al. 1991; Chin and Musson 1996; Dobrev and Parashar 1993; Doi and Takeuchi 1994; Jing 1992; Kulish 1990; Reshetikhin 1990; Sudbery 1990]. In particular, Takeuchi [1990] defined the two-parameter quantum groups associated to the general linear Lie algebras $\mathfrak{gl}_n$ and the special linear Lie algebras $\mathfrak{sl}_n$. These quantum groups are certain two-parameter deformations of the universal enveloping algebras $U(\mathfrak{g})$ of the Lie algebras $\mathfrak{g}$. Motivated by the connections

MSC2000: 17B37.

Keywords: two-parameter quantized enveloping algebras, two-parameter Ringel–Hall algebras, skew polynomial rings, completely prime ideals.

Partially supported by Fayetteville State University Faculty Development Research Grant.

to the study of down-up algebras, Takeuchi's two-parameter quantum groups have been reinvestigated by Benkart and Witherspoon [2004b]. In contrast to Takeuchi's two-parameter quantum groups, the two-parameter quantum groups they defined have the opposite coproduct.

Recently, more research efforts have been focused on finding similar constructions of the two-parameter quantum groups associated to other finite-dimensional complex simple Lie algebras $\mathfrak{g}$ and Kac–Moody algebras, and studying their ring-theoretic properties and representation theory [Bergeron et al. 2006; Benkart et al. 2006; Benkart and Witherspoon 2004a; Hu et al. 2008; Hu and Pei 2008]. For the finite-dimensional simple Lie algebras $\mathfrak{g}$, Hu and Pei [2008] formulated a uniform construction of the two-parameter quantum groups $U_{r,s}(\mathfrak{g})$ in terms of Ringel form. All these constructions and their modifications can also be unified by using the methods of Kharchenko [2002; 1999], in which a variety of quantum enveloping algebras were constructed from certain quantification matrices. The two-parameter quantum groups $U_{r,s}(\mathfrak{g})$ are similar to the one-parameter quantum groups in many aspects: They are also Hopf algebras and admit both the triangular decompositions and the Drinfeld double realizations. Indeed, they share a similar representation and structure theory with their one-parameter analogue. However, the two-parameter quantum groups possess a more complicated structure and less symmetry, which makes them more difficult to study.

To effectively study the two-parameter quantum groups $U_{r,s}(\mathfrak{g})$, it is natural to first study their important subalgebras such as $U_{r,s}^{+}(\mathfrak{g})$ and $U_{r,s}^{\geq 0}(\mathfrak{g})$, which can be regarded as the two-parameter quantized enveloping algebras of the nilpotent subalgebras $\mathfrak{n}^{+}$ and the Borel subalgebras $\mathfrak{b}^{+}$ of $\mathfrak{g}$.

In this paper, we will study these algebras from the viewpoint of two-parameter Ringel–Hall algebras. This approach has played a very important role in the study of one-parameter quantum groups $U_q(\mathfrak{g})$ [Green 1995; Lusztig 1993; 1990; Ringel 1990b; 1996; 1993; 1990c; 1990a; Xiao 1997]. It is well known that the quantized enveloping algebras $U_q(\mathfrak{g})$ can be realized as the reduced Drinfeld doubles of the extended Ringel–Hall algebras $H_{v}^{*}(\Lambda)$ of a certain finite-dimensional hereditary algebra associated to $\mathfrak{g}$. The one-parameter quantized enveloping algebra $U_q(\mathfrak{g})$ is first defined via generators and relations. Thus a first priority in the study of $U_q(\mathfrak{g})$ is to provide more information on the Hopf algebra structure of $U_q(\mathfrak{g})$ and construct good bases for $U_q(\mathfrak{g})$ as an algebra. These can be successfully fulfilled by using the Ringel–Hall algebra realization of $U_q(\mathfrak{g})$. Furthermore, this realization contributes significantly to the construction of canonical bases for $U_q^{+}(\mathfrak{g})$ via the representation theory of finite-dimensional hereditary algebras [Ringel 1996; Lusztig 1990]. For more details about Ringel–Hall algebras and their applications to the study of one-parameter quantum groups $U_q(\mathfrak{g})$, see [Green 1995; Ringel 1990b; 1996] and the references therein.

Let $\mathfrak{g}$ be a finite-dimensional complex simple Lie algebra of type A, D, E and Λ be the finite-dimensional hereditary algebra associated to $\mathfrak{g}$. Indeed, Λ is the path algebra of the corresponding Dynkin quiver associated to $\mathfrak{g}$. Reineke [2001], for the purpose of studying the monoid ring arising from generic extensions, defined a two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ of Λ. He proved that $H_{r,s}(\Lambda)$ is isomorphic to Takeuchi's two-parameter quantization $\mathcal{U}_{r,s}(\mathfrak{n}^+)$, where $\mathcal{U}_{r,s}(\mathfrak{n}^+) = U_{r,s}^+(\mathfrak{g})$ is the two-parameter quantized enveloping algebra of the nilpotent subalgebra $\mathfrak{n}^+$ of the Lie algebra $\mathfrak{g}$.

In this paper, we will generalize Reineke's definition to any finite-dimensional complex simple Lie algebra $\mathfrak{g}$. In the case of nonsimply connected Lie algebras, there are no quivers and path algebras available, so we will take Λ to be the corresponding finite-dimensional hereditary tensor algebra associated to the k-species [Dlab and Ringel 1975; 1976]. By Ringel's results [1990c; 1990a], the Hall polynomials exist for the extensions between modules in the category Λ-mod. Reineke's definition of two-parameter Ringel–Hall algebras depends solely on the existence of Hall polynomials, and it can thus be applied to all finite-dimensional complex Lie algebras. For any finite-dimensional simple Lie algebra $\mathfrak{g}$, we shall prove that the two-parameter quantized enveloping algebra $U_{r,s}^+(\mathfrak{g})$ is isomorphic to the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$, where Λ is the corresponding finite-dimensional hereditary algebra associated to $\mathfrak{g}$ [Dlab and Ringel 1975; 1976].

Following Ringel [1996], we shall carry out some standard calculations inside the algebra $H_{r,s}(\Lambda)$. As a result, we are able to prove that $H_{r,s}(\Lambda)$ can be presented as an iterated skew polynomial ring. An immediate application is that the skew-polynomial ring presentation of $H_{r,s}(\Lambda)$ will yield a natural PBW-basis for $U_{r,s}^+(\mathfrak{g})$ through the previous isomorphism. We further prove that all prime ideals of $U_{r,s}^+(\mathfrak{g})$ are completely prime based on some mild conditions on the parameters r, s. This result has also been proved in [Benkart et al. 2006] for the case of the Lie algebra $\mathfrak{g} = sl_n$ by using results in [Kharchenko 2002].

For the purpose of studying the two-parameter quantized enveloping algebra $U_{r,s}^{\geq 0}(\mathfrak{g})$, we will extend the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ by adding the torus part to it. Furthermore, we will define a Hopf algebra structure on the extended Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$. In particular, we will prove that $U_{r,s}^{\geq 0}(\mathfrak{g})$ is isomorphic to the extended Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$ as a Hopf algebra. The result gives the possibility of realizing the two-parameter quantum groups $U_{r,s}(\mathfrak{g})$ as the Drinfeld doubles of two-parameter extended Ringel–Hall algebras associated to certain finite-dimensional hereditary algebras Λ.

The paper is organized as follows. In Section 1, we recall the definition and some basic results of two-parameter quantum groups. In Section 2, we recall Reineke's construction of two-parameter Ringel–Hall algebras and prove some basic results. In Section 3, we define the extended two-parameter Ringel–Hall algebras and then propose a Hopf algebra structure for it. We establish the Hopf algebra isomorphism between the two-parameter quantization $U^{\geq 0}_{r,s}(\mathfrak{g})$ and the extended two-parameter Ringel–Hall algebra.

1. Definition and basic properties of the two-parameter quantum groups $U_{r,s}(\mathfrak{g})$

Let $\mathfrak{g}$ be a finite-dimensional complex simple Lie algebra. The two-parameter quantum groups $U_{r,s}(\mathfrak{g})$ associated to $\mathfrak{g}$ have been constructed in the literature [Bergeron et al. 2006; Benkart and Witherspoon 2004b; Takeuchi 1990]. A simpler uniform definition of $U_{r,s}(\mathfrak{g})$ in terms of Ringel form has recently been proposed in [Hu and Pei 2008]; we now recall the definition of $U_{r,s}(\mathfrak{g})$ used there and state some basic properties of them.

Let $C = (a_{ij})_{i,j\in I}$ be the Cartan matrix corresponding to the Lie algebra $\mathfrak{g}$. Let $\{d_i \mid i \in I\}$ be a set of relatively prime positive integers such that $d_i a_{ij} = d_j a_{ji}$ for $i, j \in I$. Let $\mathbb{Q}(r, s)$ be the function field in two variables r, s over the field $\mathbb{Q}$ of all rational numbers. We may also choose complex numbers $r, s \in \mathbb{C}$ so that $r^2 \neq s^2$. Let us write $r_i = r^{d_i}$ and $s_i = s^{d_i}$.

Let $\langle \cdot , \cdot \rangle$ be the corresponding bilinear form (so-called Ringel or Euler form) defined on the root lattice $\mathfrak{Q} \cong \mathbb{Z}^I$ associated to $\mathfrak{g}$. More precisely, the bilinear form is defined as follows:

$$\langle i, j\rangle := \langle \alpha_i, \alpha_j\rangle = \begin{cases} d_i a_{ij} & \text{if } i < j, \\ d_i & \text{if } i = j, \\ 0 & \text{if } i > j. \end{cases}$$

Definition 1.1 [Bergeron et al. 2006; Benkart and Witherspoon 2004b; Hu and Pei 2008]. The two-parameter quantum groups $U_{r,s}(\mathfrak{g})$ are the $\mathbb{Q}(r, s)$-algebras generated by the generators $e_i, f_i, w_i^{\pm 1}, w_i'^{\pm 1}$ subject to the relations

$$\begin{aligned} w_i^{\pm 1} w_j^{\pm 1} &= w_j^{\pm 1} w_i^{\pm 1}, & w_i'^{\pm 1} w_j'^{\pm 1} &= w_j'^{\pm 1} w_i'^{\pm 1}, \\ w_i^{\pm 1} w_j'^{\pm 1} &= w_j'^{\pm 1} w_i^{\pm 1}, & w_i^{\pm 1} w_i^{\mp 1} &= 1 = w_i'^{\pm 1} w_i'^{\mp 1}, \\ w_i e_j &= r^{\langle j,i\rangle} s^{-\langle i,j\rangle} e_j w_i, & w_i' e_j &= r^{-\langle i,j\rangle} s^{\langle j,i\rangle} e_j w_i', \\ w_i f_j &= r^{-\langle j,i\rangle} s^{\langle i,j\rangle} f_j e_i, & w_i' f_j &= r^{\langle i,j\rangle} s^{-\langle j,i\rangle} f_j w_i', \end{aligned}$$

$$e_i f_j - f_j e_i = \delta_{i,j} (w_i - w_i')/(r_i - s_i),$$

$$\sum_{k=0}^{1-a_{ij}} (-1)^k \binom{1-a_{ij}}{k}_{r_i s_i^{-1}} c_{ij}^{(k)} e_i^{1-a_{ij}-k} e_j e_i^k = 0 \quad \text{for } i \neq j,$$

$$\sum_{k=0}^{1-a_{ij}} (-1)^k \binom{1-a_{ij}}{k}_{r_i s_i^{-1}} c_{ij}^{(k)} f_i^k f_j f_i^{1-a_{ij}-k} = 0 \quad \text{for } i \neq j,$$

where $c_{ij}^{(k)} = (r_i s_i^{-1})^{k(k-1)/2} r^{k\langle j,i\rangle} s^{-k\langle i,j\rangle}$ for $i \neq j$, and for a symbol v, we set up the notation

$$(n)_v = \frac{v^n - 1}{v - 1}, \qquad (n)_v! = (1)_v (2)_v \cdots (n)_v,$$
$$\binom{n}{k}_v = \frac{(n)_v!}{(k)_v!(n-k)_v!} \quad \text{for } n \geq k \geq 0,$$

and $(0)_v! = 1$.

Remark 1.1. In the sequel, if needed, we may change the base field from the function field $\mathbb{Q}(r, s)$ to the complex number field $\mathbb{C}$ by choosing $r, s \in \mathbb{C}$ in so that $r^m s^n = 1$ implies $n = m = 0$, or we may restrict the base ring to the rational number field $\mathbb{Q}$ or the local ring $\mathbb{Q}[r, s]_{(r-1,s-1)}$.

From [Bergeron et al. 2006; Benkart and Witherspoon 2004b; Hu and Pei 2008], we know the algebra $U_{r,s}(\mathfrak{g})$ has a Hopf algebra structure with the corresponding comultiplication, counit and antipode defined as follows:

$$\begin{aligned}
\Delta(\omega_i^{\pm 1}) &= \omega_i^{\pm 1} \otimes \omega_i^{\pm 1}, & \Delta(\omega_i'^{\pm 1}) &= \omega_i'^{\pm 1} \otimes \omega_i'^{\pm 1},\\
\Delta(e_i) &= e_i \otimes 1 + \omega_i \otimes e_i, & \Delta(f_i) &= 1 \otimes f_i + f_i \otimes \omega_i',\\
\epsilon(\omega_i^{\pm 1}) &= \epsilon(\omega_i'^{\pm 1}) = 1, & \epsilon(e_i) &= \epsilon(f_i) = 0,\\
S(\omega_i^{\pm 1}) &= \omega_i^{\mp 1}, & S(\omega_i'^{\pm 1}) &= \omega_i'^{\mp 1},\\
S(e_i) &= -\omega_i^{-1} e_i, & S(f_i) &= -f_i \omega_i'^{-1}.
\end{aligned}$$

Let $U_{r,s}^+(\mathfrak{g})$ and $U_{r,s}^-(\mathfrak{g})$ be the subalgebras of $U_{r,s}(\mathfrak{g})$ generated by e_i for $i \in I$ and by f_i for $i \in I$, respectively. Let $U_{r,s}^0(\mathfrak{g})$ be the subalgebra of $U_{r,s}(\mathfrak{g})$ generated by $\omega_i^{\pm 1}, \omega_i'^{\pm 1}$ for $i \in I$. The following result about the triangular decomposition of $U_{r,s}(\mathfrak{g})$ was obtained in the papers above.

Proposition 1.1. *$U_{r,s}(\mathfrak{g})$ has the standard triangular decomposition*

$$U_{r,s}(\mathfrak{g}) \cong U_{r,s}^-(\mathfrak{g}) \otimes U_{r,s}^0(\mathfrak{g}) \otimes U_{r,s}^+(\mathfrak{g}).$$

Let us denote by $\mathbb{Z}^I$ the free abelian group of rank $|I|$ with a basis denoted by $z_1, z_2, \ldots, z_{|I|}$. Given an element $\boldsymbol{a} \in \mathbb{Z}^I$, say $\boldsymbol{a} = \sum a_i z_i$, we set $|\boldsymbol{a}| = \sum a_i$. The algebras $U_{r,s}^+(\mathfrak{g})$ and $U_{r,s}^-(\mathfrak{g})$ are $\mathbb{Z}^I$-graded algebras by assigning to the generator e_i and f_i, respectively, the degree z_i. Given $\boldsymbol{a} \in \mathbb{Z}^I$, we denote by $U_{r,s}^{\pm}(\mathfrak{g})_{\boldsymbol{a}}$ the set of homogeneous elements of degree $\boldsymbol{a}$ in $U_{r,s}^{\pm}(\mathfrak{g})$; thus we have the decomposition

$$U_{r,s}^+(\mathfrak{g}) = \bigoplus_{\boldsymbol{a}} U_{r,s}^+(\mathfrak{g})_{\boldsymbol{a}} \quad \text{and} \quad U_{r,s}^-(\mathfrak{g}) = \bigoplus_{\boldsymbol{a}} U_{r,s}^-(\mathfrak{g})_{\boldsymbol{a}}.$$

Let $U_{r,s}^{\geq 0}(\mathfrak{g})$ (respectively $U_{r,s}^{\leq 0}(\mathfrak{g})$) be the subalgebra of $U_{r,s}(\mathfrak{g})$ generated by $e_i, \omega_i^{\pm 1}$ (respectively $f_i, \omega_i'^{\pm 1}$), then we have the following result.

Proposition 1.2 [Bergeron et al. 2006; Benkart and Witherspoon 2004b; Hu and Pei 2008]. *The algebra $U_{r,s}(\mathfrak{g})$ can be realized as a Drinfeld double of Hopf subalgebras $U_{r,s}^{\geq 0}(\mathfrak{g})$ and $U_{r,s}^{\leq 0}(\mathfrak{g})$ with respect to the pairing $(\cdot\,,\cdot)$, that is,*

$$U_{r,s}(\mathfrak{g}) \cong D(U_{r,s}^{\geq 0}(\mathfrak{g}), U_{r,s}^{\leq 0}(\mathfrak{g})).$$

To better understand $U_{r,s}(\mathfrak{g})$, it is natural to further study the subalgebras $U_{r,s}^{+}(\mathfrak{g})$ and $U_{r,s}^{\geq 0}(\mathfrak{g})$. We will address this problem in the forthcoming sections via the approach of Ringel–Hall algebras.

2. Two-parameter Ringel–Hall algebras $H_{r,s}(\Lambda)$

In this section, we will first recall Reineke's construction of the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$, where Λ denotes the finite-dimensional hereditary algebra associated to a complex simple Lie algebra $\mathfrak{g}$ of type A, D, E. Indeed, Λ is the path algebra of the corresponding Dynkin quiver associated to $\mathfrak{g}$. Then we will define the corresponding two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ of the finite-dimensional hereditary algebra Λ, which is the corresponding finite-dimensional hereditary algebra associated to any finite-dimensional complex simple Lie algebra $\mathfrak{g}$. We will take Λ as the tensor algebras of the associated k-species in the nonsimply connected cases. Note that Reineke's construction is still valid due to the existence of Hall polynomials for Λ-modules.

To further study the properties of $H_{r,s}(\Lambda)$, we will carry out some calculations similar to ones done in [Ringel 1996]. These will yield a skew polynomial ring presentation of $H_{r,s}(\Lambda)$, which immediately enables us to construct a PBW-basis for $H_{r,s}(\Lambda)$. This PBW-basis will be used to construct a PBW-basis for $U_{r,s}(\mathfrak{g})$. Based on certain mild restrictions on the parameters r, s, using the stratification theory of prime ideals developed in [Goodearl and Letzter 2000], we will further prove that all prime ideals of $H_{r,s}(\Lambda)$ are completely prime. Finally, we establish the relationship between the algebra $U_{r,s}^{+}(\mathfrak{g})$ and the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ by proving that they are isomorphic to each other as algebras. Thus all the results obtained on $H_{r,s}(\Lambda)$ can be transformed to $U_{r,s}^{+}(\mathfrak{g})$ via this algebra isomorphism.

2.1. ***Preliminaries on k-species.*** In this subsection, for the reader's convenience, we shall recall some basic information about k-species. The study of k-species is a very important research topic that has generated a vast literature. We shall only briefly mention some results that relate the study of k-species to the study of finite-dimensional hereditary algebras and Lie algebras, and cite some relevant references. See [Dlab and Ringel 1975; 1976; Ringel 1976] and the references therein for a detailed account of the structure and representation theory of k-species and their connections to other subjects. In particular, the following presentation of the material is borrowed from the Dlab and Ringel references.

Gabriel [1972] observed that there is a one-to-one correspondence between the set of indecomposable representations of these graphs ("quivers") with a positive definite quadratic form and the set of positive roots of this quadratic form. Later, Bernšteĭn, Gel'fand and Ponomarev [1973] showed that this result can be proved directly, by using appropriate functors (the BGP reflection functors) to construct all indecomposable representations from the simple ones in the same way that the canonical generators of the Weyl group are used to produce all positive roots from the simple roots. Dlab and Ringel [1975; 1976] extended this method. To deal with all Dynkin diagrams (not necessarily those of type A, D or E), they further considered valued graphs (and therefore "species"). For the valued graphs of Dynkin type, they obtained the same one-to-one correspondence between the set of indecomposable representations and the set of positive roots of the corresponding quadratic form, thus generalizing Gabriel's result for the type A, D or E diagrams. In [1976], they also considered valued graphs with positive semidefinite quadratic form (that is, extended Dynkin diagrams) and described all the indecomposable representations up to homogeneous ones.

A valued graph $\Gamma := (\Gamma, d)$ consists of a finite set Γ (of vertices) together with a set d of nonnegative integers d_{ij} for all $i, j \in \Gamma$ such that $d_{ii} = 0$ and there exist positive integers $\{\epsilon_i\}_{i\in\Gamma}$ that satisfy

$$d_{ij}\epsilon_j = d_{ji}\epsilon_i \quad \text{for all } i, j \in \Gamma.$$

A pair $\{i, j\}$ of vertices of Γ is called an edge of the graph Γ if $d_{ij} \neq 0$. An orientation of Ω of a valued graph (Γ, d) is given by assigning each edge $\{i, j\}$ of Γ an order (which is denoted by an arrow $i \to j$). We usually call (Γ, d, Ω) a valued quiver. Given any orientation Ω and any vertex $i \in \Gamma$, we can define a new orientation $s_i\Omega$ of (Γ, d) by reversing the direction of the arrows along all edges containing i. A vertex $i \in \Gamma$ is called a sink (or source) with respect to the orientation Ω if $i \leftarrow j$ (or $i \to j$) for all neighbor vertices $j \in \Gamma$ of i. An orientation is said to be admissible if there is an ordering $i_1, i_2, \ldots, i_n$ of Γ such that each vertex i_t is a sink with respect to the orientation $s_{i_{t-1}} \cdots s_{i_2}s_{i_1}\Omega$ for all $1 \le t \le n$; such an ordering is called an admissible ordering for Ω.

For a given valued graph $\Gamma = (\Gamma, d)$, one can associate a symmetrizable Cartan matrix $C = (a_{ij})_{i,j\in\Gamma}$ by setting the entries of C as follows:

$$a_{ii} = 2 \quad \text{and} \quad a_{ij} = -d_{ij} \quad \text{for} \quad i \neq j \in \Gamma.$$

Conversely, for any symmetrizable Cartan matrix C, one can associate a valued graph Γ_C as well. It is easy to see that the mapping from the valued graph (Γ, d) to the Cartan matrix C defines a one-to-one correspondence between the set of valued graphs and the set of symmetrizable Cartan matrices. This correspondence indicates a close relationship between the study of valued graphs and the study of Lie algebras, as we shall further explain below.

Let k denote a finite field and let (Γ, d, Ω) be a valued graph together with an admissible orientation Ω. Following [Gabriel 1973], by definition, a k-species $\mathcal{S} = (\mathcal{M}, \Omega) := (F_i, {}_iM_j)$ of type (Γ, d, Ω) (which is also called a realization of the valued graph (Γ, d, Ω) in [Dlab and Ringel 1976]) consists of a family of $(F_i - F_j)$-bimodules ${}_iM_j$, where the fields F_i are finite field extensions of k in an algebraic closure of k such that $\dim_k F_i = \epsilon_i$ and $\dim({}_iM_j)_{F_j} = d_{ij}$. Note that $\mathcal{S}$ is called connected provided the corresponding graph is connected; an oriented cycle of $\mathcal{S}$ is given by a sequence of vertices $i_1, i_2, \dots, i_{k-1}, i_k = i_1$ such that $i_j \to i_{j+1}$ for all $1 \le j \le k-1$. From now on, we shall always assume that (Γ, d, Ω) is connected and contains no oriented cycles.

A representation $(V_i, {}_j\phi_i)$ of the k-species $\mathcal{S}$ is given by a set of vector spaces $(V_i)_{F_i}$ and F_j-linear mappings $V_i \otimes {}_iM_j \to V_j$. Such a representation is called finite-dimensional provided all the vector spaces V_i are finite-dimensional vector spaces. A homomorphism $\alpha = (\alpha_i) : (V_i, {}_j\phi_i) \to (V_i', {}_j\phi_i')$ is given by a set of F_i-linear mappings $\alpha_i : V_i \to V_i'$ such that $\alpha_j {}_j\phi_i = {}_j\phi_i'(\alpha_i \otimes 1)$. We shall denote by $\operatorname{rep} \mathcal{S} = L(\mathcal{M}, \Omega)$ the category of all finite-dimensional representations of $(\mathcal{M}, \Omega)$. It is an abelian category. A k-species $\mathcal{S}$ is said to be of finite representation type if the category $\operatorname{rep} \mathcal{S}$ has only finitely many indecomposable objects.

[Dlab and Ringel 1975, Theorem B]. *A k-species is of finite representation type if and only if its diagram is a finite union of Dynkin diagrams.*

Given a k-species $\mathcal{S}$, one denotes by $\mathbb{Q}^\Gamma$ the rational vector space of all vectors $\boldsymbol{x} = (x_i)_{i \in \Gamma}$ over the rational number field. There is a quadratic form defined on $\mathbb{Q}^n$ where $n = |\Gamma|$ as follows: For any $x \in \mathbb{Q}^n$, let

$$(x, x) = \sum \epsilon_i x_i^2 - \sum m_{ij} x_i x_j,$$

where $\epsilon_i = \dim_k F_i$ and $m_{ij} = \dim_k({}_iM_j)$. Given any representation $(V_i, {}_j\phi_i)$ of the k-species $(\mathcal{M}, \Omega)$, one can define the dimension vector mapping

$$\dim : L(\mathcal{M}, \Omega) \to \mathbb{Q}^\Gamma$$

by setting $\dim(V) = (x_i)$, where $x_i = \dim(V_i)_{F_i}$ for all $i \in \Gamma$. Dlab and Ringel [1975; 1976] proved that the k-species $\mathcal{S}$ is of finite representation type if and only if the corresponding quadratic form is positive definite, that is, the underlying graph is a Dynkin diagram. In particular, we shall quote the following two results:

[Dlab and Ringel 1976, Proposition 1.2]. (a) (Γ, d) *is a Dynkin diagram if and only if its quadratic form is positive definite.*

(b) (Γ, d) *is an extended Dynkin diagram if and only if its quadratic form is positive semi-definite.*

[Dlab and Ringel 1976, Proposition 2.6]. *Let* $(\mathcal{M}, \Omega)$ *be a realization of the valued graph* (Γ, d).

(a) *If* (Γ, d) *is a Dynkin diagram, then the mapping* dim *provides a one-to-one correspondence between all positive roots of* (Γ, d) *and all indecomposable representations in* $L(\mathcal{M}, \Omega)$.

(b) *If* (Γ, d) *is an extended Dynkin diagram, then the mapping* dim *provides a one-to-one correspondence between all positive roots of* (Γ, d) *of nonzero defect and all indecomposable representations in* $L(\mathcal{M}, \Omega)$ *of nonzero defect.*

The results on the representations of valued graphs can be translated into the representation theory of finite-dimensional associative algebras over a field, or more generally into that of certain classes of artinian rings [Dlab and Ringel 1976]. For any given artinian ring R, one can define a valued graph, we will not discuss the detailed construction here. Conversely, for any given k-species $\mathcal{S}$ on a given valued graph (Γ, d), one can define its associated tensor algebra $\Lambda = T(\mathcal{S})$ by

$$\Lambda = \bigoplus_{t \geq 0} \Lambda^{(t)}$$

where

$$\Lambda^{(0)} = \prod_{i \in \Gamma} F_i, \quad \Lambda^{(1)} = \prod_{h \in \Omega} {}_iM_j, \quad \text{and} \quad \Lambda^{(n)} = \Lambda^{(n-1)} \otimes_{\Lambda^{(0)}} \Lambda^{(1)} \quad \text{for } t \geq 2$$

with the componentwise addition and the multiplication induced by taking tensor products. Note that for an admissible orientation Ω of (Γ, d), the tensor algebra Λ of (Γ, d, Ω) is a finite-dimensional hereditary k-algebra. An algebra R is said to be of finite representation type if there are only finitely many indecomposable finite-dimensional R-modules. Each finite-dimensional hereditary k-algebra of finite representation type can be identified with the tensor algebra of some k-species. It is well known [Dlab and Ringel 1975] that the category Λ-mod of finite-dimensional Λ-modules is equivalent to the category rep$(\mathcal{S})$ of finite-dimensional representations of the k-species $\mathcal{S}$ over the field k.

[Dlab and Ringel 1975, Theorem C]. *A finite-dimensional k-algebra R is hereditary of finite representation type if and only if R is Morita equivalent to the tensor algebra $T(\mathcal{S})$, where $\mathcal{S}$ is a k-species of finite representation type.*

In the rest of this paper, we will not distinguish between these two categories.

2.2. ***Two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$.*** As above, let k be a finite field. Let Λ denote a finite-dimensional associative hereditary algebra over k. Denote by $q = |k|$ the cardinality of the base field k, and let v be a number such that $v^2 = q$. We shall also assume that Λ is finitary, that is, the cardinality of the extension group $\mathrm{Ext}^1(S, S')$ is finite for any two simple Λ-modules S and S'.

It is well known that this finitary condition is satisfied by the algebra Λ as long as Λ is a finitely generated k-algebra over a finite base field k. By $\mathcal{P}$, we will denote the set of all isomorphism classes of finite-dimensional Λ-modules. We set $\mathcal{P}_1 = \mathcal{P} - 0$, where 0 denotes the subset of $\mathcal{P}$ consisting of the only isomorphism class of the zero Λ-module. For any $\alpha \in \mathcal{P}$, we choose a module representative u_α for the isomorphism class α. We denote by a_α the order of the automorphism group $\mathrm{Aut}_\Lambda(u_\alpha)$ of the Λ-module u_α. It is easy to see that the number a_α is independent of the choices of the representatives u_α for any $\alpha \in \mathcal{P}$.

For any given three representatives $u_\alpha, u_\beta, u_\gamma$ of the elements $\alpha, \beta, \gamma \in \mathcal{P}$ respectively, we denote by $g^\gamma_{\alpha\beta}$ the number of submodules N of u_γ satisfying the conditions $N \cong u_\beta$ and $u_\gamma/N \cong u_\alpha$.

For any two given Λ-modules M, N, let us set

$$\langle M, N\rangle = \dim_k Hom(M, N) - \dim_k \mathrm{Ext}^1(M, N).$$

Since the algebra Λ is hereditary, it is well known that $\langle M, N\rangle$ depends only on the dimension vectors $\dim(M)$ and $\dim(N)$ of the Λ-modules M and N. Thus for any given two elements $\alpha, \beta \in \mathcal{P}$, we can define the notation

$$\langle \alpha, \beta\rangle = \langle u_\alpha, u_\beta\rangle$$

where u_α and u_β are any chosen representatives of α and β respectively. Note that $\langle \cdot, \cdot\rangle$ is a bilinear form that is not necessarily symmetric. However, using $\langle \cdot, \cdot\rangle$, we can also define a symmetric bilinear form $(\cdot, \cdot)$ by setting

$$(\alpha, \beta) = \langle \alpha, \beta\rangle + \langle \beta, \alpha\rangle.$$

In the rest of this paper, we will be mostly dealing with the form $\langle \cdot, \cdot\rangle$ instead.

Let Λ-mod denote the category of all finite-dimensional Λ-modules. Note that there exists a fine symmetry between elements in the category Λ-mod:

Theorem 2.1 [Green 1995, first formula]. *Assume that Λ is hereditary and finitary. Let $\alpha, \beta, \alpha', \beta' \in \mathcal{P}$. Then*

$$a_\alpha a_\beta a_{\alpha'} a_{\beta'} \sum_{\lambda\in\mathcal{P}} g^\lambda_{\alpha,\beta} g^\lambda_{\alpha'\beta'} a_\lambda^{-1} = \sum_{\rho,\sigma,\sigma',\tau\in\mathcal{P}} \frac{|\mathrm{Ext}^1(u_\rho, u_\tau)|}{|\mathrm{Hom}(u_\rho, u_\tau)|} g^\alpha_{\rho\sigma} g^{\alpha'}_{\rho\sigma'} g^\beta_{\sigma'\tau} g^{\beta'}_{\sigma\tau} a_\rho a_\sigma a_{\sigma'} a_{\tau'}.$$

Let $\mathfrak{g}$ be a finite-dimensional complex simple Lie algebra of type A, D or E; and let Λ be the finite-dimensional hereditary algebra associated to $\mathfrak{g}$. As a two-parameter twist of Ringel–Hall algebra, the two-parameter Ringel–Hall algebra

$H_{r,s}(\Lambda)$ was first defined by Reineke [2001] for the purpose of studying the monoid ring of generic extensions. We will first recall some details of its construction.

Since $\mathfrak{g}$ is a finite-dimensional complex simple Lie algebra of type A, D or E, we can associate a Dynkin quiver $\vec{\Delta}$ to the Lie algebra $\mathfrak{g}$, so that the path algebra $\Lambda := k\vec{\Delta}$ of the Dynkin quiver $\vec{\Delta}$ is a finite-dimensional hereditary algebra of finite representation type. Reineke [2001] introduced a two-parameter Ringel–Hall algebra, which was used to realize Takeuchi's two-parameter quantization $\mathcal{U}_{r,s}(\mathfrak{n}^+)$, with $\mathfrak{n}^+$ the maximal nilpotent Lie subalgebra of the Lie algebra $\mathfrak{g}$. To avoid colliding notation, we will denote Reineke's version of the two-parameter Ringel–Hall algebra by $H_{r,s}(\Lambda)$ instead of the original $H(\mathcal{Q})$. Reineke proved that the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ is indeed isomorphic to the two-parameter quantization $\mathcal{U}^+_{r,s}(\mathfrak{n}^+)$.

It is natural to extend Reineke's construction to finite-dimensional complex simple Lie algebras of other types. This can be done in terms of k-species due to the existence of Hall polynomials [Ringel 1990a; 1990c]. We will first write down the details of the formulation of the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ for any complex simple Lie algebra $\mathfrak{g}$ of other types. During the process, the algebra Λ is taken as the tensor algebra of the k-species associated to the nonsimply connected simple complex Lie algebra $\mathfrak{g}$. Then we will show that two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ is isomorphic to the two-parameter quantization $U^+_{r,s}(\mathfrak{g})$ for any Lie algebra $\mathfrak{g}$, which generalizes Reineke's result of [2001].

From now on, we will always assume that $\mathfrak{g}$ is a finite-dimensional complex simple algebra and let Λ be the corresponding hereditary path algebra (or the tensor algebra of the k-species for nonsimply connected cases). Note that there exist Hall polynomials $F^L_{M,N}(x)$ associated to modules M, N and L in Λ-mod such that for these Λ-modules, we have $g^L_{M,N} = F^L_{M,N}(q)$, where q is the cardinality of the base field k. For a detailed account of the existence and calculation of Hall polynomials in Λ-mod, see [Ringel 1996; 1993].

Recall that $\mathcal{P}$ is the set of isomorphism classes of finite-dimensional Λ-modules. Let us denote by $H_{r,s}(\Lambda)$ the free $\mathbb{Q}(r,s)$-module generated by the elements of the set $\{u_\alpha \mid \alpha \in \mathcal{P}\}$. In addition, we define a multiplication on the free $\mathbb{Q}(r,s)$-module $H_{r,s}(\Lambda)$ by

$$u_\alpha u_\beta = \sum_{\lambda\in\mathcal{P}} s^{-\langle\alpha,\beta\rangle} F^{u_\lambda}_{u_\alpha u_\beta}(rs^{-1})u_\lambda \quad \text{for any}\, \alpha, \beta \in \mathcal{P}.$$

Then it is easy to see that we have the following result:

Proposition 2.1 (see also [Reineke 2001]). *If Λ is the finite-dimensional corresponding hereditary algebra associated to the Lie algebra $\mathfrak{g}$, then the algebra $H_{r,s}(\Lambda)$ is an associative $\mathbb{Q}(r,s)$-algebra under the multiplication defined above.*

The proof is a straightforward verification and we will omit it.

2.3. ***Ring theoretical properties of $H_{r,s}(\Lambda)$.*** Now we will investigate some ring-theoretic properties of the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$. We first verify some basic identities for $H_{r,s}(\Lambda)$ following along the lines in [Ringel 1996]. These calculations are the same as those done there, with some slight modifications.

First, we introduce a new $\mathbb{Q}(r,s)$-basis for $H_{r,s}(\Lambda)$. For any chosen element $\alpha \in \mathcal{P}$, we have an element $u_\alpha \in H_{r,s}(\Lambda)$. We denote by $\epsilon(\alpha)$ the k-dimension of the endomorphism ring of the module representative u_α corresponding to α.

For any given module M of the algebra Λ, we denote the isomorphism class of M by $[M]$ and by $\dim(M)$ the dimension vector of M, which is an element of the Grothendieck group $K_0(\Lambda)$ of the category Λ-mod of all finite Λ-modules modulo the exact sequences.

According to [Bernšteĭn et al. 1973; Dlab and Ringel 1975; Gabriel 1972], there is a one-to-one correspondence between the set of all positive roots for the Lie algebra $\mathfrak{g}$ and the set of indecomposable modules in Λ-mod (see Section 2.1). Let $\boldsymbol{a} \in \Phi^+$ be any positive root; we denote by $M(\boldsymbol{a})$ the indecomposable module corresponding to $\boldsymbol{a}$. For any map $\alpha : \Phi^+ \to \mathbb{N}_0$, set

$$M(\alpha) = M_\Lambda(\alpha) = \bigoplus_{\boldsymbol{a}\in\Phi^+} \alpha(\boldsymbol{a})M(\boldsymbol{a}).$$

Then it is easy to see there is a bijection between the set $\mathcal{P}$ of isomorphism classes of finite-dimensional Λ-modules and the set of all maps $\alpha : \Phi^+ \to \mathbb{N}_0$. We will not distinguish an element $\alpha \in \mathcal{P}$ from the corresponding map associated to α, and we may denote both of them by α if no confusion arises.

For any $\alpha \in \mathcal{P}$, let us set $\mathbf{dim}\,\alpha = \sum_{\boldsymbol{a}\in\Phi^+} \alpha(\boldsymbol{a})\boldsymbol{a}$. Then we have

$$\dim(M(\alpha)) = \mathbf{dim}\,\alpha.$$

For any given $\alpha \in \mathcal{P}$, we denote by $\dim(\alpha) = \dim(u_\alpha)$ the dimension of the Λ-module u_α as a k-vector space. Furthermore, let us set

$$\langle u_\alpha \rangle = s^{\dim(u_\alpha)-\epsilon(\alpha)} u_\alpha.$$

For convenience, we may sometimes simply denote u_α by α for any $\alpha \in \mathcal{P}$ and denote $F^{u_\lambda}_{u_\alpha u_\beta}(rs^{-1})$ by $g^\lambda_{\alpha\beta}$ if no confusion arises. In particular, we will carry out all the computations in terms of α instead of u_α in the rest of this subsection.

Obviously the set $\{\langle\alpha\rangle \mid \alpha \in \mathcal{P}\}$ is also a $\mathbb{Q}(r,s)$-basis for the algebra $H_{r,s}(\Lambda)$. Note that $\langle\alpha_i\rangle = \alpha_i$ for any element $\alpha_i \in \mathcal{P}$ corresponding to the simple root α_i. Thus the multiplication in $H_{r,s}(\Lambda)$ can be rewritten in terms of this new basis as

$$\langle\alpha\rangle\langle\beta\rangle = s^{-\epsilon(\alpha)-\epsilon(\beta)-\langle\mathbf{dim}\alpha,\mathbf{dim}\beta\rangle} \sum_{\lambda\in\mathcal{P}} s^{\epsilon(\lambda)} g^\lambda_{\alpha\beta}\langle\lambda\rangle \quad \text{for any } \alpha,\beta \in \mathcal{P}.$$

Furthermore, let us write

$$e(\alpha,\beta)=\dim_k \operatorname{Hom}_\Lambda(M(\alpha),M(\beta)) \quad\text{and}\quad \zeta(\alpha,\beta)=\dim_k \operatorname{Ext}^1_\Lambda(M(\alpha),M(\beta)).$$

Then we have the following proposition, similar to the one in [Ringel 1996].

Proposition 2.2. *Let $\alpha_1,\dots,\alpha_t\in\mathcal{P}$ such that for $i<j$, we have both $\epsilon(\alpha_j,\alpha_i)=0$ and $\zeta(\alpha_i,\alpha_j)=0$. Then*

$$\Big\langle\bigoplus_{i=1}^t \alpha_i\Big\rangle=\langle\alpha_1\rangle\cdots\langle\alpha_t\rangle.$$

Proof. Without of loss of generality, we may assume that $t=2$. Let us set $\alpha_1=\alpha$ and $\alpha_2=\beta$. Since $\zeta(\alpha,\beta)=0$, we have

$$\langle\alpha,\beta\rangle=e(\alpha,\beta)-\zeta(\alpha,\beta)=e(\alpha,\beta).$$

Since $e(\beta,\alpha)=0$, we also have

$$\begin{aligned} e(\alpha\oplus\beta)&=e(\alpha,\alpha)+e(\alpha,\beta)+e(\beta,\beta)+e(\beta,\alpha)\\ &=e(\alpha,\alpha)+e(\beta,\beta)+e(\alpha,\beta).\end{aligned}$$

Thus

$$\begin{aligned}&\epsilon(\alpha\oplus\beta)-\langle\alpha,\beta\rangle-e(\alpha,\alpha)-e(\beta,\beta)\\ &\qquad=e(\alpha,\alpha)+e(\beta,\beta)+e(\alpha,\beta)-e(\alpha,\beta)+\zeta(\alpha,\beta)-e(\alpha,\alpha)-e(\beta,\beta)=0.\end{aligned}$$

Since $\zeta(\alpha,\beta)=0$, that $g^\gamma_{\alpha\beta}\neq 0$ implies that $\gamma=\alpha\oplus\beta$. Since $e(\beta,\alpha)=0$, we have $g^{\alpha\oplus\beta}_{\alpha\beta}=1$. Therefore, we may finish the proof:

$$\begin{aligned}\langle\alpha\rangle\langle\beta\rangle&=s^{\dim(\alpha)+\dim(\beta)-\epsilon(\alpha)-\epsilon(\beta)}\alpha\beta\\ &=s^{\epsilon(\alpha\oplus\beta)-\langle\alpha,\beta\rangle-\epsilon(\alpha)-\epsilon(\beta)}g^{\alpha\oplus\beta}_{\alpha\beta}\langle\alpha\oplus\beta\rangle\\ &=\langle\alpha\oplus\beta\rangle.\end{aligned}$$

□

Theorem 2.2. *Let $\alpha,\beta\in\mathcal{P}$ such that $e(\beta,\alpha)=0$ and $\zeta(\alpha,\beta)=0$. Then we have*

$$\langle\beta\rangle\langle\alpha\rangle=r^{\langle\alpha,\beta\rangle}s^{-\langle\beta,\alpha\rangle}\langle\alpha\rangle\langle\beta\rangle+\sum_{\gamma\in J(\alpha,\beta)}c_\gamma\langle\gamma\rangle$$

where the coefficients c_γ are in $\mathbb{Z}[r^{\pm1},s^{\pm1}]$ and $J(\alpha,\beta)$ is the set of all elements $\lambda\in\mathcal{P}$ that are different from $\alpha\oplus\beta$ and $g^\lambda_{\alpha\beta}\neq 0$.

Proof. First, by Proposition 2.2, we have $\langle\alpha\rangle\langle\beta\rangle=\langle\alpha\oplus\beta\rangle$.

Note that $\langle\beta\rangle\langle\alpha\rangle=\sum_\gamma c'_\gamma\gamma$. Thus we have the relationship $c'_\gamma=s^{\dim(\gamma)-\epsilon(\gamma)}c_\gamma$ between the coefficients c_γ and c'_γ. By [Ringel 1996], we also have

$$g^{\alpha\oplus\beta}_{\beta\alpha}=(rs^{-1})^{e(\alpha,\beta)}.$$

Note that $\epsilon(\alpha\oplus\beta)=\epsilon(\alpha)+\epsilon(\beta)+e(\alpha,\beta)$. Thus

$$\begin{aligned}c'_{\alpha\oplus\beta}&=s^{\dim(\beta)-\epsilon(\beta)+\dim(\alpha)-\epsilon(\alpha)}s^{-\langle\beta,\alpha\rangle}g^{\alpha\oplus\beta}_{\beta\alpha}\\&=s^{\dim(\alpha\oplus\beta)-\epsilon(\alpha)-\epsilon(\beta)+\zeta(\beta,\alpha)}(rs^{-1})^{e(\alpha,\beta)}\\&=s^{\dim(\alpha\oplus\beta)-\epsilon(\alpha)-\epsilon(\beta)+\zeta(\beta,\alpha)-e(\alpha,\beta)}r^{e(\alpha,\beta)}\\&=r^{e(\alpha,\beta)}s^{\zeta(\beta,\alpha)}s^{\dim(\alpha\oplus\beta)-\epsilon(\alpha\oplus\beta)}=r^{\langle\alpha,\beta\rangle}s^{-\langle\beta,\alpha\rangle}s^{\dim(\alpha\oplus\beta)-\epsilon(\alpha\oplus\beta)}.\end{aligned}$$

Finally, we have $c'_{\alpha\oplus\beta}\alpha\oplus\beta=r^{\langle\alpha,\beta\rangle}s^{-\langle\beta,\alpha\rangle}\langle\alpha\oplus\beta\rangle$. □

According to the representation theory of finite-dimensional hereditary algebras of finite representation type [Bernšteĭn et al. 1973; Dlab and Ringel 1976], we can give a total ordering on the set of positive roots of the Lie algebra $\mathfrak{g}$. Following Ringel [1996], we will order all positive roots in a way $\boldsymbol{a}_1,\boldsymbol{a}_2,\dots,\boldsymbol{a}_m$ so that $\mathrm{Hom}(M(\boldsymbol{a}_i),M(\boldsymbol{a}_j))\neq 0$ implies $i\leq j$, where $M(\boldsymbol{a}_i)$ is the indecomposable module corresponding to the positive root $\boldsymbol{a}_i$. Such an ordering will be called $\vec{\Delta}$-admissible.

Lemma 2.1 [Ringel 1996]. *A total ordering $\boldsymbol{a}_1,\dots,\boldsymbol{a}_m$ of all the positive roots is $\vec{\Delta}$-admissible if and only if $\langle\boldsymbol{a}_i,\boldsymbol{a}_j\rangle>0$ implies $i\leq j$. Such an ordering has the additional property that $\langle\boldsymbol{a}_i,\boldsymbol{a}_j\rangle<0$ implies $i>j$.*

From now on, we will always fix such a $\vec{\Delta}$-admissible ordering on the set of all positive roots.

Proposition 2.3. *For any $\alpha\in\mathcal{P}$, we have $\langle\alpha\rangle=\langle\alpha(\boldsymbol{a}_1)\boldsymbol{a}_1\rangle\cdots\langle\alpha(\boldsymbol{a}_m)\boldsymbol{a}_m\rangle$.*

Proof. Since the ordering of the positive roots $\boldsymbol{a}_i$ is admissible, $e(\boldsymbol{a}_j,\boldsymbol{a}_i)=0$ for any $i<j$. In addition, we also have $\zeta(\boldsymbol{a}_i,\boldsymbol{a}_j)=0$. Note that, as a module, $\alpha=\bigoplus_{i=1}^m\alpha(\boldsymbol{a}_i)M(\boldsymbol{a}_i)$; then the result follows from Proposition 2.2. □

Now let us consider the divided powers of $\langle\boldsymbol{a}\rangle$ by setting

$$\langle\boldsymbol{a}\rangle^{(t)}=\frac{1}{[t]^!_{\epsilon(\boldsymbol{a})}}\langle\boldsymbol{a}\rangle^t,\quad\text{where }[t]^!_{\epsilon(\boldsymbol{a})}=\prod_{i=1}^t\frac{r^{i\epsilon(\boldsymbol{a})}-s^{i\epsilon(\boldsymbol{a})}}{r^{\epsilon(\boldsymbol{a})}-s^{\epsilon(\boldsymbol{a})}}.$$

Lemma 2.2. *Let $\boldsymbol{a}$ be a positive root and $t\geq 0$ be an integer. Then $\langle t\boldsymbol{a}\rangle=\langle\boldsymbol{a}\rangle^{(t)}$.*

Proof. The proof is adapted from [Ringel 1996]. Let $\mathcal{S}$ be a reduced k-species, where k is a finite field. Then the number of filtrations

$$tM_{\mathcal{S}}(\boldsymbol{a})=M_0\supset M_1\cdots\supset M_t=0$$

of the module $tM_{\mathcal{S}}(\boldsymbol{a})$ with composition factors isomorphic to the module $M_{\mathcal{S}}(\boldsymbol{a})$ is given by evaluating the following polynomial in x at the number $|k|=rs^{-1}$:

$$\frac{(x^{\epsilon(\boldsymbol{a})t}-1)(x^{\epsilon(\boldsymbol{a})(t-1)}-1)\cdots(x^{\epsilon(\boldsymbol{a})}-1))}{(x^{\epsilon(\boldsymbol{a})})-1)^t}.$$

Since $\zeta(\boldsymbol{a}, \boldsymbol{a}) = 0$, we have

$$\boldsymbol{a}^t = s^{-\epsilon(\boldsymbol{a})\binom{t}{2}}[t]!_{\epsilon(\boldsymbol{a})} s^{-\epsilon(\boldsymbol{a})\binom{t}{2}} t\boldsymbol{a} = s^{-\epsilon(\boldsymbol{a})t(t-1)}[t]!_{\epsilon(\boldsymbol{a})} t\boldsymbol{a}.$$

Therefore,

$$\begin{aligned}\langle \boldsymbol{a}\rangle^{(t)} &= \frac{1}{[t]!_{\epsilon(\boldsymbol{a})}}\langle \boldsymbol{a}\rangle^t = \frac{s^{-\epsilon(\boldsymbol{a})t(t-1)}[t]!_{\epsilon(\boldsymbol{a})} s^{t(\dim(\boldsymbol{a})-\epsilon(\boldsymbol{a}))}}{[t]!_{\epsilon(\boldsymbol{a})}}\boldsymbol{a}^t \\ &= s^{-\epsilon(\boldsymbol{a})t(t-1)+t(\dim(\boldsymbol{a})-\epsilon(\boldsymbol{a}))} t\boldsymbol{a} \\ &= s^{t\dim(\boldsymbol{a})-\epsilon(\boldsymbol{a})t(t-1+1)} t\boldsymbol{a} = s^{t\dim(\boldsymbol{a})-\epsilon(\boldsymbol{a})t^2} = s^{t\dim(\boldsymbol{a})-\epsilon(t\boldsymbol{a})} t\boldsymbol{a} = \langle t\boldsymbol{a}\rangle. \qquad \square\end{aligned}$$

For each positive root $\boldsymbol{a}_i$, let us define the symbol $X_i = \langle \boldsymbol{a}_i\rangle$.

Proposition 2.4. *Let $\alpha \in \mathcal{P}$. Then α can be regarded as a map $\alpha : \Phi^+ \to \mathbb{N}_0$. Setting $\alpha(i) = \alpha(\boldsymbol{a}_i)$, we have*

$$\langle \alpha\rangle = X_1^{(\alpha(1))} \cdots X_m^{(\alpha(m))} = \left(\prod_{i=1}^{m} \frac{1}{[\alpha(i)]^!_{\epsilon(\boldsymbol{a}_i)}}\right) X_1^{\alpha(1)} \cdots X_m^{\alpha(m)}.$$

Proof. By Proposition 2.3, we have $\langle\alpha\rangle = \langle \alpha(1)\boldsymbol{a}_1\rangle \cdots \langle \alpha(m)\boldsymbol{a}_m\rangle$. By Lemma 2.2, we also have $\langle \alpha(i)\boldsymbol{a}_i\rangle = X_i^{(\alpha(i))}$. Thus the first equality holds. The second equality can be proved by using the divided powers. □

Theorem 2.3. *The monomials $X_1^{\alpha(1)} \cdots X_m^{\alpha(m)}$ with $\alpha(1), \ldots, \alpha(m) \in \mathbb{N}_0$ form a $\mathbb{Q}(r, s)$-basis of $H_{r,s}(\Lambda)$, and for $i < j$, we have*

$$\begin{aligned} X_j X_i = {} & r^{\langle \dim X_i, \dim X_j\rangle} s^{-\langle \dim X_j, \dim X_i\rangle} X_i X_j \\ & + \sum_{I(i,j)} c(a_{i+1}, \ldots, a_{j-1}) X_{i+1}^{a_{i+1}} \cdots X_{j-1}^{a_{j-1}} \end{aligned}$$

with coefficients $c(a_{i+1}, \ldots, a_{j-1})$ in $\mathbb{Q}(r, s)$. Here the index set $I(i, j)$ is the set of sequences $(a_{i+1}, \cdots a_{j-1})$ of natural numbers such that $\sum_{t=i+1}^{j-1} a_t \boldsymbol{a}_t = \boldsymbol{a}_i + \boldsymbol{a}_j$.

Proof. Given $\alpha(1), \ldots, \alpha(m) \in \mathbb{N}_0$, define an element $\alpha \in \mathcal{P}$ by setting $\alpha(\boldsymbol{a}_i) = \alpha(i)$. According to the previous proposition, we have $\langle\alpha\rangle = X_1^{(\alpha(1))} \cdots X_m^{(\alpha(m))}$. Thus the monomials given in the theorem are exactly nonzero scalar multiples of the elements in $\mathcal{P}$. Therefore, these monomials form a $\mathbb{Q}(r, s)$-basis of $H_{r,s}(\Lambda)$.

Let $i < j$. We can apply Theorem 2.2 to the positive roots $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$. We need to show that for any $\beta \in J(i, j)$, the element β is a scalar multiple of some monomials $X_{i+1}^{a_{i+1}} \cdots X_j^{a_j}$ with $\sum_{t=i+1}^{j-1} a_t \boldsymbol{a}_t = \boldsymbol{a}_i + \boldsymbol{a}_j$.

Let $\beta \in J(i, j)$, and let $\beta(t) = \beta(\boldsymbol{a}_t)$. Since $g^{\beta}_{\boldsymbol{a}_j \boldsymbol{a}_i} \neq 0$, there is an exact sequence

$$0 \to M(\boldsymbol{a}_i) \to \bigoplus_{t=1}^{m} \beta(t) M(\boldsymbol{a}_t) \to M(\boldsymbol{a}_j) \to 0.$$

Let us write $f = (f_t)_t$ with $f_t : M(\boldsymbol{a}_i) \to \beta(t)M(\boldsymbol{a}_t)$. The exact sequence does not split; otherwise, $\beta = \boldsymbol{a}_i \oplus \boldsymbol{a}_j$, which contradicts the assumption that $\beta \in J(i, j)$. Let us consider some t with $\beta(t) > 0$. We claim that $f_t \neq 0$. Otherwise, the cokernel of f would split off $\beta(t)$ copies of $M(\boldsymbol{a}_t)$; and since the cokernel of f is indecomposable, this would mean that the exact sequence splits. Since $\operatorname{Hom}(\boldsymbol{a}_i, \boldsymbol{a}_j) \neq 0$, it follows that $i \leq t$. In addition, we can exclude the case $i = t$, since in this case, f_t and therefore f would be a split monomorphism. Altogether, we have $i < t$. The dual argument applied to g shows that $t < j$. According to Proposition 2.2, we know that $\langle\beta\rangle$ is a scalar multiple of $X_{i+1}^{a_{i+1}} \cdots X_j^{a_j}$. The exact sequence exhibited above shows that $\sum_{t=i+1}^{j-1} a_t \boldsymbol{a}_t = \boldsymbol{a}_i + \boldsymbol{a}_j$. □

Now we define some algebra automorphisms and skew derivations on $H_{r,s}(\Lambda)$. Namely, for any $d \in \mathbb{Z}^n$, there is an algebra automorphism l_d of $H_{r,s}(\Lambda)$ defined by $l_d(w) = r^{\langle \dim w, d\rangle} s^{-\langle d, \dim w\rangle} w$, where w is any homogeneous element of $H_{r,s}(\Lambda)$.

Lemma 2.3 [Ringel 1996]. *Let R be a ring and let l be an endomorphism of R. For any $r \in R$, we define a map $\delta_r : R \to R$ by*

$$\delta_r(x) = rx - l(x)r \quad \text{for any } x \in R.$$

Then the map δ_r is an l-derivation.

Proof from [Ringel 1996]. First, the map δ_r is additive. In addition, for any $x, y \in R$, we have

$$\begin{aligned} \delta_r(xy) = rxy - l(xy)r &= rxy - l(x)ry + l(x)ry - l(x)l(y)r \\ &= (rx - l(x)r)y + l(x)(ry - l(y)r) \\ &= \delta_r(x)y + l(x)\delta_r(y). \end{aligned}$$

Thus the map δ_r is an l-derivation of R. □

Definition 2.1. Let R be a domain with $1 \neq 0$, let $\sigma_1 : R \to R$ be a ring homomorphism and let $\delta_1 : R \to R$ be a σ_1-derivation, so that for all $a, b \in R$, we have

- $\sigma_1(a + b) = \sigma_1(a) + \sigma_1(b)$,
- $\sigma_1(ab) = \sigma_1(a)\sigma_1(b)$,
- $\delta_1(a + b) = \delta_1(a) + \delta_1(b)$,
- $\delta_1(ab) = \delta_1(a)b + \sigma_1(a)\delta_1(b)$.

Then the *skew polynomial ring* $R[X_1, \sigma_1, \delta_1]$ is the set of noncommutative polynomials $R[X_1]$ with addition defined as commutative polynomials, and with multiplication defined distributively over addition and by the commutator rule

$$X_1 a = \sigma_1(a) X_1 + \delta_1(a),$$

valid for all $a \in R$. We set $R_1 = R[X_1, \sigma_1, \delta_1]$, and let σ_2 be a ring homomorphism of R_1 and δ_2 be a σ_2-derivation of the ring R_1. Then we can define another skew polynomial ring $R_2 = R_1[X_2, \sigma_2, \delta_2]$. Similarly, we can iterate this process to define R_n for any $n \geq 2$. These rings R_n are called *iterated skew polynomial rings.*

Let H_j denote the $\mathbb{Q}(r, s)$-subalgebra of $H_{r,s}(\Lambda)$ generated by the generators $X_1, \dots, X_j$. Thus we have $H_0 = \mathbb{Q}(r, s)$ and for any $0 \leq j \leq m$, we have

$$H_j = H_{j-1}[X_j, l_j, \delta_j]$$

with the automorphism l_j and the l_j-derivation δ_j of H_{j-1}. The automorphism l_j can be explicitly defined by

$$l_j(X_i) = r^{\langle \dim X_i, \dim X_j \rangle} s^{-\langle \dim X_j, \dim X_i \rangle} X_i \quad \text{for } i < j.$$

The skew derivation δ_j can be defined by

$$\delta_j(X_i) = X_j X_i - l_j(X_i) X_j = \sum_{I(i,j)} c(a_{i+1}, \dots, a_{j-1}) X_{i+1}^{a_{i+1}} \cdots X_{j_1}^{a_{j-1}}.$$

Theorem 2.4. *The automorphism l_j and the skew derivation δ_j satisfy the relation*

$$l_j \delta_j = r^{\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} s^{-\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} \delta_j l_j.$$

Proof. Suppose $i < j$. Then $l_j(X_i) = r^{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle} s^{-\langle \boldsymbol{a}_j, \boldsymbol{a}_i \rangle} X_i$. Thus, we have

$$\delta_j l_j(X_i) = r^{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle} s^{-\langle \boldsymbol{a}_j, \boldsymbol{a}_i \rangle} \delta_j(X_i).$$

Let us write $\boldsymbol{d} = \boldsymbol{a}_i + \boldsymbol{a}_j$. Note that $\delta_j(X_i)$ is a linear combination of monomials of the form

$$X_{i+1}^{a_{i+1}} \cdots X_{j-1}^{a_{j-1}} \quad \text{where } \textstyle\sum_{t=i+1}^{j-1} a_t \boldsymbol{a}_t = \boldsymbol{a}_i + \boldsymbol{a}_j = \boldsymbol{d}.$$

Thus we know that $\delta_j(X_i)$ belongs to $H_{r,s}(\Lambda)$. Since we have

$$\langle \boldsymbol{d}, \boldsymbol{a}_j \rangle = \langle \boldsymbol{a}_i + \boldsymbol{a}_j, \boldsymbol{a}_j \rangle = \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle + \langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle,$$
$$\langle \boldsymbol{a}_j, \boldsymbol{d} \rangle = \langle \boldsymbol{a}_j, \boldsymbol{a}_i + \boldsymbol{a}_j \rangle = \langle \boldsymbol{a}_j, \boldsymbol{a}_i \rangle + \langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle,$$

it follows that

$$\begin{aligned} l_j \delta_j(X_i) &= r^{\langle \boldsymbol{d}, \boldsymbol{a}_j \rangle} s^{-\langle \boldsymbol{d}, \boldsymbol{a}_j \rangle} \delta_j(X_i) \\ &= r^{\langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle + \langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} s^{-\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle - \langle a_j, \boldsymbol{a}_j \rangle} \delta_j(X_i) \\ &= r^{\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} s^{-\langle \boldsymbol{a}_j, \boldsymbol{a}_j \rangle} \delta_j l_j(X_i). \end{aligned}$$ □

Theorem 2.5. *The two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$ can be presented as an iterated skew polynomial ring.*

Proof. The proof follows from the previous theorem. □

Let R be a ring. Recall that an ideal $P \subset R$ is said to be prime if $P \neq R$ and if whenever the product AB of two ideals A and B of R is contained in P, at least one of A and B is contained in P. An ideal $P \subset R$ is called completely prime if $P \neq R$ and if whenever the product ab of two elements of R is contained in P, at least one of the elements a and b is contained in P. In the case of commutative rings, prime ideals are exactly completely prime ideals. In the case of noncommutative rings, a completely prime ideal is a prime ideal, but a prime ideal is not necessarily a completely prime ideal. Concerning prime ideals, we have the following result for the algebra $H_{r,s}(\Lambda)$.

Corollary 2.1. *Suppose the multiplicative group generated by r and s is torsion-free. Then any prime ideal of $H_{r,s}(\Lambda)$ is completely prime.*

Proof. The proof follows directly from a result on prime ideals of iterated skew polynomial rings, due to Goodearl and Letzter [2000]. □

2.4. ***An algebra isomorphism from $U^+_{r,s}(\mathfrak{g})$ onto $H_{r,s}(\Lambda)$.*** In this subsection, we introduce an algebra isomorphism from the two-parameter quantized enveloping algebra $U^+_{r,s}(\mathfrak{g})$ onto the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$. Via this isomorphism, all results established in the previous subsection on $H_{r,s}(\Lambda)$ will be transferred to the two-parameter quantized enveloping algebra $U^+_{r,s}(\mathfrak{g})$. For the convenience of this paper, we will awkwardly denote the nontwisted Hall algebra multiplication by $\circ$ in the one-parameter nontwisted generic Ringel–Hall algebra $H_v(\Lambda)$ (which can be defined due to the existence of Hall polynomials). Recall also that $v^2 = q$.

First, we need an important lemma on a two-parameter version of the quantum Serre relations, which was proved to hold in the case of one-parameter nontwisted Ringel–Hall algebra $H_v(\Lambda)$.

Lemma 2.4. *Let $\alpha_i \in \mathcal{P}$ correspond to the simple module S_i. Then we have the identities*

$$\sum_{k=0}^{1-a_{ij}} (-1)^k \binom{1-a_{ij}}{k}_{r_i s_i^{-1}} c_{ij}^{(k)} u_{\alpha_i}^{1-a_{ij}-k} u_{\alpha_j} u_{\alpha_i}^k = 0 \quad \textit{for } i \neq j$$

in $H_{r,s}(\Lambda)$, where $c_{ij}^{(k)} = (r_i s_i^{-1})^{k(k-1)/2} r^{k\langle j,i\rangle} s^{-k\langle i,j\rangle}$ for $i \neq j$.

Proof. The idea of the proof is to reduce these identities to those that have been proved in [Ringel 1990b] to hold for the one-parameter nontwisted generic Ringel–Hall algebra $H_v(\Lambda)$. Though this reduction is straightforward, we will provide the details. For convenience, we shall set $m = 1 - a_{ij}$ in the rest of this proof.

First, we assume that $i < j$. Then we have $\langle i, j\rangle = d_i a_{ij}$ and $\langle j, i\rangle = 0$. Therefore,

$$\begin{aligned}
&\sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} c_{ij}^{(k)} u_{\alpha_i}^{m-k} u_{\alpha_j} u_{\alpha_i}^{k}\\
&= \sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} (r_i s_i^{-1})^{k(k-1)/2} r^{k\langle j,i\rangle} s^{-k\langle i,j\rangle} u_{\alpha_i}^{m-k} u_{\alpha_j} u_{\alpha_i}^{k}\\
&= \sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} (r_i s_i^{-1})^{k(k-1)/2} r^{k\langle j,i\rangle} s^{-k\langle i,j\rangle} s^{-(m(m-1)/2\langle i,i\rangle+(m-k)\langle i,j\rangle+k\langle j,i\rangle)}\\
&\qquad\qquad u_{\alpha_i}^{\circ(m-k)} \circ u_{\alpha_j} \circ u_{\alpha_i}^{\circ(k)}\\
&= s^{-(\langle i,i\rangle m^2-m/2+m\langle i,j\rangle)} \sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} (r_i s_i^{-1})^{k(k-1)/2} u_{\alpha_i}^{\circ(m-k)} \circ u_{\alpha_j} \circ u_{\alpha_i}^{\circ(k)}.
\end{aligned}$$

Note that the following result was proved for the nontwisted generic Ringel–Hall algebra $H_v(\Lambda)$ in [Ringel 1990b]:

$$\sum_{k=0}^{m}(-1)^k \binom{m}{k}_{q^i} (q^i)^{k(k-1)/2} u_{\alpha_i}^{\circ(m-k)} \circ u_{\alpha_j} \circ u_{\alpha_i}^{\circ(k)} = 0.$$

Due to the existence of Hall polynomials [Ringel 1990a; 1990c], we can set $rs^{-1} = q$. Thus we have proved that the statement is true for $i < j$, as desired.

Now let us assume that $i > j$. Then $\langle i, j\rangle = 0$ and $\langle j, i\rangle = d_j a_{ji} = d_i a_{ij} = d_i(1-m)$. Furthermore, we have

$$\begin{aligned}
&(-1)^m \sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} c_{ij}^{(k)} u_{\alpha_i}^{m-k} u_{\alpha_j} u_{\alpha_i}^{k}\\
&= \sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} (r_i s_i^{-1})^{(m-k)(m-k-1)/2} r^{(m-k)\langle j,i\rangle} s^{-(m-k)\langle i,j\rangle} u_{\alpha_i}^{k} u_{\alpha_j} u_{\alpha_i}^{m-k}\\
&= \sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} (r_i s_i^{-1})^{(m-k)(m-k-1)/2} r^{(m-k)\langle j,i\rangle}\\
&\qquad\qquad s^{-(\frac{1}{2}m(m-1)\langle i,i\rangle+k\langle i,j\rangle+(m-k)\langle j,i\rangle)} u_{\alpha_i}^{\circ(k)} \circ u_{\alpha_j} \circ u_{\alpha_i}^{\circ(m-k)}\\
&= r^{\frac{1}{2}(m-m^2)\langle i,i\rangle} \sum_{k=0}^{m}(-1)^k \binom{m}{k}_{r_i s_i^{-1}} (r_i s_i^{-1})^{k(k-1)/2} u_{\alpha_i}^{\circ(k)} \circ u_{\alpha_j} \circ u_{\alpha_i}^{\circ(m-k)}.
\end{aligned}$$

The following result was proved for the nontwisted generic Ringel–Hall algebra $H_v(\Lambda)$ in [Ringel 1990b]:

$$\sum_{k=0}^{m}(-1)^k \binom{m}{k}_{q^i} (r^i)^{k(k-1)/2} u_{\alpha_i}^{\circ(k)} \circ u_{\alpha_j} \circ u_{\alpha_i}^{\circ(m-k)} = 0.$$

Once again, thanks to the existence of Hall polynomials, we can set $rs^{-1} = q$ and thus the result follows as desired. □

The next result was first proved by Reineke for the case of a finite-dimensional complex simple Lie algebra $\mathfrak{g}$ of type A, D or E in [Reineke 2001], and we show that the result holds for all finite-dimensional complex simple Lie algebras $\mathfrak{g}$. The proof is more or less the same as the one used in [Reineke 2001].

Theorem 2.6 (see also [Reineke 2001]). *With the multiplication defined above, $H_{r,s}(\Lambda)$ is an associative $\mathbb{Q}(r, s)$-algebra. In particular, the map*

$$\eta : e_i \to \alpha_i$$

extends to a $\mathbb{Q}(r, s)$-algebra isomorphism

$$\eta : U^+_{r,s}(\mathfrak{g}) \to H_{r,s}(\Lambda).$$

Proof. First, note that the quantum Serre relations of $U^+_{r,s}(\mathfrak{g})$ are preserved by the map η. Thus the map η indeed defines an algebra homomorphism from the two-parameter quantized enveloping algebra $U^+_{r,s}(\mathfrak{g})$ into the two-parameter Ringel–Hall algebra $H_{r,s}(\Lambda)$. It only remains to show that the map η is a bijection.

We first show that the map η is surjective by verifying that the algebra $H_{r,s}(\Lambda)$ is generated by the elements u_i that correspond to the simple modules S_i of the algebra Λ. Let u_α be any element in $H_{r,s}(\Lambda)$. Then we have

$$u_\alpha = \left(\prod_{i=1}^{m} \frac{1}{[\alpha(i)]^{!}_{\epsilon(\boldsymbol{a}_i)}}\right) u_{\boldsymbol{a}_1}^{\alpha(\boldsymbol{a}_1)} \cdots u_{\boldsymbol{a}_m}^{\alpha(\boldsymbol{a}_m)}.$$

Now it suffices to prove that u_α is generated by u_i for any α corresponding to an indecomposable module. We prove this claim by using induction. Note that $\zeta(\alpha, \alpha) = 0$. Thus, we have

$$u_\alpha = u_1^{d_1} \cdots u_n^{d_n} - \sum_{\substack{\beta \neq \alpha \\ \dim \beta = \dim \alpha}} s^{\langle \beta, \beta \rangle} u_\beta.$$

However, one sees that the dimension of the module u_β is less than that of the module u_α. Thus by induction on the dimension, we can reduce to the case where $\dim(u_\alpha) = 1$. In this case, the only possibility is that $u_\alpha = u_i$ for some i. Thus we have proved the statement that every u_α is generated by u_i, which further implies that the map η is a surjective map. We also note that the map η is a graded map.

Finally, we show that the map η is injective. Let $B := \mathbb{Q}[r, s]_{(r-1),s-1)}$ denote the localization of the polynomial ring $\mathbb{Q}[r, s]$ at the maximal ideal $(r - 1, s - 1)$. Then we know that $B = \mathbb{Q}[r, s]_{(r-1,s-1)}$ is a local ring with residue field $\mathbb{Q}$ and fractional field $\mathbb{Q}(r, s)$. Let U^+_B denote the free B-algebra generated by the generators e_i

subject to the quantum Serre relations holding in $U^+_{r,s}(\mathfrak{g})$. Also let $U^+_{\mathbb{Q}}(\mathfrak{g})$ denote the universal enveloping algebra of $\mathfrak{n}^+$ defined over the base field $\mathbb{Q}$. Then we have

$$U^+_{r,s}(\mathfrak{g}) = \mathbb{Q}(r,s)\otimes_B U^+_B \quad\text{and}\quad U^+_{\mathbb{Q}}(\mathfrak{g}) = \mathbb{Q}\otimes_B U^+_B.$$

For any $\beta \in \mathbb{N}^n$, we have the following result via Nakayama's lemma:

$$\begin{aligned}\dim_{\mathbb{Q}} U^+_{\mathbb{Q}}(\mathfrak{g})_\beta &= \dim_{\mathbb{Q}}(\mathbb{Q}\otimes_B U^+_B))_\beta\\ &\ge \dim_{\mathbb{Q}(r,s)}(\mathbb{Q}(r,s)\otimes_B U^+_B)_\beta = \dim_{\mathbb{Q}(r,s)} U^+_{r,s}(\mathfrak{g})_\beta\\ &\ge \dim_{\mathbb{Q}(r,s)} H_{r,s}(\Lambda)_\beta.\end{aligned}$$

Using [Ringel 1993, Corollary 2] and the PBW-theorem, we also have

$$\dim_{\mathbb{Q}} U^+_{\mathbb{Q}}(\mathfrak{g})_\beta = \dim_{\mathbb{Q}(r,s)} H_{r,s}(\Lambda)_\beta.$$

Thus we have proved that the map η is injective. Therefore, the map η is an algebra isomorphism from $U^+_{r,s}(\mathfrak{g})$ onto $H_{r,s}(\Lambda)$, as desired. □

Based on the previous theorem, the following corollary is in order:

Corollary 2.2. *The algebra $U^+_{r,s}(\mathfrak{g})$ has a $\mathbb{Q}(r,s)$-basis parameterized by the isomorphism classes of finite-dimensional representations of the algebra Λ.*

Theorem 2.7. *All prime ideals of $U^+_{r,s}(\mathfrak{g})$ are completely prime under the condition that the multiplicative group generated by r and s is torsion-free.*

Proof. This follows since $U_{r,s}(\mathfrak{g})$ is isomorphic to $H_{r,s}(\Lambda)$ as an algebra and since all prime ideals of $H_{r,s}(\Lambda)$ are completely prime under the condition. □

3. The extended two-parameter Ringel–Hall algebras $\overline{H_{r,s}(\Lambda)}$

In the one-parameter quantum group case, the torus part was added to the Ringel–Hall algebra for the purpose of realizing the Borel subalgebra $U^{\ge 0}_q(\mathfrak{g})$ of the one-parameter quantum group $U_q(\mathfrak{g})$. In the two-parameter case, we can do the same. Here, we spell out the details. In particular, we will first define the extended Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$ by adding the torus part. Then we propose a Hopf algebra structure on this extended two-parameter Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$. As an application, we will prove that $U^{\ge 0}_{r,s}(\mathfrak{g})$ is isomorphic to the extended two-parameter Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$ as a Hopf algebra. The approach used here is very similar to those in [Ringel 1996; Green 1995; Xiao 1997]. In addition, an analogous result is obtained for the algebra $U^{\le 0}_{r,s}(\mathfrak{g})$. By patching them together via the triangular decomposition of $U_{r,s}(\mathfrak{g})$, we derive a PBW-basis of $U_{r,s}(\mathfrak{g})$.

3.1. ***Extended Ringel–Hall algebras*** $\overline{\boldsymbol{H_{r,s}}(\boldsymbol{\Lambda})}$**.** To realize the Borel subalgebras $U_{r,s}^{\geq 0}(\mathfrak{g})$ and $U_{r,s}^{\leq 0}(\mathfrak{g})$ of the two-parameter quantum group $U_{r,s}(\mathfrak{g})$, we need to enlarge the Ringel–Hall algebras $H_{r,s}(\Lambda)$ defined in the previous section. We are going to enlarge $H_{r,s}(\Lambda)$ by adding the torus part to it. Namely, we will define $\overline{H_{r,s}(\Lambda)}$ to be the free $\mathbb{Q}(r,s)$-module with the basis

$$\{k_\alpha u_\lambda \mid \alpha \in \mathbb{Z}[I],\ \lambda \in \mathcal{P}\}.$$

In addition, we are going to define an algebra structure for $\overline{H_{r,s}(\Lambda)}$ by

$$\begin{aligned} u_\alpha u_\beta &= \textstyle\sum_{\lambda\in\mathcal{P}} s^{-\langle\alpha,\beta\rangle} F^{u_\lambda}_{u_\alpha,u_\beta}(rs^{-1})u_\lambda && \text{for any } \alpha,\beta\in\mathcal{P},\\ k_\alpha u_\beta &= r^{\langle\beta,\alpha\rangle}s^{-\langle\alpha,\beta\rangle}u_\beta k_\alpha && \text{for any } \alpha\in\mathbb{Z}[I],\ \beta\in\mathcal{P},\\ k_\alpha k_\beta &= k_\beta k_\alpha && \text{for any } \alpha,\beta\in\mathbb{Z}[I]. \end{aligned}$$

Lemma 3.1. *For any elements $x, y, z \in \mathbb{Z}[I]$ and $\alpha, \beta, \gamma \in \mathcal{P}$, we have*

$$[(k_x u_\alpha)(k_y u_\beta)](k_x u_\alpha) = (k_x u_\alpha)[(k_y u_\beta)(k_z u_\gamma)].$$

In particular, the multiplication defined in $\overline{H_{r,s}(\Lambda)}$ is associative.

Proof. First, we have

$$\begin{aligned} (u_\alpha u_\beta)u_\gamma &= \Big(\sum_{\lambda\in\mathcal{P}} s^{-\langle\alpha,\beta\rangle} g^{\lambda}_{\alpha\beta}u_\lambda\Big)u_\gamma = \sum_{\lambda'\in\mathcal{P}} s^{-\langle\alpha,\beta\rangle-\langle\alpha+\beta,\gamma\rangle} g^{\lambda'}_{\alpha\beta\gamma}u_{\lambda'},\\ u_\alpha(u_\beta u_\gamma) &= u_\alpha\Big(\sum_{\lambda\in\mathcal{P}} s^{-\langle\beta,\gamma\rangle} g^{\lambda}_{\beta\gamma}u_\lambda\Big) = \sum_{\lambda'\in\mathcal{P}} s^{-\langle\alpha,\beta+\gamma\rangle-\langle\beta,\gamma\rangle} g^{\lambda'}_{\alpha\beta\gamma}u_{\lambda'}. \end{aligned}$$

So we have just proved that $(u_\alpha u_\beta)u_\gamma = u_\alpha(u_\beta u_\gamma)$. In addition, we have the results

$$\begin{aligned} [(k_x u_\alpha)(k_y u_\beta)](k_z u_\gamma) &= r^{-\langle\alpha,y\rangle-\langle\alpha+\beta,z\rangle} s^{\langle y,\alpha\rangle+\langle z,\alpha+\beta\rangle} k_{x+y+x}u_\alpha u_\beta u_\gamma,\\ (k_x u_\alpha)[(k_y u_\beta)(k_z u_\gamma)] &= r^{-\langle\alpha,y+z\rangle-\langle\beta,z\rangle} s^{\langle y+z,\alpha\rangle+\langle z,\beta\rangle} k_{x+y+z}u_\alpha u_\beta u_\gamma. \end{aligned}$$

Also, we have $(k_x u_\alpha)[(k_y u_\beta)(k_z u_\gamma)] = [(k_x u_\alpha)(k_y u_\beta)](k_z u_\gamma)$, which further implies that the multiplication is associative. □

Proposition 3.1. *With the above defined multiplication, $\overline{H_{r,s}(\Lambda)}$ is an associative $\mathbb{Q}(r,s)$-algebra.*

Proof. This follows directly from the previous lemma. □

Theorem 3.1. *The map η extends to a $\mathbb{Q}(r,s)$-algebra isomorphism from $U_{r,s}^{\geq 0}(\mathfrak{g})$ onto $\overline{H_{r,s}(\Lambda)}$ via the map $\eta(\omega_i) = k_i$ and $\eta(e_i) = u_{\alpha_i}$.*

Proof. The proof is straightforward. □

Corollary 3.1. *The set $\boldsymbol{B}^+ = \{\omega_\alpha \eta^{-1}(u_\lambda) \mid \alpha \in \mathbb{Z}[I], \lambda \in \mathcal{P}\}$ is a $\mathbb{Q}(r,s)$-basis of $U_{r,s}^{\geq 0}(\mathfrak{g})$.*

3.2. ***A Hopf algebra structure on $\overline{H_{r,s}(\Lambda)}$.*** Now we are going to introduce a Hopf algebra structure on the extended two-parameter Ringel–Hall algebra $\overline{H_{r,s}(\Lambda)}$.

Theorem 3.2. *The algebra $\overline{H_{r,s}(\Lambda)}$ is a Hopf algebra with the Hopf algebra structure defined as follows.*

(1) *Multiplication*:

$$\begin{aligned} u_\alpha u_\beta &= \textstyle\sum_{\lambda\in\mathcal{P}} s^{-\langle\alpha,\beta\rangle} g^\lambda_{\alpha\beta} u_\lambda && \textit{for any } \alpha,\beta\in\mathcal{B},\\ k_\alpha u_\beta &= r^{\langle\beta,\alpha\rangle} s^{-\langle\alpha,\beta\rangle} u_\beta k_\alpha && \textit{for any } \alpha\in\mathbb{Z}[I],\ \beta\in\mathcal{P},\\ k_\alpha k_\beta &= k_\beta k_\alpha && \textit{for any } \alpha,\beta\in\mathbb{Z}[I]. \end{aligned}$$

(2) *Comultiplication*:

$$\begin{aligned} \Delta(u_\lambda) &= \textstyle\sum_{\alpha,\beta\in\mathcal{P}} r^{\langle\alpha,\beta\rangle}(a_\alpha a_\beta/a_\lambda) g^\lambda_{\alpha\beta} u_\alpha k_\beta\otimes u_\beta && \textit{for any } \lambda\in\mathcal{P},\\ \Delta(k_\alpha) &= k_\alpha\otimes k_\alpha && \textit{for any } \alpha\in\mathbb{Z}[I]. \end{aligned}$$

(3) *Counit*: $\epsilon(u_\lambda)=0$ *for all* $\lambda\neq 0$ *and* $\epsilon(k_\alpha)=1$ *for any* $\alpha\in\mathcal{P}$.

(4) *Antipode*:

$$\begin{aligned} \sigma(u_\lambda) = \delta_{\lambda,0} + \sum_{m\geq 1}(-1)^m \\ \times \sum_{\pi\in\mathcal{P},\lambda_1,\lambda_2,\dots,\lambda_m\in\mathcal{P}_1} (rs^{-1})^{\sum_{i<j}\langle\lambda_i,\lambda_j\rangle} \frac{a_{\lambda_1}\cdots a_{\lambda_m}}{a_\lambda} g^\lambda_{\lambda_1\cdots\lambda_m} g^\pi_{\lambda_1\cdots\lambda_m} k_{-\lambda} u_\pi \end{aligned}$$

for any element $\lambda\in\mathcal{P}$ *and* $\sigma(k_\alpha)=k_{-\alpha}$ *for any* $\alpha\in\mathbb{Z}[I]$.

The proof of this theorem consists of two lemmas.

Lemma 3.2. *The comultiplication Δ is an algebra endomorphism of $\overline{H_{r,s}(\Lambda)}$.*

Proof. First, $\Delta(k_xk_y)=\Delta(k_{x+y})=k_{x+y}\otimes k_{x+y}$. Thus, $\Delta(k_xk_y)=\Delta(k_x)\Delta(k_y)$. To prove that Δ is an algebra homomorphism of $H_{r,s}(\Lambda)$, it suffices to show that $\Delta(u_\alpha u_\beta)=\Delta(u_\alpha)\Delta(u_\beta)$. Since

$$u_{\alpha'}u_{\beta'} = \sum_{\lambda\in\mathcal{P}} s^{-\langle\alpha',\beta'\rangle} g^\lambda_{\alpha'\beta'} u_\lambda,$$

we have

$$\begin{aligned} \Delta(u_{\alpha'}u_{\beta'}) &= \Delta\Big(\sum_{\lambda\in\mathcal{P}} s^{-\langle\alpha,\beta\rangle} g^\lambda_{\alpha'\beta'} u_\lambda\Big)\\ &= \sum_{\lambda\in\mathcal{P}} s^{-\langle\alpha',\beta'\rangle} g^\lambda_{\alpha'\beta'} \Delta(u_\lambda)\\ &= \sum_{\lambda,\alpha,\beta\in\mathcal{P}} s^{-\langle\alpha',\beta'\rangle} r^{\langle\alpha,\beta\rangle} g^\lambda_{\alpha'\beta'} g^\lambda_{\alpha\beta} \frac{a_\alpha a_\beta}{a_\lambda} u_\alpha k_\beta\otimes u_\beta \end{aligned}$$

and

$$\begin{aligned}\Delta(u_{\alpha'})\Delta(u_{\beta'}) &= \Big(\sum_{\rho,\sigma\in\mathcal{P}} r^{\langle\rho,\sigma\rangle}\frac{a_\rho a_\sigma}{a_{\alpha'}} g^{\alpha'}_{\rho\sigma} u_\rho k_\sigma \otimes u_\sigma\Big)\\ &\qquad\qquad \times\Big(\sum_{\rho',\sigma'\in\mathcal{P}} r^{\langle\rho',\sigma'\rangle}\frac{a_{\rho'}a_{a_{\sigma'}}}{a_{\beta'}} g^{\beta'}_{\rho'\sigma'} u_{\rho'} k_{\sigma'} \otimes u_{\sigma'}\Big)\\ &= \sum_{\rho,\sigma,\rho',\sigma'\in\mathcal{P}} r^{\langle\rho,\sigma\rangle+\langle\rho',\sigma'\rangle}\frac{a_\rho a_\sigma a_{\rho'} a_{\sigma'}}{a_{\alpha'}a_{\beta'}} \times g^{\alpha'}_{\rho\sigma} g^{\beta'}_{\rho'\sigma'} u_\rho k_\sigma u_{\rho'} k_{\sigma'} \otimes u_\sigma u_{\sigma'}\\ &= \sum_{\rho,\sigma,\rho',\sigma'\in\mathcal{P}} r^{\langle\rho,\sigma\rangle+\langle\rho',\sigma'\rangle+\langle\rho',\sigma\rangle} s^{-\langle\sigma,\rho'\rangle-\langle\sigma,\sigma'\rangle-\langle\rho,\rho'\rangle}\frac{a_\rho a_\sigma a_{\rho'} a_\sigma}{a_{\alpha'}a_{\beta'}}\\ &\qquad\qquad \times g^{\alpha'}_{\rho\sigma} g^{\beta'}_{\rho'\sigma'} g^{\alpha}_{\rho\rho'} g^{\beta}_{\sigma\sigma'} u_\alpha k_\beta \otimes u_\beta.\end{aligned}$$

Note that $\dim(u_\alpha)+\dim(u_\beta)=\dim(u_\lambda)$. Thus, we have

$$\dim(u_\alpha)+\dim(u_\beta)=\dim(u_\lambda)=\dim(u_{\alpha'})+\dim(u_{\beta'})$$

and

$$\begin{aligned}&\dim(u_\rho)+\dim(u_\sigma)=\dim(u_{\alpha'}), \quad \dim(u_{\rho'})+\dim(u_{\sigma'})=\dim(u_{\beta'});\\ &\dim(u_\rho)+\dim(u_{\rho'})=\dim(u_\alpha), \quad\ \dim(u_\sigma)+\dim(u_{\sigma'})=\dim(u_\beta).\end{aligned}$$

In addition, we have $k_\beta = k_\sigma k_{\sigma'}$. Thus we have

$$\begin{aligned}\langle\alpha,\beta\rangle &= \langle\rho,\sigma\rangle+\langle\rho,\sigma'\rangle+\langle\rho',\sigma\rangle+\langle\sigma,\sigma'\rangle,\\ \langle\alpha',\beta'\rangle &= \langle\rho,\rho'\rangle+\langle\rho,\sigma'\rangle+\langle\sigma,\rho'\rangle+\langle\sigma,\sigma'\rangle.\end{aligned}$$

Therefore, we only need to show that

$$\sum_{\lambda\in\mathcal{P}} g^{\lambda}_{\alpha\beta} g^{\lambda}_{\alpha'\beta'}\frac{a_\alpha a_\beta a_{\alpha'} a_{\beta'}}{a_\lambda} = \sum_{\rho,\sigma,\rho',\sigma'\in\mathcal{P}} (rs^{-1})^{-\langle\rho,\sigma'\rangle} g^{\sigma}_{\rho\rho'} g^{\beta}_{\sigma\sigma'} g^{\alpha'}_{\rho\sigma} g^{\beta'}_{\rho'\sigma'} a_\rho a_\sigma a_{\rho'} a_{\sigma'},$$

but this is true according to Green's formula. □

Lemma 3.3. *For any $\lambda\in\mathcal{P}$, we have*

$$\mu(\sigma\otimes 1)\Delta(u_\lambda)=\delta_{\lambda 0} \quad \textit{and} \quad \mu(1\otimes\sigma)\Delta(u_\lambda)=\delta_{\lambda 0}.$$

Proof. First of all, we have

$$\Delta(u_\lambda)=\sum_{\lambda',\lambda_{m+1}\in\mathcal{P}} \frac{a_{\lambda'}a_{\lambda_{m+1}}}{a_\lambda} g^{\lambda}_{\lambda'\lambda_{m+1}} u_{\lambda'} k_{\lambda_{m+1}} \otimes u_{\lambda_{m+1}}.$$

Thus we further have

$$\mu(\sigma\otimes 1)\Delta(u_\lambda)=\sum_{\lambda',\lambda_{m+1}\in\mathcal{P}} r^{\langle\lambda',\lambda_{m+1}\rangle}\frac{a_{\lambda'}a_{\lambda_{m+1}}}{a_\lambda} g^{\lambda}_{\lambda'\lambda_{m+1}} k_{-\lambda_{m+1}}\sigma(u_{\lambda'})u_{\lambda_{m+1}}.$$

To prove the first identity, it suffices to prove that

$$\sigma(u_\lambda) = \delta_{\lambda 0} - \sum_{\lambda' \in \mathcal{P}, \lambda_{m+1} \in \mathcal{P}_1} r^{\langle \lambda', \lambda_{m+1} \rangle} \frac{a_{\lambda'} a_{\lambda_{m+1}}}{a_\lambda} g^{\lambda}_{\lambda' \lambda_{m+1}} k_{-\lambda_{m+1}} \sigma(u_{\lambda'}) u_{\lambda_{m+1}}.$$

Since we have

$$\sigma(u_{\lambda'}) = \delta_{\lambda' 0} - \sum_{m \geq 1} (-1)^m \sum_{\pi' \in \mathcal{P}, \lambda_1, \ldots, \lambda_m \in \mathcal{P}_1} (rs^{-1})^{\sum_{i<j} \langle \lambda_i, \lambda_j \rangle} \frac{a_{\lambda_1} \cdots a_{\lambda_m}}{a_{\lambda'}} \times g^{\lambda'}_{\lambda_1 \cdots \lambda_m} g^{\pi'}_{\lambda_1 \cdots \lambda_m} k_{-\lambda'} u_{\pi'},$$

we will have the result

$$\begin{aligned}\sigma(u_\lambda) &= \delta_{\lambda 0} - k_{-\lambda} u_\lambda - \sum_{\lambda', \lambda_{m+1} \in \mathcal{P}_1} r^{\langle \lambda', \lambda_{m+1} \rangle} \frac{a_{\lambda'} a_{\lambda_{m+1}}}{a_\lambda} g^{\lambda}_{\lambda' \lambda_{m+1}} k_{-\lambda_{m+1}} \\ &\quad \times \sum_{m \geq 1} (-1)^m \sum_{\substack{\pi' \in \mathcal{P}, \\ \lambda_1, \ldots, \lambda_m \in \mathcal{P}_1}} (rs^{-1})^{\sum_{i<j} \langle \lambda_i, \lambda_j \rangle} \frac{a_{\lambda_1} \cdots a_{\lambda_m}}{a_{\lambda'}} \\ &\quad g^{\lambda'}_{\lambda_1 \cdots \lambda_m} g^{\pi'}_{\lambda_1 \ldots, \lambda_m} \times k_{-\lambda'} u_{\pi'} u_{\lambda_{m+1}} \\ &= \delta_{\lambda 0} - k_{-\lambda} u_\lambda - \sum_{m \geq 1} (-1)^m \sum_{\substack{\pi' \in \mathcal{P} \\ \lambda', \lambda_1, \ldots, \lambda_m \in \mathcal{P}_1}} (rs^{-1})^{\sum_{k<j} \langle \lambda_i, \lambda_j \rangle} r^{\langle \lambda', \lambda_{m+1} \rangle} \\ &\quad \times \frac{a_{\lambda_1 \cdots a_{\lambda_m}} a_{\lambda_{m+1}}}{\lambda} g^{\lambda'}_{\lambda_1 \cdots \lambda_m} g^{\pi'}_{\lambda' \lambda} k_{\lambda_{m+1}} k_{-\lambda'} \sum_{\pi \in \mathcal{P}} s^{\langle \pi', \lambda_{m+1} \rangle} g^{\pi}_{\pi' \lambda_{m+1}} u_\pi \\ &= \delta_{\lambda 0} - \sum_{m \geq 1} (-1)^m \sum_{\substack{\pi \in \mathcal{P} \\ \lambda_1, \ldots, \lambda_m \in \mathcal{P}_1}} \frac{a_{\lambda_1} \cdots a_{\lambda_m} a_{\lambda_{m+1}}}{a_\lambda} g^{\lambda}_{\lambda_1 \cdots \lambda_m \lambda_{m+1}} k_{-\lambda} u_\pi.\end{aligned}$$

Since $g^{\lambda}_{\alpha_1 \cdots \alpha_i} \neq 0$ implies $\dim u_\lambda = \dim u_{\alpha_1} + \cdots \dim u_{\lambda_i}$, we may assume that

$$\begin{aligned}\dim(u_{\lambda'}) &= \dim(u_{\lambda_1}) + \cdots + \dim(u_{\lambda_m}), \quad &\dim(u_{\lambda'}) + \dim(u_{\lambda_{m+1}}) = \dim(u_\lambda), \\ \dim(u_{\pi'}) &= \dim(u_{\lambda_1}) + \cdots + \dim(u_{\lambda_m}), \quad &\dim(u_{\pi'}) + \dim(u_{\lambda_{m+1}}) = \dim(u_\pi).\end{aligned}$$

Therefore, we have the result

$$\begin{aligned}\sigma(u_\lambda) &= \delta_{\lambda 0} - k_{-\lambda} u_\lambda + \sum_{m \geq 2} (-1)^m \\ &\quad \sum_{\substack{\pi \in \mathcal{P} \\ \lambda_1, \ldots, \lambda_m \in \mathcal{P}_1}} (rs^{-1})^{\sum_{i<j} \langle \lambda_i, \lambda_j \rangle} \frac{a_{\lambda_1} \cdots a_{\lambda_m}}{a_\lambda} g^{\lambda}_{\lambda_1 \cdots \lambda_m} g^{\pi}_{\lambda_1 \cdots \lambda_m} k_{-\lambda} u_\pi \\ &= \delta_{\lambda 0} + \sum_{m \geq 1} (-1)^m \sum_{\substack{\pi \in \mathcal{P} \\ \lambda_1, \ldots, \lambda_m \in \mathcal{P}_1}} (rs^{-1})^{\sum_{i<j} \langle \lambda_i, \lambda_j \rangle} \frac{a_{\lambda_1} \cdots a_{\lambda m}}{a_\lambda} \\ &\quad \times g^{\lambda}_{\lambda_1 \cdots \lambda_m} g^{\pi}_{\lambda_1 \cdots \lambda_m} k_{-\lambda} u_\pi.\end{aligned}$$

So we have proved the statement by the definition of σ. Similarly, we can verify that $\mu(1 \otimes \sigma)\Delta(u_\lambda) = \delta_{\lambda 0}$. □

Remark 3.1. The proofs of the two lemmas above are slightly modified versions of those in [Xiao 1997].

3.3. ***A Hopf algebra isomorphism from $U_{r,s}^{\geq 0}(\mathfrak{g})$ onto $\overline{H_{r,s}(\Lambda)}$.*** Here we prove that the Borel subalgebras $U_{r,s}^{\geq 0}(\mathfrak{g})$ and $U_{r,s}^{\leq 0}(\mathfrak{g})$ of the two-parameter quantum group $U_{r,s}(\mathfrak{g})$ can be realized as the extended two-parameter Ringel–Hall algebras $\overline{H_{r,s}(\Lambda)}$ and $\overline{H_{s^{-1},r^{-1}}(\Lambda)}$ as Hopf algebras. As a result, we shall derive a PBW-basis for the algebra $U_{r,s}(\mathfrak{g})$.

Theorem 3.3. *We have*

$$U_{r,s}^{\geq 0}(\mathfrak{g}) \cong \overline{H_{r,s}(\Lambda)} \quad \textit{and} \quad U_{r,s}^{\leq 0}(\mathfrak{g}) \cong \overline{H_{s^{-1},r^{-1}}(\Lambda)}$$

as Hopf algebras.

Proof. Let us define a map $\phi : U_{r,s}^{\geq 0}(\mathfrak{g}) \to \overline{H_{r,s}(\Lambda)}$ by setting $\phi(E_i) = u_{S_i}$ and $\phi(\omega_i) = k_i$. Then it is easy to verify that ϕ is a bijection and respects the Hopf algebra structures. Thus it is a Hopf algebra isomorphism. Similarly, we can prove that $U_{r,s}^{\leq 0}(\mathfrak{g})$ is isomorphic to $\overline{H_{s^{-1},r^{-1}}(\Lambda)}$ as a Hopf algebra. □

Let $\boldsymbol{B}^-$ be the basis constructed for $U_{r,s}^{\leq 0}(\mathfrak{g})$ via the algebra $\overline{H_{s^{-1},r^{-1}}(\Lambda)}$; then we have the following:

Corollary 3.2. *The set $\boldsymbol{B}^+ \cup \boldsymbol{B}^-$ is a $\mathbb{Q}(r, s)$-basis for the two-parameter quantum groups $U_{r,s}(\mathfrak{g})$.*

Acknowledgment

I would like to thank the referee for various comments and suggestions, which have dramatically helped to improve the presentation of the material in this paper.

References

[Artin et al. 1991] M. Artin, W. Schelter, and J. Tate, "Quantum deformations of GL_n", *Comm. Pure Appl. Math.* **44**:8-9 (1991), 879–895. MR 92i:17014 Zbl 0753.17015

[Benkart and Witherspoon 2004a] G. Benkart and S. Witherspoon, "Restricted two-parameter quantum groups", pp. 293–318 in *Representations of finite dimensional algebras and related topics in Lie theory and geometry* (Toronto, 2002), edited by V. Dlab and C. M. Ringel, Fields Inst. Commun. **40**, Amer. Math. Soc., Providence, RI, 2004. MR 2005b:17027 Zbl 1048.16020

[Benkart and Witherspoon 2004b] G. Benkart and S. Witherspoon, "Two-parameter quantum groups and Drinfel'd doubles", *Algebr. Represent. Theory* **7**:3 (2004), 261–286. MR 2005g:17028 Zbl 1113.16041

[Benkart et al. 2006] G. Benkart, S.-J. Kang, and K.-H. Lee, "On the centre of two-parameter quantum groups", *Proc. Roy. Soc. Edinburgh Sect. A* **136**:3 (2006), 445–472. MR 2007a:17019 Zbl 1106.17013

[Bergeron et al. 2006] N. Bergeron, Y. Gao, and N. Hu, "Drinfel'd doubles and Lusztig's symmetries of two-parameter quantum groups", *J. Algebra* **301**:1 (2006), 378–405. MR 2007e:17010 Zbl 1148.17007

[Bernšteĭn et al. 1973] I. N. Bernšteĭn, I. M. Gel'fand, and V. A. Ponomarev, "Coxeter functors, and Gabriel's theorem", *Uspehi Mat. Nauk* **28**:2(170) (1973), 19–33. In Russian. MR 52 #13876 Zbl 0269.08001

[Chin and Musson 1996] W. Chin and I. M. Musson, "Multiparameter quantum enveloping algebras", *J. Pure Appl. Algebra* **107**:2-3 (1996), 171–191. MR 97c:17016 Zbl 0859.17004

[Dlab and Ringel 1975] V. Dlab and C. M. Ringel, "On algebras of finite representation type", *J. Algebra* **33** (1975), 306–394. MR 50 #9974 Zbl 0332.16014

[Dlab and Ringel 1976] V. Dlab and C. M. Ringel, *Indecomposable representations of graphs and algebras*, Mem. Amer. Math. Soc. **173**, American Mathematical Society, Providence, RI, 1976. MR 56 #5657 Zbl 0332.16015

[Dobrev and Parashar 1993] V. K. Dobrev and P. Parashar, "Duality for multiparametric quantum GL(n)", *J. Phys. A* **26**:23 (1993), 6991–7002. MR 95d:81053 Zbl 0821.17009

[Doi and Takeuchi 1994] Y. Doi and M. Takeuchi, "Multiplication alteration by two-cocycles — the quantum version", *Comm. Algebra* **22**:14 (1994), 5715–5732. MR 95j:16043 Zbl 0821.16038

[Gabriel 1972] P. Gabriel, "Unzerlegbare Darstellungen, I", *Manuscripta Math.* **6** (1972), 71–103. Corrected in **6** (1972), 309. MR 48 #11212 Zbl 0232.08001

[Gabriel 1973] P. Gabriel, "Indecomposable representations, II", pp. 81–104 in *Convegno di Algebra Commutativa, INDAM* (Rome, 1971), Symposia Mathematica **11**, Academic Press, London, 1973. MR 49 #5132 Zbl 0276.16001

[Goodearl and Letzter 2000] K. R. Goodearl and E. S. Letzter, "The Dixmier–Moeglin equivalence in quantum coordinate rings and quantized Weyl algebras", *Trans. Amer. Math. Soc.* **352**:3 (2000), 1381–1403. MR 2000j:16040 Zbl 0978.16040

[Green 1995] J. A. Green, "Hall algebras, hereditary algebras and quantum groups", *Invent. Math.* **120**:2 (1995), 361–377. MR 96c:16016 Zbl 0836.16021

[Hu and Pei 2008] N. Hu and Y. Pei, "Notes on 2-parameter quantum groups, I", *Sci. China Ser. A* **51**:6 (2008), 1101–1110. MR 2009c:17016 Zbl 1145.81381

[Hu et al. 2008] N. Hu, M. Rosso, and H. Zhang, "Two-parameter quantum affine algebra $U_{r,s}(\widehat{\mathfrak{sl}_n})$, Drinfel'd realization and quantum affine Lyndon basis", *Comm. Math. Phys.* **278**:2 (2008), 453–486. MR 2009b:17033

[Jing 1992] N. H. Jing, "Quantum groups with two parameters", pp. 129–138 in *Deformation theory and quantum groups with applications to mathematical physics* (Amherst, MA, 1990), edited by M. Gerstenhaber and J. Stasheff, Contemp. Math. **134**, Amer. Math. Soc., Providence, RI, 1992. MR 94b:17022 Zbl 0774.17015

[Kharchenko 1999] V. K. Kharchenko, "A quantum analogue of the Poincaré–Birkhoff–Witt theorem", *Algebra Log.* **38**:4 (1999), 476–507, 509. In Russian; translated in *Algebra and Logic* **38**:4 (1999), 259–276. MR 2001f:16075

[Kharchenko 2002] V. K. Kharchenko, "A combinatorial approach to the quantification of Lie algebras", *Pacific J. Math.* **203**:1 (2002), 191–233. MR 2003b:17018 Zbl 1069.17008

[Kulish 1990] P. P. Kulish, "A two-parameter quantum group and a gauge transformation", *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov.* (*LOMI*) **180**:Voprosy Kvant. Teor. Polya i Statist. Fiz. 9 (1990), 89–93, 180. In Russian; translated in *J. Math. Sci., New York* **68**:2 (1994), 220–222. MR 91j:81029 Zbl 0792.17011

[Lusztig 1990] G. Lusztig, "Canonical bases arising from quantized enveloping algebras", *J. Amer. Math. Soc.* **3**:2 (1990), 447–498. MR 90m:17023 Zbl 0703.17008

[Lusztig 1993] G. Lusztig, *Introduction to quantum groups*, Progress in Mathematics **110**, Birkhäuser, Boston, MA, 1993. MR 94m:17016 Zbl 0788.17010

[Reineke 2001] M. Reineke, "Generic extensions and multiplicative bases of quantum groups at $q = 0$", *Represent. Theory* **5** (2001), 147–163. MR 2002c:17029 Zbl 1050.17015

[Reshetikhin 1990] N. Reshetikhin, "Multiparameter quantum groups and twisted quasitriangular Hopf algebras", *Lett. Math. Phys.* **20**:4 (1990), 331–335. MR 91k:17012 Zbl 0719.17006

[Ringel 1976] C. M. Ringel, "Representations of K-species and bimodules", *J. Algebra* **41**:2 (1976), 269–302. MR 54 #10340 Zbl 0338.16011

[Ringel 1990a] C. M. Ringel, "Hall algebras", pp. 433–447 in *Topics in algebra* (Warsaw, 1988), vol. 1, edited by S. Balcerzyk et al., Banach Center Publ. **26**, PWN, Warsaw, 1990. MR 93f:16027 Zbl 0778.16004

[Ringel 1990b] C. M. Ringel, "Hall algebras and quantum groups", *Invent. Math.* **101**:3 (1990), 583–591. MR 91i:16024 Zbl 0735.16009

[Ringel 1990c] C. M. Ringel, "Hall polynomials for the representation-finite hereditary algebras", *Adv. Math.* **84**:2 (1990), 137–178. MR 92e:16010 Zbl 0799.16013

[Ringel 1993] C. M. Ringel, "Hall algebras revisited", pp. 171–176 in *Quantum deformations of algebras and their representations* (Rehovot, 1991/1992), edited by A. Joseph and S. Shnider, Israel Math. Conf. Proc. **7**, Bar-Ilan Univ., Ramat Gan, 1993. MR 94k:16021 Zbl 0852.17009

[Ringel 1996] C. M. Ringel, "PBW-bases of quantum groups", *J. Reine Angew. Math.* **470** (1996), 51–88. MR 97d:17009 Zbl 0840.17010

[Sudbery 1990] A. Sudbery, "Consistent multiparameter quantisation of GL(n)", *J. Phys. A* **23**:15 (1990), L697–L704. MR 91m:17022 Zbl 0722.17007

[Takeuchi 1990] M. Takeuchi, "A two-parameter quantization of GL(n) (summary)", *Proc. Japan Acad. Ser. A Math. Sci.* **66**:5 (1990), 112–114. MR 92f:16049

[Xiao 1997] J. Xiao, "Drinfeld double and Ringel–Green theory of Hall algebras", *J. Algebra* **190**:1 (1997), 100–144. MR 98a:16018 Zbl 0874.16026

Received June 23, 2009. Revised January 5, 2010.

XIN TANG
DEPARTMENT OF MATHEMATICS & COMPUTER SCIENCE
FAYETTEVILLE STATE UNIVERSITY
1200 MURCHISON ROAD
FAYETTEVILLE, NC 28301

xtang@uncfsu.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 247, No. 1, 2010

A NEW PROBABILITY DISTRIBUTION WITH APPLICATIONS

MINGJIN WANG

We introduce a new probability distribution, which is useful in the study of basic hypergeometric series. As applications, we give probabilistic derivations of the q-binomial theorem, the q-Gauss summation formula, a new multiple identity, and an extension of the Rogers–Ramanujan identities.

1. Introduction

The probabilistic method is a useful tool in the study of basic hypergeometric series [Chapman 2005; Evans 2002; Fulman 2001; Rawlings 1997]. In this paper, we introduce a new probability distribution and then demonstrate the applications of this distribution in q-series. We begin with recall some definitions, notations and known results in [Andrews et al. 1999; Gasper and Rahman 1990; Liu 2003]. Throughout the paper, we suppose that $0 < q < 1$. The q-shifted factorials are defined as

$$(a;q)_0 = 1, \quad (a;q)_n = \prod_{k=0}^{n-1}(1-aq^k), \quad (a;q)_\infty = \prod_{k=0}^{\infty}(1-aq^k).$$

We also adopt a compact notation for multiple q-shifted factorials:

$$(a_1, a_2, \ldots, a_m; q)_n = (a_1;q)_n(a_2;q)_n \cdots (a_m;q)_n,$$

where n is an integer or ∞. The q-binomial coefficient is defined by

$$\begin{bmatrix} n \\ k \end{bmatrix} = \frac{(q;q)_n}{(q;q)_k(q;q)_{n-k}}.$$

In 1846, Heine introduced the ${}_{r+1}\phi_r$ basic hypergeometric series, which is defined by

$${}_{r+1}\phi_r\left(\begin{matrix} a_1, a_2, \ldots, a_{r+1} \\ b_1, b_2, \ldots, b_r \end{matrix}; q, x\right) = \sum_{n=0}^{\infty} \frac{(a_1, a_2, \ldots, a_{r+1}; q)_n x^n}{(q, b_1, b_2, \ldots, b_r; q)_n}.$$

MSC2000: primary 60E05; secondary 33D15, 05A10.
Keywords: probability distribution, basic hypergeometric series, Andrews–Askey integral, Al-Salam–Carlitz polynomials, Lebesgue's dominated convergence theorem, Tannery's theorem, Rogers–Ramanujan identities.

F. H. Jackson [1910] defined the q-integral by

$$\int_0^d f(t)d_qt = d(1-q)\sum_{n=0}^{\infty} f(dq^n)q^n, \tag{1-1}$$

and

$$\int_c^d f(t)d_qt = \int_0^d f(t)d_qt - \int_0^c f(t)d_qt. \tag{1-2}$$

The q-integrals are important in the theory and application of basic hypergeometric series. For example, the author gives some applications of the q-integral in [Wang 2008; 2009b; 2009a; 2010b; 2010a]. The Andrews–Askey [1981] integral is

$$\int_c^d \frac{(qt/c, qt/d; q)_\infty}{(at, bt; q)_\infty} d_qt = \frac{d(1-q)(q, dq/c, c/d, abcd; q)_\infty}{(ac, ad, bc, bd; q)_\infty}, \tag{1-3}$$

which can be derived from Ramanujan's ${}_1\psi_1$ summation provided that no zero factors occur in the denominator of the integral.

The Al-Salam–Carlitz polynomials $\varphi_n^{(a)}(x\,|\,q)$ are defined by

$$\varphi_n^{(a)}(x\,|\,q) = \sum_{k=0}^{n} \begin{bmatrix} n \\ k \end{bmatrix} x^k (a;q)_k,$$

[Srivastava and Jain 1989] and have the q-integral representation [Wang 2009b]

$$\varphi_n^{(a)}(x\,|\,q) = \frac{(ax, a; q)_\infty}{(1-q)(q, q/x, x; q)_\infty} \int_x^1 \frac{(qt/x, qt; q)_\infty t^n}{(at; q)_\infty} d_qt \tag{1-4}$$

provided that no zero factors occur in the denominator.

We frequently use the following well-known theorems:

Theorem (analytic continuation theorem). *If f and g are analytic at z_0 and agree at infinitely many points, which include z_0 as an accumulation point, then $f = g$.*

Theorem (Lebesgue's dominated convergence theorem). *Suppose that $\{X_n, n \geq 1\}$ is a sequence of random variables such that $X_n \to X$ pointwise almost everywhere as $n \to \infty$, and such that $|X_n| \leq Y$ for all n, where the random variable Y is integrable. Then X is integrable, and*

$$\lim_{n\to\infty} EX_n = EX,$$

where $E(\cdot)$ denotes expected value.

Tannery's theorem is a special case of Lebesgue's dominated convergence theorem on the sequence space L^1.

Theorem [Tannery 1904]. *If* $s(n) = \sum_{k\geq 0} f_k(n)$ *is a finite sum* (*or a convergent series*) *for each* n,

$$\lim_{n\to\infty} f_k(n) = f_k, \quad |f_k(n)| \leq M_k, \quad \textit{and} \quad \sum_{k=0}^{\infty} M_k < \infty$$

then

$$\lim_{n\to\infty} s(n) = \sum_{k=0}^{\infty} f_k.$$

2. A new probability distribution

In order to use Lebesgue's dominated convergence theorem to get q-identities, we need to find some special probability distributions. In this section, we introduce a useful probability distribution.

The main method of this paper as follows: First, we define a probability distribution by q-shifted factorials; its expected value can be easily obtained. Then we construct a sequence of random variables with this probability distribution. Finally, we use Lebesgue's dominated convergence theorem to obtain a q-identity.

Lemma 2.1. *Suppose* x *is a real such that* $x < 0$; *then we have*

$$\frac{(-x)^n (x^{n-1}q^{k+1}, x^n q^{k+1}; q)_\infty q^k}{(q, q/x, x; q)_\infty} \geq 0 \tag{2-1}$$

and

$$\sum_{n=0}^{1}\sum_{k=0}^{\infty} \frac{(-x)^n (x^{n-1}q^{k+1}, x^n q^{k+1}; q)_\infty q^k}{(q, q/x, x; q)_\infty} = 1, \tag{2-2}$$

where $n = 0, 1$ *and* $k = 0, 1, 2, \ldots$.

Proof. Inequality (2-1) is obvious by the definition of the q-shifted factorials and the assumption that $x < 0$. We only need to prove (2-2).

Since

$$\begin{aligned} &\sum_{n=0}^{1}\sum_{k=0}^{\infty} \frac{(-x)^n (x^{n-1}q^{k+1}, x^n q^{k+1}; q)_\infty q^k}{(q, q/x, x; q)_\infty} \\ &= \frac{1}{(1-q)(q, q/x, x; q)_\infty} \\ &\qquad \times \biggl((1-q)\sum_{k=0}^{\infty}(q^{k+1}/x, q^{k+1}; q)_\infty q^k - x(1-q)\sum_{k=0}^{\infty}(q^{k+1}, xq^{k+1}; q)_\infty q^k\biggr), \end{aligned} \tag{2-3}$$

using the definition of the q-integral gives

$$(1-q)\sum_{k=0}^{\infty}(q^{k+1}/x, q^{k+1}; q)_\infty q^k = \int_0^1 (qt/x, qt; q)_\infty d_q t$$

and

$$x(1-q)\sum_{k=0}^{\infty}(q^{k+1}, xq^{k+1}; q)_\infty q^k = \int_0^x (qt/x, qt; q)_\infty d_q t.$$

Consequently, we have

$$(1-q)\sum_{k=0}^{\infty}(q^{k+1}/x, q^{k+1}; q)_\infty q^k \tag{2-4}$$
$$= \int_0^1 (qt/x, qt; q)_\infty d_q t - \int_0^x (qt/x, qt; q)_\infty d_q t = \int_x^1 (qt/x, qt; q)_\infty d_q t.$$

Employing the Andrews–Askey integral (1-3) gives

$$\int_x^1 (qt/x, qt; q)_\infty d_q t = (1-q)(q, q/x, x; q)_\infty. \tag{2-5}$$

Substituting (2-4) and (2-5) into (2-3) gives (2-2). □

Definition 2.2. A random variable ξ has distribution $W(x; q)$ if

$$P(\xi = x^n q^k) = \frac{(-x)^n (x^{n-1}q^{k+1}, x^n q^{k+1}; q)_\infty q^k}{(q, q/x, x; q)_\infty},$$

where $x < 0$, $0 < q < 1$, $n = 0, 1$ and $k = 0, 1, 2, \dots$.

The distribution $W(x; q)$ has some applications in the study of basic hypergeometric series.

Before giving applications, we need the following lemmas.

Lemma 2.3. *Let $-1 < x < 0$ and $|a| < 1$. Let ξ denote a random variable having with $W(x; q)$. Then we have*

$$E\left(\frac{\xi^m}{(a\xi; q)_\infty}\right) = \frac{1}{(a, ax; q)_\infty}\, \varphi_m^{(a)}(x\,|\,q) \quad \textit{for } m = 0, 1, 2, \dots. \tag{2-6}$$

Proof. Using the definition of the q-integral (1-1), (1-2) and the q-integral representation of the Al-Salam–Carlitz polynomials (1-4), we have

$$
\begin{aligned}
&E\left(\frac{\xi^m}{(a\xi;q)_\infty}\right)\\
&\quad=\sum_{n=0}^{1}\sum_{k=0}^{\infty}\frac{(-x)^n(x^{n-1}q^{k+1},x^nq^{k+1};q)_\infty q^k}{(q,q/x,x;q)_\infty}\cdot\frac{x^{nm}q^{km}}{(ax^nq^k;q)_\infty}\\
&\quad=\frac{1}{(1-q)(q,q/x,x;q)_\infty}\left((1-q)\sum_{k=0}^{\infty}(q^{k+1}/x,q^{k+1};q)_\infty\cdot\frac{q^{k(m+1)}}{(aq^k;q)_\infty}\right.\\
&\qquad\qquad\left.-x(1-q)\sum_{k=0}^{\infty}(q^{k+1},xq^{k+1};q)_\infty\cdot\frac{x^mq^{k(m+1)}}{(axq^k;q)_\infty}\right)\\
&\quad=\frac{1}{(1-q)(q,q/x,x;q)_\infty}\\
&\qquad\qquad\times\left(\int_0^1\frac{(qt/x,qt;q)_\infty t^m}{(at;q)_\infty}d_qt-\int_0^x\frac{(qt/x,qt;q)_\infty t^m}{(at;q)_\infty}d_qt\right)\\
&\quad=\frac{1}{(1-q)(q,q/x,x;q)_\infty}\int_x^1\frac{(qt/x,qt;q)_\infty t^m}{(at;q)_\infty}d_qt\\
&\quad=\frac{1}{(a,ax;q)_\infty}\,\varphi_m^{(a)}(x\,|\,q).\qquad\square
\end{aligned}
$$

Lemma 2.4. *Let $-1<x<0$ and $|a|<1$. Let ξ denote a random variable having distribution $W(x;q)$. Then we have*

$$E\left(\frac{1}{(a\xi,b\xi;q)_\infty}\right)=\frac{(abx,;q)_\infty}{(a,b,ax,bx;q)_\infty}.$$

Proof. Using the definition of the q-integral (1-1), (1-2) and the Andrews–Askey integral (1-3), we have

$$
\begin{aligned}
&E\left(\frac{1}{(a\xi,b\xi;q)_\infty}\right)\\
&\quad=\sum_{n=0}^{1}\sum_{k=0}^{\infty}\frac{(-x)^n(x^{n-1}q^{k+1},x^nq^{k+1};q)_\infty q^k}{(q,q/x,x;q)_\infty}\cdot\frac{1}{(ax^nq^k,bx^nq^k;q)_\infty}\\
&\quad=\frac{1}{(1-q)(q,q/x,x;q)_\infty}\left((1-q)\sum_{k=0}^{\infty}(q^{k+1}/x,q^{k+1};q)_\infty\cdot\frac{q^k}{(aq^k,bq^k;q)_\infty}\right.\\
&\qquad\qquad\left.-x(1-q)\sum_{k=0}^{\infty}(q^{k+1},xq^{k+1};q)_\infty\cdot\frac{q^k}{(axq^k,bx^nq^k;q)_\infty}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{(1-q)(q,q/x,x;q)_\infty}\left(\int_0^1 \frac{(qt/x,qt;q)_\infty}{(at,bt;q)_\infty}d_qt - \int_0^x \frac{(qt/x,qt;q)_\infty}{(at,bt;q)_\infty}d_qt\right)\\
&= \frac{1}{(1-q)(q,q/x,x;q)_\infty}\int_x^1 \frac{(qt/x,qt;q)_\infty}{(at,bt;q)_\infty}d_qt = \frac{(abx,;q)_\infty}{(a,b,ax,bx;q)_\infty},
\end{aligned}
$$

which completes the proof. □

Lemma 2.5. *Let* $|x| < 1$. *Then*

$$
\lim_{n\to\infty} \varphi_n^{(a)}(x\,|\,q) = \sum_{k=0}^{\infty} \frac{(a;q)_k x^k}{(q;q)_k}. \tag{2-7}
$$

Proof. Let $f_k(n) = {n \brack k} x^k (a;q)_k$ if $k \le n$ and $f_k(n) = 0$ if $k > n$. We have

$$
\varphi_n^{(a)}(x\,|\,q) = \sum_{k=0}^{\infty} f_k(n).
$$

Since

$$
\lim_{n\to\infty} f_k(n) = \frac{(a;q)_k x^k}{(q;q)_k}, \quad |f_k(n)| \le \frac{|(a;q)_k x^k|}{(q;q)_\infty}, \quad \sum_{k=0}^{\infty} \frac{|(a;q)_k x^k|}{(q;q)_\infty} < \infty,
$$

by Tannery's theorem we know (2-7) holds. □

3. The q-binomial theorem

One of the most important summation formulas for basic hypergeometric series is the q-binomial theorem, which was derived by Cauchy in 1843, Heine in 1847, and by other mathematicians. There are many proofs. By using the probability distribution $W(x;q)$ and the Lebesgue dominated convergence theorem, we give a probabilistic derivation; see also [Andrews et al. 1999; Gasper and Rahman 1990].

Theorem 3.1. $\displaystyle\sum_{n=0}^{\infty} \frac{(a;q)_n x^n}{(q;q)_n} = \frac{(ax;q)_\infty}{(x;q)_\infty}$ *for* $|x| < 1$.

Proof. Let ξ be a random variable having distribution $W(x;q)$, where $-1 < x < 0$. We consider the sequence

$$
\left\{ \frac{\xi^n}{(a\xi;q)_\infty} \right\}_{n=1}^{\infty} \quad \text{for } |a| < 1
$$

of random variables (on a probability space). It is easy to see that ξ^n converges to $I_{(\xi=1)}$, which has Binomial distribution $B(1, 1/(x;q)_\infty)$ and

$$
\lim_{n\to\infty} \frac{\xi^n}{(a\xi;q)_\infty} = \frac{I_{(\xi=1)}}{(a;q)_\infty},
$$

where I_Ω is the indicator function defined by

$$I_\Omega(x) = \begin{cases} 1 & \text{if } x \in \Omega, \\ 0 & \text{if } x \notin \Omega. \end{cases}$$

Since

$$\left| \frac{\xi^n}{(a\xi; q)_\infty} \right| \le \frac{1}{(|a|; q)_\infty},$$

using Lebesgue's dominated convergence theorem gives

$$\lim_{n\to\infty} E\left(\frac{\xi^n}{(a\xi; q)_\infty} \right) = E\left(\frac{I_{(\xi=1)}}{(a; q)_\infty} \right). \tag{3-1}$$

Employing (1-4) and using Tannery's theorem gives

$$\begin{aligned} \lim_{m\to\infty} E\left(\frac{\xi^m}{(a\xi; q)_\infty} \right) &= \frac{1}{(a, ax; q)_\infty} \lim_{m\to\infty} \varphi_m^{(a)}(x \,|\, q) \\ &= \frac{1}{(a, ax; q)_\infty} \sum_{m=0}^{\infty} \frac{(a; q)_m x^m}{(q; q)_m}. \end{aligned} \tag{3-2}$$

By direct calculation,

$$E\left(\frac{I_{(\xi=1)}}{(a; q)_\infty} \right) = \frac{1}{(a, x; q)_\infty}. \tag{3-3}$$

Substituting (3-2) and (3-3) into (3-1) gives

$$\sum_{n=0}^{\infty} \frac{(a; q)_n x^n}{(q; q)_n} = \frac{(ax; q)_\infty}{(x; q)_\infty},$$

where $-1 < x < 0$ and $|a| < 1$. By analytic continuation, we may replace the assumptions $-1 < x < 0$ by $|a| < 1$ by $|x| < 1$. Thus, we get Theorem 3.1. □

4. The q-Gauss summation formula

In 1847, Heine derived a q-analogue of Gauss's summation formula. We show that this result can be recovered with the probability distribution $W(x; q)$.

Theorem 4.1. $_2\phi_1\left(\begin{matrix} a, b \\ c \end{matrix} ; q, \frac{c}{ab} \right) = \frac{(c/a, c/b; q)_\infty}{(c, c/ab; q)_\infty}$ *for* $|c/(ab)| < 1$.

Proof. Let ξ and η denote two independent random variables having distributions $W(x; q)$ and $W(y; q)$, respectively, where we set $-1 < x, y < 0$. We consider the following sequence of random variables (on a probability space):

$$\left\{ \frac{\eta^n}{(a\xi\eta; q)_\infty} \right\}_{n=1}^{\infty} \quad \text{for } |a| < 1.$$

Clearly η^n converges to $I_{(\eta=1)}$ having binomial distribution $B(1, 1/((y;q)_\infty))$ and

$$\lim_{n\to\infty}\frac{\eta^n}{(a\xi\eta;q)_\infty}=\frac{I_{(\eta=1)}}{(a\xi;q)_\infty},$$

where I_Ω is the indicator function.

Since

$$\left|\frac{\eta^n}{(a\xi\eta;q)_\infty}\right|\le\frac{1}{(|a|;q)_\infty},$$

using Lebesgue's dominated convergence theorem gives

$$\lim_{n\to\infty}E\left(\frac{\eta^n}{(a\xi\eta;q)_\infty}\right)=E\left(\frac{I_{(\eta=1)}}{(a\xi;q)_\infty}\right). \tag{4-1}$$

Observe that

$$\begin{aligned}E\left(\frac{\eta^n}{(a\xi\eta;q)_\infty}\right)&=E\left(E\left(\frac{\eta^n}{(a\xi\eta;q)_\infty}|\xi\right)\right)\\&=E\left(\frac{1}{(a\xi,ay\xi;q)_\infty}\,\varphi_n^{(a\xi)}(x\,|\,q)\right)\\&=\sum_{k=0}^{n}\begin{bmatrix}n\\k\end{bmatrix}y^k\cdot E\left(\frac{1}{(a\xi q^k,ay\xi;q)_\infty}\right)\\&=\sum_{k=0}^{n}\begin{bmatrix}n\\k\end{bmatrix}y^k\cdot\frac{(a^2xyq^k;q)_\infty}{(aq^k,axq^k,ay,axy;q)_\infty}\\&=\frac{(a^2xy;q)_\infty}{(a,ax,ay,axy;q)_\infty}\sum_{k=0}^{n}\begin{bmatrix}n\\k\end{bmatrix}\cdot\frac{(a,ax;q)_\infty y^k}{(a^2xy;q)_\infty}.\end{aligned}$$

Hence, we get the left hand side of (4-1):

$$\lim_{n\to\infty}E\left(\frac{\eta^n}{(a\xi\eta;q)_\infty}\right)=\frac{(a^2xy;q)_\infty}{(a,ax,ay,axy;q)_\infty}\sum_{k=0}^{\infty}\frac{(a,ax;q)_\infty y^k}{(q,a^2xy;q)_\infty}. \tag{4-2}$$

On the other hand, the right hand side of (4-1) equals

$$E\left(\frac{I_{(\eta=1)}}{(a\xi;q)_\infty}\right)=p(\eta=1)E\left(\frac{1}{(a\xi;q)_\infty}\right)=\frac{1}{(a,ax,y;q)_\infty}. \tag{4-3}$$

Substituting (4-2) and (4-3) into (4-1) gives

$$\sum_{k=0}^{\infty}\frac{(a,ax;q)_\infty y^k}{(q,a^2xy;q)_\infty}=\frac{(ay,axy;q)_\infty}{(a^2xy,y;q)_\infty},$$

which is equivalent to the q-Gauss theorem, Theorem 4.1, by analytic continuation. □

5. A multiple identity

Multiple basic hypergeometric series have been investigated by various authors [Milne 1997; Wang 2009a; Zhang 2006; Zhang and Liu 2006]. We will use the distribution $W(x;q)$ to prove the following multiple identity.

Theorem 5.1. *Let* $|a|<1$. *Then for any positive integers* m *and* n, *we have*

$$\sum_{y_1+\cdots+y_m\geq n}\begin{bmatrix}y_1+\cdots+y_m\\ n\end{bmatrix}_q q^{y_2+2y_3+\cdots+(m-1)y_m}a^{y_1+\cdots+y_m} = \frac{a^n}{(a;q)_{n+m}}\begin{bmatrix}n+m-1\\ n\end{bmatrix}. \tag{5-1}$$

Proof. Let ξ denote a random variable with distribution $W(x;q)$, where $-1<x<0$. For any positive integer m, we consider the sequence

$$\left\{\frac{(1-(a\xi)^n)(1-(aq\xi)^n)\cdots(1-(aq^{m-1}\xi)^n)}{(a\xi;q)_\infty}\right\}_{n=1}^{\infty}\quad\text{for } |a|<1$$

of random variables (on a probability space). It is easy to see that

$$\lim_{n\to\infty}\frac{(1-(a\xi)^n)(1-(aq\xi)^n)\cdots(1-(aq^{m-1}\xi)^n)}{(a\xi;q)_\infty}=\frac{1}{(a\xi;q)_\infty}.$$

Since $|(1-(a\xi)^n)(1-(aq\xi)^n)\cdots(1-(aq^{m-1}\xi)^n)/(a\xi;q)_\infty|\leq 1/(|a|;q)_\infty$, using Lebesgue's dominated convergence theorem gives

$$\lim_{n\to\infty}E\left(\frac{(1-(a\xi)^n)(1-(aq\xi)^n)\cdots(1-(aq^{m-1}\xi)^n)}{(a\xi;q)_\infty}\right)=E\left(\frac{1}{(a\xi;q)_\infty}\right). \tag{5-2}$$

Employing (2-6), we get the right hand side of (5-2):

$$E\left(\frac{1}{(a\xi;q)_\infty}\right)=\frac{1}{(a,ax;q)_\infty}. \tag{5-3}$$

On the other hand, observing that

$$\begin{aligned}
&\frac{(1-(a\xi)^n)(1-(aq\xi)^n)\cdots(1-(aq^{m-1}\xi)^n)}{(a\xi;q)_\infty}\\
&\quad=\frac{1-(a\xi)^n}{1-a\xi}\cdot\frac{1-(aq\xi)^n}{1-aq\xi}\cdots\frac{1-(aq^{m-1}\xi)^n}{1-aq^{m-1}\xi}\cdot\frac{1}{(aq^m\xi;q)_\infty}\\
&\quad=\sum_{y_1=0}^{n-1}(a\xi)^{y_1}\cdot\sum_{y_2=0}^{n-1}(aq\xi)^{y_2}\cdots\sum_{y_m=0}^{n-1}(aq^{m-1}\xi)^{y_m}\cdot\frac{1}{(aq^m\xi;q)_\infty}\\
&\quad=\sum_{0\leq y_1,\ldots,y_m\leq n-1}q^{y_2+2y_3+\cdots+(m-1)y_m}a^{y_1+\cdots+y_m}\cdot\frac{\xi^{y_1+\cdots+y_m}}{(aq^m\xi;q)_\infty},
\end{aligned}$$

we have

$$E\left(\frac{[1-(a\xi)^n][1-(aq\xi)^n]\cdots[1-(aq^{m-1}\xi)^n]}{(a\xi;q)_\infty}\right)$$
$$=\sum_{0\le y_1,\dots,y_m\le n-1} q^{y_2+2y_3+\cdots+(m-1)y_m}a^{y_1+\cdots+y_m}E\left(\frac{\xi^{y_1+\cdots+y_m}}{(aq^m\xi;q)_\infty}\right)$$
$$=\frac{1}{(aq^m,axq^m;q)_\infty}\times\sum_{0\le y_1,\dots,y_m\le n-1} q^{y_2+2y_3+\cdots+(m-1)y_m}a^{y_1+\cdots+y_m}\,\varphi^{(aq^m)}_{y_1+\cdots+y_m}(x\,|\,q).$$

Hence, we get the left hand side of (5-2):

$$\lim_{n\to\infty}E\left(\frac{[1-(a\xi)^n][1-(aq\xi)^n]\cdots[1-(aq^{m-1}\xi)^n]}{(a\xi;q)_\infty}\right)=\frac{1}{(aq^m,axq^m;q)_\infty}\times\sum_{y_1,\dots,y_m\ge 0} q^{y_2+2y_3+\cdots+(m-1)y_m}a^{y_1+\cdots+y_m}\,\varphi^{(aq^m)}_{y_1+\cdots+y_m}(x\,|\,q).\tag{5-4}$$

Substituting (5-3) and (5-4) into (5-2) gives

$$\sum_{y_1,\dots,y_m\ge 0} q^{y_2+2y_3+\cdots+(m-1)y_m}a^{y_1+\cdots+y_m}\,\varphi^{(aq^m)}_{y_1+\cdots+y_m}(x\,|\,q)=\frac{1}{(a,ax;q)_m}.\tag{5-5}$$

Using Theorem 3.1 with $a=q^m$ and $x=ax$ gives

$$\sum_{k=0}^{\infty}\begin{bmatrix}m+k-1\\k\end{bmatrix}a^kx^k=\frac{1}{(ax;q)_m}.\tag{5-6}$$

Substituting (5-6) into (5-5) and comparing the coefficients of x^n gives (5-1). □

6. An extension of the Rogers–Ramanujan identities

The well-known Rogers–Ramanujan identities are

$$\sum_{m=0}^{\infty}\frac{q^{m^2}}{(q;q)_m}=\frac{1}{(q,q^4;q^5)_\infty},\tag{6-1}$$

$$\sum_{m=0}^{\infty}\frac{q^{m^2+m}}{(q;q)_m}=\frac{1}{(q^2,q^3;q^5)_\infty}.\tag{6-2}$$

There are many proofs of this beautiful pair of identities. Baxter's [1982] is based on the statistical mechanics, and the proof of Lepowsky and Milne [1978] uses the character formula on an infinite dimensional Lie algebra. We use our probability distribution to derive an extension of the Rogers–Ramanujan identities.

Theorem 6.1. *We have*

$$\sum_{m=n}^{\infty}\frac{q^{m^2}}{(q;q)_{m-n}}=\frac{1}{(q;q)_\infty}$$
$$\times\Bigg(q^n+\sum_{k=1}^{\infty}\sum_{l=0}^{n}\begin{bmatrix}n\\l\end{bmatrix}\frac{(-1)^k(q^k;q)_l(q;q)_{2k}}{(1-q^k)(q;q)_{2k+1-l}}\,q^{5\binom{k}{2}+k(n+2-l)}(1-q^{2k}+q^{4k+1})$$
$$-\sum_{k=1}^{\infty}\sum_{l=0}^{n}\begin{bmatrix}n\\l\end{bmatrix}\frac{(-1)^k(q^k;q)_l(q;q)_{2k}}{(1-q^k)(q;q)_{2k+1-l}}\,q^{5\binom{k}{2}+k(n+4-l)+1-l}\Bigg).$$

Proof. By Watson's q-Whipple transformation formula [Watson 1929],

$${}_8\phi_7\left(\begin{matrix}a,q\sqrt{a},-q\sqrt{a},b,c,d,e,q^{-n}\\ \sqrt{a},-\sqrt{a},qa/b,qa/c,qa/d,qa/e,q^{n+1}a\end{matrix};q,\frac{q^{2+n}a^2}{bcde}\right)$$
$$=\frac{(qa,qa/bc;q)_n}{(qa/b,qa/c)_n}\,{}_4\phi_3\left(\begin{matrix}q^{-n},b,c,qa/de\\ qa/d,qa/e,q^{-n}bc/a\end{matrix};q,q\right).$$

Letting $b,c,d,e,n\to\infty$ in this equation gives

$$\sum_{m=0}^{\infty}\frac{q^{m^2}a^m}{(q;q)_m}=\frac{1}{(aq;q)_\infty}+\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+2k}}{(q;q)_k}\cdot\frac{a^{2k}}{(aq^k;q)_\infty}$$
$$-\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+4k}}{(q;q)_k}\cdot\frac{a^{2k+1}}{(aq^k;q)_\infty}\quad\text{for } |a|\le 1.$$

Then letting $a=\xi$ gives

$$\sum_{m=0}^{\infty}\frac{q^{m^2}\xi^m}{(q;q)_m}=\frac{1}{(\xi q;q)_\infty}+\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+2k}}{(q;q)_k}\cdot\frac{\xi^{2k}}{(\xi q^k;q)_\infty}$$
$$-\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+4k}}{(q;q)_k}\cdot\frac{\xi^{2k+1}}{(\xi q^k;q)_\infty}.$$

where ξ is a random variable with distribution $W(x;q)$ and $-1<x<0$. Applying the expectation operator E to both sides of the above, we get

$$E\left(\sum_{m=0}^{\infty}\frac{q^{m^2}\xi^m}{(q;q)_m}\right)=E\left(\frac{1}{(\xi q;q)_\infty}\right)+E\left(\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+2k}}{(q;q)_k}\cdot\frac{\xi^{2k}}{(\xi q^k;q)_\infty}\right)\tag{6-3}$$
$$-E\left(\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+4k}}{(q;q)_k}\cdot\frac{\xi^{2k+1}}{(\xi q^k;q)_\infty}\right).$$

Since $|q^{m^2}\xi^m/(q;q)_m| \le q^{m^2}/(q;q)_m$ and the series $\sum_{m=0}^{\infty} q^{m^2}/(q;q)_m$ converges absolutely, using Lebesgue's dominated convergence theorem and (2-6) gives the left hand side of (6-3):

$$E\biggl(\sum_{m=0}^{\infty}\frac{q^{m^2}\xi^m}{(q;q)_m}\biggr)=\sum_{m=0}^{\infty}\frac{q^{m^2}E\{\xi^m\}}{(q;q)_m}=\sum_{m=0}^{\infty}\frac{q^{m^2}h_m(x\,|\,q)}{(q;q)_m}. \tag{6-4}$$

On the other hand, using (2-6) gives

$$E\biggl(\frac{1}{(\xi q;q)_\infty}\biggr)=\frac{1}{(q,qx;q)_\infty}, \tag{6-5}$$

$$E\biggl(\frac{\xi^{2k}}{(\xi q^k;q)_\infty}\biggr)=\frac{1}{(q^k,q^kx;q)_\infty}\,\varphi_{2k}^{(q^k)}(x\,|\,q), \tag{6-6}$$

$$E\biggl(\frac{\xi^{2k+1}}{(\xi q^k;q)_\infty}\biggr)=\frac{1}{(q^k,q^kx;q)_\infty}\,\varphi_{2k+1}^{(q^k)}(x\,|\,q). \tag{6-7}$$

It is easy to see that

$$\biggl|\frac{(-1)^kq^{5\binom{k}{2}+4k}}{(q;q)_k}\cdot\frac{\xi^{2k+1}}{(\xi q^k;q)_\infty}\biggr|\le\biggl|\frac{(-1)^kq^{5\binom{k}{2}+2k}}{(q;q)_k}\cdot\frac{\xi^{2k}}{(\xi q^k;q)_\infty}\biggr|\le\frac{q^{5\binom{k}{2}+2k}}{(q;q)_k(q;q)_\infty},$$

and the series $\sum_{k=0}^{\infty} q^{5\binom{k}{2}+2k}/((q;q)_k(q;q)_\infty)$ is converges absolutely. Using Lebesgue's dominated convergence theorem and (6-5), (6-6) and (6-7) gives the right hand side of (6-3):

$$\begin{aligned}
&E\biggl(\frac{1}{(\xi q;q)_\infty}\biggr)+E\biggl(\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+2k}}{(q;q)_k}\cdot\frac{\xi^{2k}}{(\xi q^k;q)_\infty}\biggr)\\
&\qquad\qquad -E\biggl(\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+4k}}{(q;q)_k}\cdot\frac{\xi^{2k+1}}{(\xi q^k;q)_\infty}\biggr)\\
&=E\biggl(\frac{1}{(\xi q;q)_\infty}\biggr)+\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+2k}}{(q;q)_k}E\biggl(\frac{\xi^{2k}}{(\xi q^k;q)_\infty}\biggr)\\
&\qquad\qquad -\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+4k}}{(q;q)_k}E\biggl(\frac{\xi^{2k+1}}{(\xi q^k;q)_\infty}\biggr)\\
&=\frac{1}{(q,qx;q)_\infty}+\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+2k}}{(q;q)_k}\frac{1}{(q^k,q^kx;q)_\infty}\,\varphi_{2k}^{(q^k)}(x\,|\,q)\\
&\qquad\qquad -\sum_{k=1}^{\infty}\frac{(-1)^kq^{5\binom{k}{2}+4k}}{(q;q)_k}\frac{1}{(q^k,q^kx;q)_\infty}\,\varphi_{2k+1}^{(q^k)}(x\,|\,q)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{(q, qx; q)_\infty} \\
&\qquad + \frac{1}{(q, x; q)_\infty} \sum_{k=1}^{\infty} \frac{(-1)^k (x; q)_k}{1-q^k}\, q^{5\binom{k}{2}+2k} \big(\varphi_{2k}^{(q^k)}(x \,|\, q) - q^{2k} \varphi_{2k+1}^{(q^k)}(x \,|\, q)\big).
\end{aligned}
$$

Substituting this and (6-4) into (6-3) gives

$$
\begin{aligned}
\sum_{m=0}^{\infty} \frac{q^{m^2} h_m(x \,|\, q)}{(q; q)_m} &= \frac{1}{(q, qx; q)_\infty} \\
&\qquad + \frac{1}{(q, x; q)_\infty} \sum_{k=1}^{\infty} \frac{(-1)^k (x; q)_k}{1-q^k}\, q^{5\binom{k}{2}+2k} \big(\varphi_{2k}^{(q^k)}(x \,|\, q) - q^{2k} \varphi_{2k+1}^{(q^k)}(x \,|\, q)\big),
\end{aligned}
$$

where $-1 < x < 0$. By analytic continuation, we may replace the assumption $-1 < x < 0$ by $|x| < 1$.

Substituting the expansion

$$
\frac{1}{(z; q)_\infty} = \sum_{l=0}^{\infty} \frac{z^l}{(q; q)_l}
$$

into the last, we have

$$
\begin{aligned}
\sum_{m=0}^{\infty} \frac{q^{m^2} h_m(x \,|\, q)}{(q; q)_m} &= \frac{1}{(q; q)_\infty} \sum_{l=0}^{\infty} \frac{q^l x^l}{(q; q)_l} \\
&\quad + \frac{1}{(q; q)_\infty} \sum_{k=1}^{\infty} \sum_{l=0}^{\infty} \bigg(\frac{q^{kl} x^l}{(q; q)_l} \cdot \frac{(-1)^k q^{5\binom{k}{2}+2k}}{1-q^k} \big(\varphi_{2k}^{(q^k)}(x \,|\, q) - q^{2k} \varphi_{2k+1}^{(q^k)}(x \,|\, q)\big) \bigg).
\end{aligned}
$$

Comparing the coefficients of x^n in this identity gives

$$
\begin{aligned}
\sum_{m=n}^{\infty} \frac{q^{m^2} \big[{m \atop n}\big]}{(q; q)_m} &= \frac{q^n}{(q; q)_\infty (q; q)_n} \\
&\quad + \frac{1}{(q; q)_\infty} \sum_{k=1}^{\infty} \sum_{l=0}^{n} \frac{(-1)^k (q^k; q)_l}{(1-q^k)(q; q)_{n-l}}\, q^{5\binom{k}{2}+k(n+2-l)} \bigg(\begin{bmatrix} 2k \\ l \end{bmatrix} + q^{2k} \begin{bmatrix} 2k+1 \\ l \end{bmatrix} \bigg),
\end{aligned}
$$

which can be written as Theorem 6.1. □

The Rogers–Ramanujan identities are special cases of Theorem 6.1. Letting $n = 0$ and then applying the Jacobi triple product identity [Andrews et al. 1999]

$$
\sum_{n=-\infty}^{\infty} (-1)^n q^{\binom{n}{2}} x^n = (q, x, q/x; q)_\infty
$$

leads to the Rogers–Ramanujan identity (6-1). In fact, when $n = 0$, we have

$$\begin{aligned}\sum_{m=0}^{\infty}\frac{q^{m^2}}{(q;q)_m} &= \frac{1}{(q;q)_\infty}\left(1+\sum_{k=1}^{\infty}(-1)^k(1+q^k)q^{5\binom{k}{2}+2k}\right)\\ &= \frac{1}{(q;q)_\infty}\sum_{k=-\infty}^{\infty}(-1)^k q^{5\binom{k}{2}+2k}\\ &= \frac{(q^5,q^2,q^3;q^5)_\infty}{(q;q)_\infty} = \frac{1}{(q,q^4;q^5)_\infty}.\end{aligned}$$

Similarly, the case $n = 1$ of Theorem 6.1 results in another identity due to Rogers and Ramanujan:

$$\begin{aligned}\sum_{m=0}^{\infty}\frac{q^{m^2+m}}{(q;q)_m} &= \sum_{m=0}^{\infty}\frac{q^{m^2}}{(q;q)_m} - \sum_{m=1}^{\infty}\frac{q^{m^2}}{(q;q)_{m-1}}\\ &= \frac{1}{(q;q)_\infty}\left(1+\sum_{k=1}^{\infty}(1-q^{2k+1})q^{5\binom{k}{2}+4k}\right)\\ &= \frac{1}{(q;q)_\infty}\sum_{k=-\infty}^{\infty}(-1)^k q^{5\binom{k}{2}+4k}\\ &= \frac{(q,q^4,q^5;q^5)_\infty}{(q;q)_\infty} = \frac{1}{(q^2,q^3;q^5)_\infty}.\end{aligned}$$

References

[Andrews and Askey 1981] G. E. Andrews and R. Askey, "Another q-extension of the beta function", *Proc. Amer. Math. Soc.* **81**:1 (1981), 97–100. MR 81j:33001 Zbl 0471.33001

[Andrews et al. 1999] G. E. Andrews, R. Askey, and R. Roy, *Special functions*, Encyclopedia of Mathematics and its Applications **71**, Cambridge University Press, 1999. MR 2000g:33001 Zbl 0920.33001

[Baxter 1982] R. J. Baxter, *Exactly solved models in statistical mechanics*, Academic Press, London, 1982. MR 86i:82002a Zbl 0538.60093

[Chapman 2005] R. Chapman, "A probabilistic proof of the Andrews–Gordon identities", *Discrete Math.* **290**:1 (2005), 79–84. MR 2005i:11145 Zbl 1081.11066

[Evans 2002] S. N. Evans, "Elementary divisors and determinants of random matrices over a local field", *Stochastic Process. Appl.* **102**:1 (2002), 89–102. MR 2004c:15041 Zbl 1075.15500

[Fulman 2001] J. Fulman, "A probabilistic proof of the Rogers–Ramanujan identities", *Bull. London Math. Soc.* **33**:4 (2001), 397–407. MR 2002b:11146 Zbl 1040.11074

[Gasper and Rahman 1990] G. Gasper and M. Rahman, *Basic hypergeometric series*, Encyclopedia of Mathematics and its Applications **35**, Cambridge University Press, 1990. MR 91d:33034 Zbl 0695.33001

[Jackson 1910] F. H. Jackson, "On q-definite integrals", *Quart. J. Pure and Appl. Math* **41** (1910), 101–112. JFM 41.0317.04

[Lepowsky and Milne 1978] J. Lepowsky and S. Milne, “Lie algebraic approaches to classical partition identities”, *Adv. in Math.* **29**:1 (1978), 15–59. MR 82f:17005 Zbl 0384.10008

[Liu 2003] Z.-G. Liu, “Some operator identities and q-series transformation formulas”, *Discrete Math.* **265**:1-3 (2003), 119–139. MR 2004c:33034 Zbl 1021.05010

[Milne 1997] S. C. Milne, “Balanced ${}_3\phi_2$ summation theorems for U(n) basic hypergeometric series”, *Adv. Math.* **131**:1 (1997), 93–187. MR 99d:33025 Zbl 0886.33014

[Rawlings 1997] D. Rawlings, “Absorption processes: models for q-identities”, *Adv. in Appl. Math.* **18**:2 (1997), 133–148. MR 98b:05010 Zbl 0867.05003

[Srivastava and Jain 1989] H. M. Srivastava and V. K. Jain, “Some multilinear generating functions for q-Hermite polynomials”, *J. Math. Anal. Appl.* **144**:1 (1989), 147–157. MR 91g:33024a Zbl 0665.33008

[Tannery 1904] J. Tannery, *Introduction à la théorie des fonctions d’une variable, I: Nombres irrationnels, ensembles, limites, fonctions élémentaires, dérivées*, 2nd ed., A. Hermann, Paris, 1904. JFM 35.0374.01

[Wang 2008] M. Wang, “A remark on Andrews–Askey integral”, *J. Math. Anal. Appl.* **341**:2 (2008), 1487–1494. MR 2009e:33056 Zbl 1142.33006

[Wang 2009a] M. Wang, “Generalizations of Milne’s U($n+1$) q-binomial theorems”, *Comput. Math. Appl.* **58**:1 (2009), 80–87. MR 2010f:33027

[Wang 2009b] M. Wang, “q-integral representation of the Al-Salam–Carlitz polynomials”, *Appl. Math. Lett.* **22**:6 (2009), 943–945. MR 2523611 Zbl 1173.33305

[Wang 2010a] M. Wang, “An extension of the q-beta integral with applications”, *J. Math. Anal. Appl.* **365**:2 (2010), 653–658. MR 2587068 Zbl 05676137

[Wang 2010b] M. Wang, “A recurring q-integral formula”, *Appl. Math. Lett.* **23**:3 (2010), 256–260. MR 2565186 Zbl 1183.33036

[Watson 1929] G. N. Watson, “A new proof of the Rogers–Ramanujan identities”, *J. London Math. Soc.* **4** (1929), 4–9. JFM 55.0219.09

[Zhang 2006] Z. Zhang, “Operator identities and several $U(n+1)$ generalizations of the Kalnins–Miller transformations”, *J. Math. Anal. Appl.* **324**:2 (2006), 1152–1167. MR 2008b:33043 Zbl 1113.33020

[Zhang and Liu 2006] Z. Zhang and M. Liu, “Applications of operator identities to the multiple q-binomial theorem and q-Gauss summation theorem”, *Discrete Math.* **306**:13 (2006), 1424–1437. MR 2007f:33025 Zbl 1095.05002

Received February 21, 2009. Revised December 12, 2009.

MINGJIN WANG
DEPARTMENT OF APPLIED MATHEMATICS
CHANGZHOU UNIVERSITY
CHANGZHOU 213164
CHINA
wmj@cczu.edu.cn

Guidelines for Authors

Authors may submit manuscripts at pjm.math.berkeley.edu/about/journal/submissions.html and choose an editor at that time. Exceptionally, a paper may be submitted in hard copy to one of the editors; authors should keep a copy.

By submitting a manuscript you assert that it is original and is not under consideration for publication elsewhere. Instructions on manuscript preparation are provided below. For further information, visit the web address above or write to pacific@math.berkeley.edu or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use LaTeX, but papers in other varieties of TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a website in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PACIFIC JOURNAL OF MATHEMATICS

Volume 247 No. 1 September 2010

0030-8730(201009)247:1;1-8