

*Pacific  
Journal of  
Mathematics*

Volume 265 No. 1

September 2013

# PACIFIC JOURNAL OF MATHEMATICS

msp.org/pjm

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

## EDITORS

V. S. Varadarajan (Managing Editor)  
Department of Mathematics  
University of California  
Los Angeles, CA 90095-1555  
pacific@math.ucla.edu

Paul Balmer  
Department of Mathematics  
University of California  
Los Angeles, CA 90095-1555  
balmer@math.ucla.edu

Don Blasius  
Department of Mathematics  
University of California  
Los Angeles, CA 90095-1555  
blasius@math.ucla.edu

Vyjayanthi Chari  
Department of Mathematics  
University of California  
Riverside, CA 92521-0135  
chari@math.ucr.edu

Daryl Cooper  
Department of Mathematics  
University of California  
Santa Barbara, CA 93106-3080  
cooper@math.ucsb.edu

Robert Finn  
Department of Mathematics  
Stanford University  
Stanford, CA 94305-2125  
finn@math.stanford.edu

Kefeng Liu  
Department of Mathematics  
University of California  
Los Angeles, CA 90095-1555  
liu@math.ucla.edu

Jiang-Hua Lu  
Department of Mathematics  
The University of Hong Kong  
Pokfulam Rd., Hong Kong  
jhlu@maths.hku.hk

Sorin Popa  
Department of Mathematics  
University of California  
Los Angeles, CA 90095-1555  
popa@math.ucla.edu

Jie Qing  
Department of Mathematics  
University of California  
Santa Cruz, CA 95064  
qing@cats.ucsc.edu

Paul Yang  
Department of Mathematics  
Princeton University  
Princeton NJ 08544-1000  
yang@math.princeton.edu

## PRODUCTION

Silvio Levy, Scientific Editor, production@msp.org

## SUPPORTING INSTITUTIONS

ACADEMIA SINICA, TAIPEI  
CALIFORNIA INST. OF TECHNOLOGY  
INST. DE MATEMÁTICA PURA E APLICADA  
KEIO UNIVERSITY  
MATH. SCIENCES RESEARCH INSTITUTE  
NEW MEXICO STATE UNIV.  
OREGON STATE UNIV.

STANFORD UNIVERSITY  
UNIV. OF BRITISH COLUMBIA  
UNIV. OF CALIFORNIA, BERKELEY  
UNIV. OF CALIFORNIA, DAVIS  
UNIV. OF CALIFORNIA, LOS ANGELES  
UNIV. OF CALIFORNIA, RIVERSIDE  
UNIV. OF CALIFORNIA, SAN DIEGO  
UNIV. OF CALIF., SANTA BARBARA

UNIV. OF CALIF., SANTA CRUZ  
UNIV. OF MONTANA  
UNIV. OF OREGON  
UNIV. OF SOUTHERN CALIFORNIA  
UNIV. OF UTAH  
UNIV. OF WASHINGTON  
WASHINGTON STATE UNIVERSITY

These supporting institutions contribute to the cost of publication of this Journal, but they are not owners or publishers and have no responsibility for its contents or policies.

---

See inside back cover or [msp.org/pjm](http://msp.org/pjm) for submission instructions.

---

The subscription price for 2013 is US \$400/year for the electronic version, and \$485/year for print and electronic. Subscriptions, requests for back issues and changes of subscribers address should be sent to Pacific Journal of Mathematics, P.O. Box 4163, Berkeley, CA 94704-0163, U.S.A. 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, 798 Evans Hall #3840, Berkeley, CA 94720-3840, is published twelve times a year. 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

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2013 Mathematical Sciences Publishers

## GENUS-TWO GOERITZ GROUPS OF LENS SPACES

SANGBUM CHO

**Given a genus- $g$  Heegaard splitting of a 3-manifold, the Goeritz group is defined to be the group of isotopy classes of orientation-preserving homeomorphisms of the manifold that preserve the splitting. In this work, we show that the Goeritz groups of genus-2 Heegaard splittings for lens spaces  $L(p, 1)$  are finitely presented, and give explicit presentations of them.**

### 1. Introduction

It is well known that every closed orientable 3-manifold  $M$  can be decomposed into two handlebodies of the same genus. This is what we call a Heegaard splitting of the manifold, and the genus of the handlebodies is called the genus of the splitting. Given a genus- $g$  Heegaard splitting of  $M$ , the *Goeritz group* of the splitting, which we will denote by  $\mathcal{G}_g$ , is the group of isotopy classes of orientation-preserving homeomorphisms of  $M$  that preserve each of the handlebodies of the splitting setwise. In particular, this group is interesting when the manifold is the 3-sphere or a lens space since it is well known from [Waldhausen 1968; Bonahon 1983; Bonahon and Otal 1983] that they have unique Heegaard splittings for each genus up to isotopy. In this case, each Goeritz group depends only on the genus of the splitting, and so we can define the *genus- $g$  Goeritz group*  $\mathcal{G}_g$  of each of those manifolds without mentioning a specific Heegaard splitting. For the 3-sphere, it was shown in [Goeritz 1933; Scharlemann 2004] that  $\mathcal{G}_2$  is finitely generated, and subsequently in [Akbas 2008; Cho 2008] that  $\mathcal{G}_2$  is finitely presented and its finite presentation was introduced. Further, in [Koda 2011], a natural generalization of a Goeritz group is studied, namely, the group of isotopy classes of orientation-preserving homeomorphisms of the 3-sphere preserving an embedded genus-two handlebody which is possibly knotted.

---

This work is supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) and funded by the Ministry of Education, Science, and Technology (2012006520).

*MSC2010:* primary 57N10; secondary 57M60.

*Keywords:* Heegaard splitting, Goeritz group, lens space, disk complex.

In this work, we show that the Goeritz group  $\mathcal{G}_2$  of each of the lens spaces  $L(p, 1)$  is finitely presented. In the main theorem, Theorem 5.4, their explicit presentations are given. For the genus-2 Goeritz groups of the other lens spaces, and for the higher genus Goeritz groups of the 3-sphere and lens spaces, it is conjectured that they are all finitely presented, but it is still known to be an open problem.

We generalize the method developed in [Cho 2008]. We find a tree on which  $\mathcal{G}_2$  for  $L(p, 1)$  acts such that the quotient of the tree by the action of  $\mathcal{G}_2$  is a single edge, and then apply the well known theory of groups acting on trees due to Bass and Serre (see [Serre 1980]). Such a tree will be found in the barycentric subdivision of the disk complex for one of the handlebodies of the splitting. For arbitrary lens spaces  $L(p, q)$ , finding such trees, if they exist, is a much more complicated problem than  $L(p, 1)$ , which will be fully discussed in [Cho and Koda 2012].

Throughout the paper, we simply denote by  $\mathcal{G}$  the genus-2 Goeritz group  $\mathcal{G}_2$  of a lens space. We use the standard notation  $L(p, q)$  with  $p \geq 2$  for a lens space with its basic properties found in standard textbooks. For an example, we refer to [Rolfsen 1976]. For a genus-1 Heegaard splitting of  $L(p, 1)$ , any oriented meridian circle of a solid torus of the splitting is identified with a  $(p, 1)$ -curve (or a  $(p, p - 1)$ -curve) on the boundary of the other solid torus after a suitable choice of oriented longitude and meridian of the other solid torus is made. The triple  $(V, W; \Sigma)$  will denote a genus-2 Heegaard splitting of a lens space  $L = L(p, q)$ . That is,  $L = V \cup W$  and  $V \cap W = \partial V = \partial W = \Sigma$ , where  $V$  and  $W$  are handlebodies of genus two.

The disks  $D$  and  $E$  in a handlebody are always assumed to be properly embedded, and their intersection is transverse and minimal up to isotopy. In particular, if  $D$  intersects  $E$ , then  $D \cap E$  is a collection of pairwise disjoint arcs that are properly embedded in both  $D$  and  $E$ . Finally,  $\text{Nbd}(X)$  will denote a regular neighborhood of  $X$ , and  $\text{cl}(X)$  the closure of  $X$  for a subspace  $X$  of a polyhedral space where the ambient space will always be clear from the context.

## 2. Primitive elements of the free group of rank two

The fundamental group of the genus-2 handlebody is the free group  $\mathbb{Z} * \mathbb{Z}$  of rank two. We call an element of  $\mathbb{Z} * \mathbb{Z}$  *primitive* if it is a member of a generating pair of  $\mathbb{Z} * \mathbb{Z}$ . Primitive elements of  $\mathbb{Z} * \mathbb{Z}$  have been well understood. For an example we refer [Osborne and Zieschang 1981] to the reader. A key property of the primitive elements of the free group of rank two is the following, which is a direct consequence of Corollary 3.3 in [Osborne and Zieschang 1981]:

**Proposition 2.1.** *Fix a generating pair  $\{x, y\}$  of  $\mathbb{Z} * \mathbb{Z}$ , and let  $w$  be a primitive element of  $\mathbb{Z} * \mathbb{Z}$ . Then for some  $\epsilon \in \{1, -1\}$  and some  $n \in \mathbb{Z}$ , some cyclically reduced form of  $w$  is a product of terms of the form  $x^\epsilon y^n$  or  $x^\epsilon y^{n+1}$ , or else a product of terms of the form  $y^\epsilon x^n$  or  $y^\epsilon x^{n+1}$ .*

From the proposition, the cyclically reduced forms of a primitive element are very restrictive. For example, if  $w$  is a primitive element of  $\mathbb{Z} * \mathbb{Z}$ , then no cyclically reduced form of  $w$  in terms of  $x$  and  $y$  can contain  $x$  and  $x^{-1}$  (and  $y$  and  $y^{-1}$ ) simultaneously.

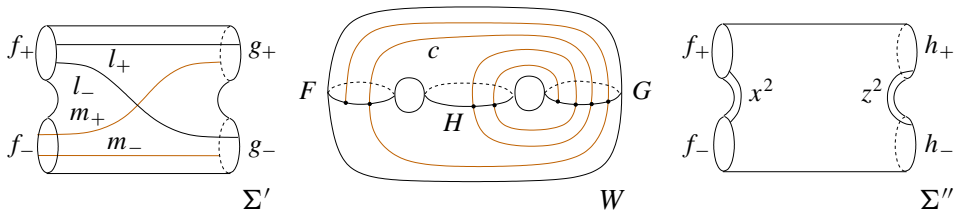
A simple closed curve in the boundary of a genus-2 handlebody  $W$  represents an element of  $\pi_1(W) = \mathbb{Z} * \mathbb{Z}$ . We call a pair of essential disks in  $W$  a *complete meridian system* for  $W$  if the union of the two disks cuts up  $W$  into a 3-ball. Given a complete meridian system  $\{F, G\}$ , assign symbols  $x$  and  $y$  to circles  $\partial F$  and  $\partial G$  respectively. Suppose that an oriented simple closed curve  $l$  on  $\partial W$  meets  $\partial F \cup \partial G$  transversely and minimally. Then  $l$  determines a word in terms of  $x$  and  $y$  which can be read off from the intersections of  $l$  with  $\partial F$  and  $\partial G$  (after a choice of orientations of  $\partial F$  and  $\partial G$ ), and hence  $l$  represents an element of the free group  $\pi_1(W) = \langle x, y \rangle$ .

In this set up, the following is a simple criterion for the primitiveness of the elements represented by such a simple closed curve:

**Lemma 2.2.** *With a suitable choice of orientations of  $\partial F$  and  $\partial G$ , if a word determined by the simple closed curve  $l$  contains one of the subwords  $yx y^{-1}$  or  $xy x y^n$  for  $n \geq 3$ , then any element in  $\pi_1(W)$  represented by  $l$  cannot be a primitive element.*

*Proof.* Let  $\Sigma'$  be the 4-holed sphere cut up from  $\partial W$  along  $\partial F \cup \partial G$ , and denote by  $f_+$  and  $f_-$  (respectively  $g_+$  and  $g_-$ ) the boundary circles of  $\Sigma'$  that came from  $\partial F$  (respectively  $\partial G$ ).

Suppose first that a word represented by  $l$  contains a subword of the form  $yx y^{-1}$ . Then we may assume that there are two arcs  $l_+$  and  $l_-$  of  $l \cap \Sigma'$  such that  $l_+$  connects  $f_+$  and  $g_+$ , and  $l_-$  connects  $f_+$  and  $g_-$  as in Figure 1, left. Since  $|l \cap f_+| = |l \cap f_-|$  and  $|l \cap g_+| = |l \cap g_-|$ , we must have two other arcs  $m_+$  and  $m_-$  of  $l \cap \Sigma'$  such that  $m_+$  connects  $f_-$  and  $g_+$ , and  $m_-$  connects  $f_-$  and  $g_-$ . We see then that there exists no arc component of  $l \cap \Sigma'$  that meets only one of  $f_+$ ,  $f_-$ ,  $g_+$  or  $g_-$ . That is, any word determined by  $l$  contains neither  $x^{\pm 1} x^{\mp 1}$  nor  $y^{\pm 1} y^{\mp 1}$ , and so each word is cyclically reduced, but a word determined by  $l$  already contains both  $y$  and  $y^{-1}$ , and so  $l$  cannot represent a primitive element of  $\pi_1(W)$  by Proposition 2.1.



**Figure 1.** The 4-holed spheres  $\Sigma'$  and  $\Sigma''$ .

Next, suppose that a word represented by  $l$  contains a subword of the form  $xyxy^n$  for  $n \geq 3$ . We may assume there is an arc  $c$  of  $l \cap \Sigma'$  connecting  $f_+$  and  $g_+$  in  $\Sigma'$ . Consider the circle which is the frontier of a regular neighborhood of  $f_+ \cup c \cup g_+$  in  $\Sigma'$ . This circle bounds a disk  $H$  in  $W$ , and  $\{F, H\}$  forms a complete meridian system of  $W$ . Assigning symbols  $x$  and  $z$  to  $\partial F$  and  $\partial H$  respectively, the circle  $l$  represents an element of  $\pi_1(W) = \langle x, z \rangle$  (see Figure 1, middle).

Let  $\Sigma''$  be the 4-holed sphere cut up from  $\partial W$  along  $\partial F \cup \partial H$ , and denote by  $f_+$  and  $f_-$  (respectively  $h_+$  and  $h_-$ ) the boundary circles of  $\Sigma''$  that came from  $\partial F$  (respectively  $\partial H$ ). There are two arcs of  $l \cap \Sigma''$  such that one connects  $f_+$  and  $f_-$ , and the other one connects  $h_+$  and  $h_-$ . We may assume that these two arcs represent subwords of the form  $x^2$  and  $z^2$  (see Figure 1, right). Thus there exists no arc component of  $l \cap \Sigma''$  that meets only one of  $f_+$ ,  $f_-$ ,  $h_+$  and  $h_-$ . That is, each word represented by  $l$  is cyclically reduced. But a word determined by  $l$  already contains both  $x^2$  and  $z^2$ , and so  $l$  cannot represent a primitive element of  $\pi_1(W)$  by Proposition 2.1 again.  $\square$

### 3. Primitive disks in a handlebody

Recall that  $(V, W; \Sigma)$  denotes a genus-two Heegaard splitting of a lens space  $L = L(p, q)$  with  $p \geq 2$ . We call an essential disk  $E$  in  $V$  *primitive* if there exists an essential disk  $E'$  in  $W$  such that  $\partial E$  intersects  $\partial E'$  transversely in a single point. Such a disk  $E'$  is called a *dual disk* of  $E$ . Note that  $E'$  is also primitive in  $W$  with a dual disk  $E$ , and  $W \cup \text{Nbd}(E)$  and  $V \cup \text{Nbd}(E')$  are both solid tori. Primitive disks are necessarily nonseparating. We call a pair of disjoint, nonisotopic primitive disks in  $V$  a *primitive pair* in  $V$ . Similarly, a triple of pairwise disjoint, nonisotopic, primitive disks (if it exists) is a *primitive triple*.

A nonseparating disk  $E_0$  properly embedded in  $V$  is called *semiprimitive* if there is a primitive disk  $E'$  in  $W$  such that  $\partial E'$  is disjoint from  $\partial E_0$ . With a suitable choice of oriented meridian and longitude circles on the boundary of the solid torus obtained by cutting up  $W$  along  $E'$ , the oriented boundary circle  $\partial E_0$  can be considered a  $(p, 1)$ -curve on the boundary of the solid torus, if  $q = 1$ .

Any simple closed curve on the boundary of  $W$  represents an element of  $\pi_1(W)$ , which is the free group of rank two. We can interpret primitive disks algebraically as follows, which is a direct consequence of [Gordon 1987]:

**Lemma 3.1.** *Let  $D$  be a nonseparating disk in  $V$ . Then  $D$  is primitive if and only if  $\partial D$  represents a primitive element of  $\pi_1(W)$ .*

Note that no disk can be both primitive and semiprimitive since the boundary circle of a semiprimitive disk in  $V$  represents the  $p$ -th power of a primitive element of  $\pi_1(W)$ .

Let  $D$  and  $E$  be essential disks in  $V$ , and suppose that  $D$  intersects  $E$  transversely and minimally. Let  $C \subset D$  be a disk cut up from  $D$  by an outermost arc  $\beta$  of  $D \cap E$  in  $D$  such that  $C \cap E = \beta$ . We call such a  $C$  an *outermost subdisk* of  $D$  cut up by  $D \cap E$ . The arc  $\beta$  cuts  $E$  into two disks, say  $G$  and  $H$ . Then we have two essential disks  $E_1$  and  $E_2$  in  $V$  which are isotopic to disks  $G \cup C$  and  $H \cup C$  respectively. We call  $E_1$  and  $E_2$  the *disks from surgery* on  $E$  along the outermost subdisk  $C$  of  $D$  cut up by  $D \cap E$ . Observe that  $E_1$  and  $E_2$  each have fewer arcs of intersection with  $D$  than  $E$  had, since at least the arc  $\beta$  no longer counts.

Since  $E$  and  $D$  are assumed to intersect minimally,  $E_1$  and  $E_2$  are isotopic to neither  $E$  nor  $D$ . In particular, if both  $D$  and  $E$  are nonseparating, then the resulting disks  $E_1$  and  $E_2$  are both nonseparating and they are not isotopic to each other. Further,  $E_1$  and  $E_2$  are meridian disks of the solid torus  $V$  cut up by  $E$ , and the boundary circles  $\partial E_1$  and  $\partial E_2$  are not isotopic to each other in the two holed torus  $\partial V$  cut up by  $\partial E$ .

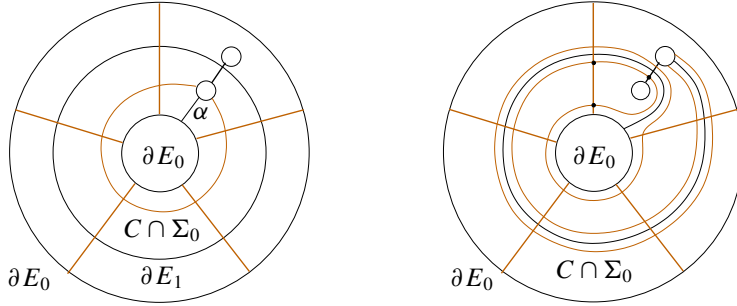
**Theorem 3.2.** *Let  $(V, W; \Sigma)$  be the genus-two Heegaard splitting of the lens space  $L = L(p, 1)$  with  $p \geq 2$ . Let  $D$  and  $E$  be primitive disks in  $V$  which intersect each other transversely and minimally. Then one of the two disks from surgery on  $E$  along an outermost subdisk of  $D$  cut up by  $D \cap E$  is primitive. Furthermore, it has a common dual disk with  $E$ .*

*Proof.* We will prove the theorem only for  $p \geq 5$ . The cases of  $p \in \{2, 3, 4\}$  will be similar but simpler.

Let  $C$  be an outermost subdisk of  $D$  cut up by  $D \cap E$ . The choice of a dual disk  $E'$  of  $E$  determines a unique semiprimitive disk  $E_0$  in  $V$ , namely, the meridian disk  $E_0$  of  $V$  disjoint from  $E \cup E'$ . Among all the dual disks of  $E$ , choose one, denoted by  $E'$  again, so that the semiprimitive  $E_0$  determined by  $E'$  intersects  $C$  minimally. Further, there is a unique semiprimitive disk  $E'_0$  in  $W$  disjoint from  $E \cup E'$ . We give symbols  $x$  and  $y$  to oriented  $\partial E'$  and  $\partial E'_0$  respectively to have  $\pi_1(W) = \langle x, y \rangle$ . For convenience, we simply identify the boundary circles  $\partial E'$  and  $\partial E'_0$  with the assigned symbols  $x$  and  $y$  respectively. Notice that the circle  $y$  is disjoint from  $\partial E$  and intersects  $\partial E_0$  in  $p$  points in the same direction, and  $x$  is disjoint from  $\partial E_0$  and intersects  $\partial E$  in a single point. Thus we may assume that  $\partial E_0$  and  $\partial E$  determine the words  $y^p$  and  $x$  respectively.

Let  $\Sigma_0$  be the 4-holed sphere  $\partial V$  cut up by  $\partial E \cup \partial E_0$ . We regard  $\Sigma_0$  as a 2-holed annulus where the two boundary circles came from  $\partial E_0$  and the two holes came from  $\partial E$ . Then  $y \cap \Sigma_0$  is the union of  $p$  spanning arcs which cut  $\Sigma_0$  into  $p$  rectangles, and  $x$  is a single arc connecting two holes which are contained in a single rectangle. See Figure 2, left.

Suppose first that  $C$  is disjoint from  $E_0$ . Note that one of the disks from surgery on  $E$  along  $C$  is  $E_0$ , which is semiprimitive. The arc  $C \cap \Sigma_0$  is the frontier of



**Figure 2.** The 2-holed annulus  $\Sigma_0$  in  $L(5, 1)$ .

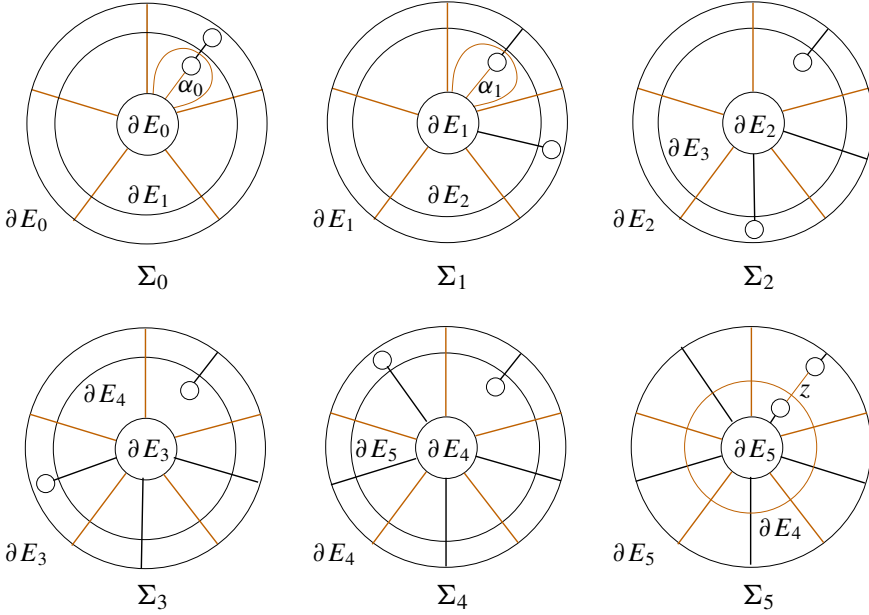
a regular neighborhood of the union of one boundary circle of  $\Sigma_0$  and an arc  $\alpha$  connecting the boundary circle to a hole. Observe that the arc  $\alpha$  is disjoint from  $y \cap \Sigma_0$ , otherwise a word of  $\partial D$  must contain  $yx y^{-1}$  (after changing orientation if necessary) which contradicts that  $D$  is primitive, by Lemma 2.2. See Figure 2, right. Consequently, if we denote by  $E_1$  the disk from surgery that is not  $E_0$ , then  $\partial E_1$  intersects  $\partial E'$  in a single point. That is, the resulting disk  $E_1$  is primitive with the common dual disk  $E'$  of  $E$ . See Figure 2, left.

From now on, we assume that  $C$  intersects  $E_0$ . Let  $C_0$  be an outermost subdisk of  $C$  cut up by  $C \cap E_0$ . The arc  $C_0 \cap \Sigma_0$  is the frontier of a regular neighborhood of one hole of  $\Sigma_0$  and an arc, say  $\alpha_0$ , connecting the hole to a boundary circle of  $\Sigma_0$ . By the same reasoning as in the case of  $\alpha$ , the arc  $\alpha_0$  is disjoint from  $y \cap \Sigma_0$ . Thus one of the disks from surgery on  $E_0$  along  $C_0$  is  $E$ , and the other one, denoted by  $E_1$  again, is primitive since  $\partial E_1$  intersects  $\partial E'$  in a single point as in the previous case. Note that  $|C \cap E_1| < |C \cap E_0|$  from the surgery construction. See  $\Sigma_0$  in Figure 3.

Let  $\Sigma_1$  be the 4-holed sphere  $\partial V$  cut up by  $\partial E \cup \partial E_1$ . We regard  $\Sigma_1$  as a 2-holed annulus, like  $\Sigma_0$ , where the two boundary circles came from  $\partial E_1$  and the two holes came from  $\partial E$ . Then  $y \cap \Sigma_1$  is the union of  $p$  spanning arcs which cut  $\Sigma_1$  into  $p$  rectangles as in the case of  $\Sigma_0$ , but the two holes, which came from  $\partial E$ , are now contained in different consecutive rectangles, and  $x \cap \Sigma_1$  is the union of two arcs each joining a hole and a boundary circle of  $\Sigma_1$  as in Figure 3. If the original subdisk  $C$  is disjoint from  $E_1$ , then we are done since  $E_1$  is the desired primitive disk resulting from the surgery.

Suppose that  $C$  also intersects  $E_1$ , and let  $C_1$  be an outermost subdisk of  $C$  cut up by  $C \cap E_1$ . Then  $C_1 \cap \Sigma_1$  is the frontier of a regular neighborhood of the union of one hole of  $\Sigma_1$  and an arc, say  $\alpha_1$ , connecting the hole to a boundary circle. The arc  $\alpha_1$  is also disjoint from  $y \cap \Sigma_1$  by the same reasoning as for  $\alpha_0$ . Thus if we denote by  $E_2$  the disk from surgery on  $E_1$  along  $C_1$  that is not  $E$ , then  $\partial E_2$  represents a word  $xyxy^{p-1}$ . See  $\Sigma_1$  in Figure 3.





**Figure 3.** The sequence of 2-holed annuli from the consecutive surgeries for  $L(5, 1)$ .

We continue such a construction repeatedly whenever  $C$  also intersects the next disk. For each  $1 \leq j \leq p - 1$ , if  $C$  intersects  $E_j$ , then we obtain the disk  $E_{j+1}$  from surgery on  $E_j$  along an outermost subdisk  $C_j$  cut up by  $C \cap E_j$ . We see that  $|C \cap E_{j+1}| < |C \cap E_j|$  from the surgery construction. In the 2-holed annulus  $\Sigma_j$ , the arc  $C_j \cap \Sigma_j$  is the frontier of a regular neighborhood of the union of a hole of  $\Sigma_j$  and an arc  $\alpha_j$  connecting the hole to a boundary circle. The arc  $\alpha_j$  is disjoint from  $y \cap \Sigma_j$ , and so  $\partial E_{j+1}$  represents a word of the form  $(xy)^j xy^{p-j}$ . In particular, notice that the disk  $E_p$  is semiprimitive and  $E_{p-1}$  is primitive, since there is a primitive disk  $E''$  in  $W$  disjoint from  $\partial E_p$  that intersects  $\partial E_{p-1}$  in a single point. Such an  $E''$  is not hard to find. In the final 2-holed annulus  $\Sigma_5$  in Figure 3, the arc  $z$  is the boundary circle of  $E''$  in  $\Sigma_p$ . Note that  $z$  is disjoint from  $x \cup y$ , and so it does bound a disk  $E''$  in the 3-ball  $W$  cut up by  $E' \cup E'_0$ . Also,  $z$  intersects  $\partial E_{p-1}$  in a single point and is disjoint from  $\partial E_p$ .

We remark that each of the arcs  $\alpha_j$ ,  $j \in \{0, 1, \dots, p - 1\}$ , is disjoint from the circle  $y$  due to the fact that  $D$  is primitive. There are infinitely many arcs  $\alpha_0$  that are not isotopic to each other in  $\Sigma_0$ , but each arc  $\alpha_j$  in  $\Sigma_j$  with  $j \geq 1$  is unique up to isotopy. Therefore, once  $E_1$  is determined, we have the unique sequence of disks  $E_2, E_3, \dots, E_p$  only under the condition that each  $\alpha_j$  is disjoint from  $y$ .

**Claim.** For each  $j \in \{2, 3, \dots, p - 1\}$ , the subdisk  $C$  intersects  $E_j$ .

*Proof of claim.* Suppose not, and let  $E_j$  be the first disk disjoint from  $C$ . First, suppose that  $j \in \{2, 3, \dots, p-3\}$ . Then  $C$  is disjoint from  $E_j$  and intersects  $E_{j-1}$ , and so the arc  $\partial C \cap \Sigma_j$  gives a subword of  $\partial D$  of the form  $(yx)^j y^{p-j}$  which implies that  $D$  is not primitive by Lemma 2.2 again, which is a contradiction. Next, suppose that  $j = p-2$ . That is,  $C$  is disjoint from  $E_{p-2}$  and intersects  $E_{p-3}$ . Then one of the resulting disks from surgery on  $E$  along  $C$  is  $E_{p-2}$ , and the other one is exactly  $E_{p-1}$ , which is a disk in the sequence of disks in the previous construction. The subdisk  $C$  is disjoint from  $E_{p-2} \cup E_{p-1}$ , and consequently,  $C$  necessarily intersects the semiprimitive disk  $E_p$  in the previous construction in a single arc. That is,  $|C \cap E_p| = 1$ . But from the consecutive surgery constructions for  $j \in \{2, 3, \dots, p-3\}$ , we have  $1 \leq |C \cap E_{p-3}| < |C \cap E_0|$ , which contradicts the minimality of  $|C \cap E_0|$ . Similarly, if  $j = p-1$ , then we have the same contradiction on the minimality, since  $C$  is disjoint from  $E_p$  in this case. This proves the claim.

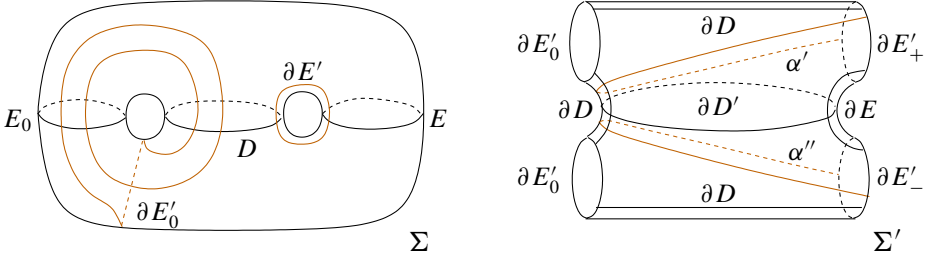
By the claim, we can do surgery on  $E_{p-1}$  along  $C_{p-1}$  and one resulting disk from surgery is  $E_p$ , the semiprimitive disk. But  $|C \cap E_{j+1}| < |C \cap E_j|$  for each  $j \in \{1, 2, \dots, p-1\}$ , and consequently  $|C \cap E_p| < |C \cap E_0|$ , which contradicts the minimality of  $|C \cap E_0|$  again.

Therefore the primitive disk  $E_1$  is a disk from surgery on  $E$  along  $C$ , and  $E'$  is also a dual disk of  $E_1$ , and so we complete the proof. We note that the other disk from surgery is either  $E_0$  or  $E_2$  depending on whether  $C$  is disjoint from  $E_0$  or not.  $\square$

**Theorem 3.3.** *Let  $(V, W; \Sigma)$  be the genus-two Heegaard splitting of the lens space  $L = L(p, 1)$  with  $p \geq 2$ . Then, for every primitive pair  $\{D, E\}$  of  $V$ ,  $D$  and  $E$  have a common dual disk. In particular, the two disks of each primitive pair have a unique common dual disk if  $p \geq 3$ , and have exactly two common dual disks if  $p = 2$  which form a primitive pair in  $W$ .*

*Proof.* The proof of the existence of a common dual disk goes almost in the same way as that of Theorem 3.2, by taking the primitive disk  $D$  disjoint from  $E$  instead of the outermost subdisk  $C$  in Theorem 3.2. That is, when we choose a dual disk  $E'$  of  $E$  so that  $|\partial D \cap \partial E_0|$  is minimal where  $E_0$  is the unique semiprimitive disk in  $V$  disjoint from  $\text{Nbd}(E \cup E')$ , the primitive disk  $D$  must be  $E_1$ , having the common dual disk  $E'$  of  $E$ .

Now, let  $E'$  be a common dual disk of  $D$  and  $E$ . Let  $E_0$  and  $E'_0$  be the unique meridian disks of  $V$  and  $W$  respectively that are disjoint from  $\text{Nbd}(E \cup E')$  (see Figure 4, left). Cut the surface  $\Sigma$  along  $\partial E' \cup \partial E'_0$  to obtain the 4-holed sphere  $\Sigma'$ . Then  $\partial E \cap \Sigma'$  is a single arc in  $\Sigma'$  connecting the two holes coming from  $\partial E'$ , and  $\partial D \cap \Sigma'$  consists of  $p-1$  parallel arcs connecting the two holes coming from  $\partial E'_0$  and two arcs connecting the holes coming from  $\partial E'$  to  $\partial E'_0$  on opposite sides, as in Figure 4.



**Figure 4.** The surfaces  $\Sigma$  and  $\Sigma'$  for  $L(2, 1)$ .

Let  $D'$  be a common dual disk of  $D$  and  $E$  which is not isotopic to  $E'$ . Then an outermost subdisk  $C'$  of  $D'$  cut up by  $D' \cap (E' \cup E'_0)$  would intersect  $\partial D$  if  $C'$  is incident to  $E'$ . Denote by  $\partial E'_+$  and  $\partial E'_-$  the two holes of  $\Sigma'$  which came from  $\partial E'$ . We may assume that the endpoints of the arc  $\alpha' = C' \cap \Sigma'$  meet  $\partial E'_+$ . Since  $|\partial D' \cap \partial E'_+| = |\partial D' \cap \partial E'_-|$ , we must have one more arc component  $\alpha''$  of  $\partial D' \cap \Sigma'$  other than  $C' \cap \Sigma'$  whose endpoints meet  $\partial E'_-$  (see Figure 4, right). The arc  $\alpha''$  also intersects  $\partial D$ , and so  $\partial D'$  intersects  $\partial D$  in more than one point, which contradicts that  $D'$  is a dual disk of  $D$ . Similarly, if  $C'$  is incident to  $E'_0$ , then  $D'$  cannot be a dual disk of  $E$ . Thus we see that  $D'$  is disjoint from  $E' \cup E'_0$ .

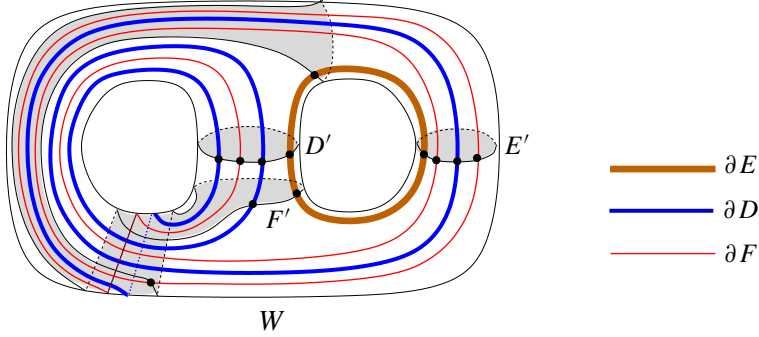
If  $p \geq 3$ , there is no possibility of such a disk  $D'$  which is disjoint from  $E' \cup E'_0$  and is not isotopic to  $E'$ , and so  $E'$  is the unique common dual disk. If  $p = 2$ , there is a unique circle in  $\Sigma'$  which is not boundary parallel and which intersects  $\partial E$  and  $\partial D$  exactly once (see the circle  $\partial D'$  in Figure 4, right). So we have exactly two common dual disks  $D'$  and  $E'$  and in this case they are disjoint from each other.  $\square$

Given a primitive disk  $D$  in  $V$ , there are infinitely many (nonisotopic) primitive disks each of which forms a primitive pair together with  $D$ . But any primitive pair can be contained in at most one primitive triple, proved as follows:

**Theorem 3.4.** *Let  $(V, W; \Sigma)$  be the genus-two Heegaard splitting of the lens space  $L = L(p, 1)$  with  $p \geq 2$ . Then there is a primitive triple of  $V$  if and only if  $p = 3$ . In this case, every primitive pair is contained in a unique primitive triple.*

*Proof.* Let  $\{E, E_1\}$  be a primitive pair of  $V$ . Choose a common dual disk  $E'$  of  $E$  and  $E_1$  given by Theorem 3.3. There are unique semiprimitive disks  $E_0$  in  $V$  and  $E'_0$  in  $W$  disjoint from  $\text{Nbd}(E \cup E')$ . Let  $\Sigma_1$  be the 4-holed sphere  $\partial V$  cut up by  $\partial E \cup \partial E_1$ , and as in Figure 3 again, consider  $\Sigma_1$  as a 2-holed annulus with two boundary circles coming from  $\partial E_1$  and two holes from  $\partial E$ . We give symbols  $x$  and  $y$  to  $\partial E'$  and  $\partial E'_0$  respectively as in the proof of Theorem 3.2.

The boundary of any primitive disk  $E_2$  in  $V$  disjoint from  $E$  and  $E_1$ , if it exists, lies in  $\Sigma_1$ , and it is the frontier of a regular neighborhood of the union of a boundary circle, a hole of  $\Sigma_1$  and an arc  $\alpha_1$  connecting them. This arc is disjoint from the



**Figure 5.** The primitive triple  $\{D', E', F'\}$  of  $W$  in  $L(3, 1)$  with the boundary circles  $\partial D$ ,  $\partial E$ , and  $\partial F$  of the disks in the primitive triple of  $V$ .

arcs  $y \cap \Sigma_1$ , otherwise  $\partial E_2$  represents a word containing  $yx y^{-1}$ ; that is,  $E_2$  is not primitive. Consequently,  $\partial E_2$  is uniquely determined and it represents a word of the form  $xyxy^{p-1}$ , and so it is primitive if and only if  $p = 3$ . Thus, only when  $p = 3$ , we have the unique primitive triple  $\{E, E_1, E_2\}$  containing the pair  $\{E, E_1\}$ .  $\square$

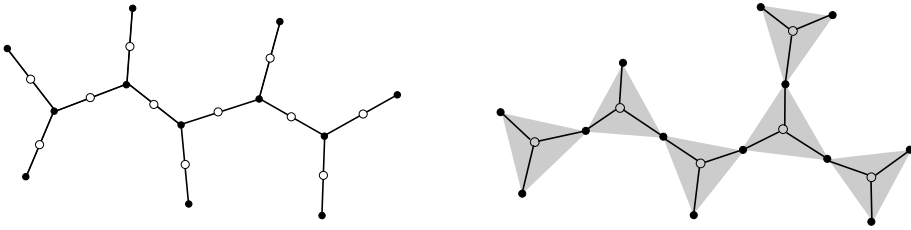
**Remark 3.5.** For any primitive triple  $\{D, E, F\}$  of  $V$  in  $L(3, 1)$ , by Theorem 3.3, there exist unique common dual disks  $D'$ ,  $E'$ , and  $F'$  of the disks in the pairs  $\{E, F\}$ ,  $\{F, D\}$ , and  $\{D, E\}$  respectively. In fact, the disks  $D'$ ,  $E'$ , and  $F'$  form a primitive triple of  $W$ . Furthermore, we have  $|\partial D' \cap \partial D| = |\partial E' \cap \partial E| = |\partial F' \cap \partial F| = 2$ . Figure 5 illustrates the triple  $\{D', E', F'\}$  of  $W$  together with the boundary circles of  $D$ ,  $E$  and  $F$  in  $\partial W = \Sigma$ .

#### 4. The complex of primitive disks

Let  $M$  be an irreducible 3-manifold with compressible boundary. The *disk complex* of  $M$  is a simplicial complex defined as follows: The vertices of the disk complex are isotopy classes of essential disks in  $M$ , and a collection of  $k + 1$  vertices spans a  $k$ -simplex if and only if it admits a collection of representative disks which are pairwise disjoint. In particular, if  $M$  is a handlebody of genus  $g \geq 2$ , then the disk complex is  $(3g - 4)$ -dimensional and is not locally finite. The following is a key property of a disk complex:

**Theorem 4.1.** *If  $\mathcal{K}$  is a full subcomplex of the disk complex satisfying the following condition, then  $\mathcal{K}$  is contractible:*

*Let  $E$  and  $D$  be disks in  $M$  representing vertices of  $\mathcal{K}$ . If  $E$  and  $D$  intersect transversely and minimally, then at least one of the disks from surgery on  $E$  along an outermost subdisk of  $D$  cut up by  $D \cap E$  represents a vertex of  $\mathcal{K}$ .*



**Figure 6.** Small portions of primitive disk complexes  $\mathcal{P}(V)$  for  $p \neq 3$  (left) and  $p = 3$  (right).

In [Cho 2008], the above theorem is proved in the case that  $M$  is a handlebody, but the proof is still valid for an arbitrary irreducible manifold with compressible boundary. From the theorem, we see that the disk complex itself is contractible.

Now consider the genus-two Heegaard splitting  $(V, W; \Sigma)$  of a lens space  $L(p, 1)$  with  $p \geq 2$ . We define the *primitive disk complex*, denoted by  $\mathcal{P}(V)$ , to be the full subcomplex of the disk complex spanned by the vertices of primitive disks in  $V$ . We already know that every primitive disk is a member of infinitely many primitive pairs, and so every vertex of  $\mathcal{P}(V)$  has infinite valency. The following is our main theorem, a direct consequence of Theorems 3.2, 3.4 and 4.1:

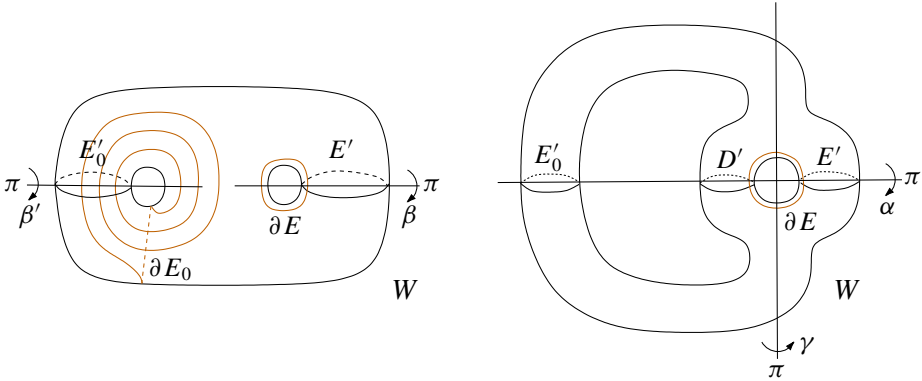
**Theorem 4.2.** *Let  $(V, W; \Sigma)$  be the genus-two Heegaard splitting of the lens space  $L = L(p, 1)$  with  $p \geq 2$ . The primitive disk complex  $\mathcal{P}(V)$  is contractible. In particular, if  $p \neq 3$  it is a tree, and if  $p = 3$  it is 2-dimensional and every edge is contained in a unique 2-simplex.*

Figure 6 illustrates portions of the primitive disk complexes. The black vertices are the vertices of  $\mathcal{P}(V)$  while the white ones are the barycenters of edges when  $p \neq 3$  and of 2-simplices when  $p = 3$ . Observe that the 2-dimensional  $\mathcal{P}(V)$  deformation retracts to a tree in its barycentric subdivision, as in the figure.

## 5. Genus-two Goeritz groups of lens spaces $L(p, 1)$

In this section, we give explicit presentation of the genus-two Goeritz group  $\mathcal{G}$  of each lens space  $L(p, 1)$ . From Theorem 4.2, if  $p \neq 3$ , the primitive disk complex  $\mathcal{P}(V)$  is a tree, and if  $p = 3$ , then  $\mathcal{P}(V)$  is 2-dimensional but deformation retracts to a tree. We simply denote by  $\mathcal{T}$  the barycentric subdivision of the tree  $\mathcal{P}(V)$  if  $p \neq 3$  and the deformation retract of  $\mathcal{P}(V)$  if  $p = 3$ . Each of the trees  $\mathcal{T}$  is bipartite, as in Figure 6, with the black vertices of (countably) infinite valency, and the white vertices of valence 2 if  $p \neq 3$  and of valence 3 if  $p = 3$ .

Each black vertex of  $\mathcal{T}$  is represented by a primitive disk, while each white vertex is represented by a primitive pair if  $p \neq 3$  and by a primitive triple if  $p = 3$ .



**Figure 7.** Generators of the stabilizer subgroup  $\mathcal{G}_{\{E\}}$ .

An element of the group  $\mathcal{G}$  can be considered a simplicial automorphism of  $\mathcal{T}$ . The tree  $\mathcal{T}$  is invariant under the action of  $\mathcal{G}$  for each  $L(p, 1)$ . In particular,  $\mathcal{G}$  acts transitively on the set of black vertices and on the set of white vertices, and hence the quotient of  $\mathcal{T}$  by the action of  $\mathcal{G}$  is a single edge of which one end vertex is black and another one white. Thus, by the theory of groups acting on trees due to Bass and Serre (see [Serre 1980]), the group  $\mathcal{G}$  can be expressed as the free product of the stabilizer subgroups of two end vertices with the amalgamated stabilizer subgroup of the edge.

First, we find a presentation of the stabilizer subgroup of a black vertex of  $\mathcal{T}$ ; that is, of (the isotopy class of) a primitive disk in  $V$ . For convenience, we will not distinguish disks (pairs and triples of disks) and homeomorphisms from their isotopy classes in their notations. Throughout the section,  $\mathcal{G}_{\{A_1, A_2, \dots, A_k\}}$  will denote the subgroup of  $\mathcal{G}$  of elements preserving  $A_1, A_2, \dots, A_k$  setwise, where  $A_i$  will be (isotopy classes of) disks or unions of disks in  $V$  or in  $W$ .

**Lemma 5.1.** *Let  $E$  be a primitive disk in  $V$ . The stabilizer subgroup  $\mathcal{G}_{\{E\}}$  of  $E$  has the presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \beta, \gamma \mid \gamma^2 = 1 \rangle$ , where the generators  $\alpha, \beta$  and  $\gamma$  are described in Figure 7.*

*Proof.* Let  $\mathcal{P}'(W)$  be the full subcomplex of the primitive disk complex  $\mathcal{P}(W)$  for  $W$  spanned by the vertices of dual disks of  $E$ . There is a unique semiprimitive disk  $E'_0$  in  $W$  disjoint from  $\partial E$ , and it is easy to show that any dual disk of  $E$  is disjoint from  $E'_0$ . Thus  $\mathcal{P}'(W)$  is 1-dimensional and further, by a similar argument used for  $\mathcal{P}(V)$ , we have that  $\mathcal{P}'(W)$  is a tree whose vertices have infinite valence. That is, when two dual disks of  $E$  intersect each other, one of the two disks from the surgery construction is  $E'_0$  and the other one is again a dual disk of  $E$ . Denote by  $\mathcal{T}'$  the barycentric subdivision of  $\mathcal{P}'(W)$ . The tree  $\mathcal{T}'$  is invariant under the action of the stabilizer subgroup  $\mathcal{G}_{\{E\}}$ , and the quotient of  $\mathcal{T}'$  by the action is a single edge. One vertex of this edge corresponds to a dual disk  $E'$  of  $E$ , and the other one to a

primitive pair  $\{E', D'\}$  of dual disks of  $E$ . Thus  $\mathcal{G}_{\{E\}}$  can be expressed as the free product of the stabilizer subgroups  $\mathcal{G}_{\{E, E'\}} * \mathcal{G}_{\{E, E' \cup D'\}}$  amalgamated by  $\mathcal{G}_{\{E, E', D'\}}$ .

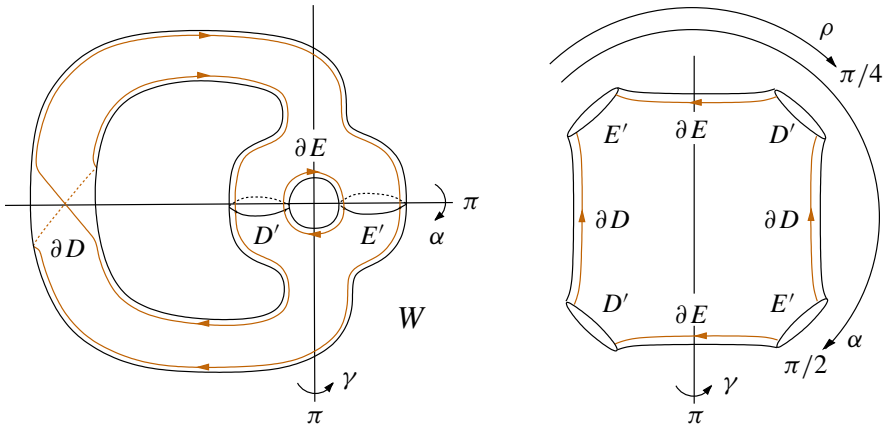
Consider the subgroup  $\mathcal{G}_{\{E, E'\}}$  first. Any element of  $\mathcal{G}_{\{E, E'\}}$  also preserves the disks  $E_0$  and  $E'_0$ , which are unique meridian disks disjoint from  $E \cup E'$  in  $V$  and in  $W$  respectively. Since  $V$  cut up by  $E \cup E_0$  and  $W$  cut up by  $E' \cup E'_0$  are all 3-balls, the group  $\mathcal{G}_{\{E, E'\}}$  is identified with the group of isotopy classes of orientation-preserving homeomorphisms of  $\Sigma = \partial V = \partial W$  which preserve  $\partial E$ ,  $\partial E'$ ,  $\partial E_0$ , and  $\partial E'_0$ . This group has a presentation  $\langle \beta, \beta' \mid (\beta\beta')^2 = 1, \beta\beta' = \beta'\beta \rangle$ , where the generators  $\beta$  and  $\beta'$  are  $\pi$ -rotations (half Dehn twists) described in Figure 7, left.

Next, consider the subgroup  $\mathcal{G}_{\{E, E' \cup D'\}}$ . Any element of this group preserves  $E' \cup D'$  in  $W$ , and further it preserves  $E$  and  $E_0 \cup D_0$  in  $V$  where  $E_0$  and  $D_0$  are unique meridian disks in  $V$  disjoint from  $E \cup E'$  and  $E \cup D'$  respectively. Thus  $\mathcal{G}_{\{E, E' \cup D'\}}$  is generated by two elements  $\alpha$  and  $\gamma$ , where  $\alpha$  is the hyperelliptic involution, and  $\gamma$  is the element of order 2 exchanging  $E'$  and  $D'$  described in Figure 7, right. Thus  $\mathcal{G}_{\{E, E' \cup D'\}}$  has the presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \gamma \mid \gamma^2 = 1 \rangle$ . Similarly,  $\mathcal{G}_{\{E, E', D'\}}$  has the presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle$ . Observing that  $\alpha$  satisfies  $\beta\beta' = \alpha$ , we have the desired presentation of  $\mathcal{G}_{\{E\}}$ .  $\square$

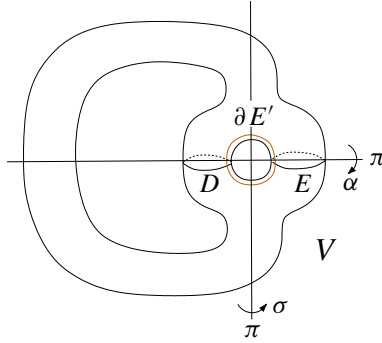
Thus the stabilizer subgroups of black vertices have the same presentation for each  $p \geq 2$ , but for white vertices, we have the following cases depending on  $p$ :

**Lemma 5.2.** *A white vertex of  $\mathcal{T}$  corresponds to a primitive pair if  $p \neq 3$  and to a primitive triple if  $p = 3$ .*

- (1) *Let  $\{D, E\}$  be a primitive pair of  $V$  in  $L(p, 1)$ . Then the stabilizer subgroup  $\mathcal{G}_{\{D \cup E\}}$  has the presentation  $\langle \rho, \gamma \mid \rho^4 = \gamma^2 = (\rho\gamma)^2 = 1 \rangle$  if  $p = 2$ , and  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \sigma \mid \sigma^2 = 1 \rangle$  if  $p \geq 3$ , where the generators are described in Figures 8 and 9.*



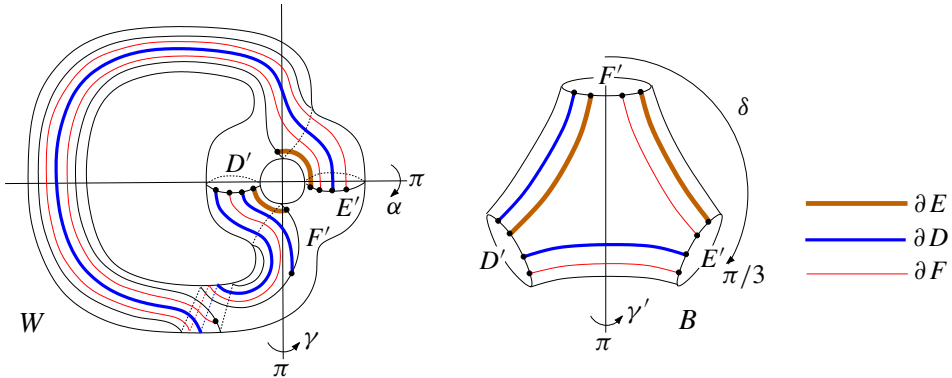
**Figure 8.** Generators of the stabilizer subgroup  $\mathcal{G}_{\{D \cup E\}}$  for  $L(2, 1)$ .



**Figure 9.** Generators of the stabilizer subgroup  $\mathcal{G}_{\{D \cup E\}}$  for  $L(p, 1)$ , with  $p \geq 3$ .

- (2) Let  $\{D, E, F\}$  be a primitive triple of  $V$  in  $L(3, 1)$ . The stabilizer subgroup  $\mathcal{G}_{\{D \cup E \cup F\}}$  has the presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \delta, \gamma \mid \delta^3 = \gamma^2 = (\gamma\delta)^2 = 1 \rangle$ , where the generators are described in Figure 10.

*Proof.* (1) First, let  $\{D, E\}$  be a primitive pair of  $V$  in  $L(2, 1)$ . Then, by Theorem 3.3, there is a unique primitive pair  $\{D', E'\}$  of  $W$  such that  $D'$  and  $E'$  are common dual disks of  $D$  and  $E$ . Any element of  $\mathcal{G}_{\{D \cup E\}}$  preserves  $D' \cup E'$ , and hence  $\mathcal{G}_{\{D \cup E\}}$  is identified with the stabilizer subgroup  $\mathcal{G}_{\{D \cup E, D' \cup E'\}}$ . Since  $D \cup E$  and  $D' \cup E'$  cut up  $V$  and  $W$  into 3-balls, the group  $\mathcal{G}_{\{D \cup E, D' \cup E'\}}$  is identified with the group of isotopy classes of orientation-preserving homeomorphisms of  $\Sigma = \partial V = \partial W$  which preserve  $\partial D \cup \partial E$  and  $\partial D' \cup \partial E'$ . This is the dihedral group  $D_8$  of order 8 with generators  $\rho$  and  $\gamma$  described in Figure 8. The 3-ball in Figure 8, right, is obtained by cutting up  $W$  along  $D' \cup E'$ . Figure 8 gives two descriptions of the



**Figure 10.** Left: The primitive triple  $\{D', E', F'\}$  of  $W$  and the arcs  $(\partial D \cup \partial E \cup \partial F) \cap \partial B$ . Right: The 3-ball  $B$ .



elements  $\alpha$  and  $\gamma$ . Thus we have the presentation  $\langle \rho, \gamma \mid \rho^4 = \gamma^2 = (\rho\gamma)^2 = 1 \rangle$ . We remark that the hyperelliptic involution  $\alpha$  equals  $\rho^2$ .

Next, let  $\{D, E\}$  be a primitive pair of  $V$  in  $L(p, 1)$  with  $p \geq 3$ . There is a unique common dual disk  $E'$  of  $D$  and  $E$  by Theorem 3.3, and hence  $\mathcal{G}_{\{D \cup E\}}$  is identified with the stabilizer subgroup  $\mathcal{G}_{\{D \cup E, E'\}}$ . As in the case of  $\mathcal{G}_{\{E, E' \cup D'\}}$  in the proof of Lemma 5.1, this group is generated by two elements: One is the hyperelliptic involution  $\alpha$ , and the other one is the element, denoted by  $\sigma$ , of order 2 exchanging  $D$  and  $E$  described in Figure 9. Thus we have the presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \sigma \mid \sigma^2 = 1 \rangle$ .

(2) Let  $\{D, E, F\}$  be a primitive triple of  $V$  in  $L(3, 1)$ . Then there exists a unique primitive triple  $\{D', E', F'\}$  of  $W$  as described in Remark 3.5 and Figure 5. Thus the stabilizer subgroup  $\mathcal{G}_{\{D \cup E \cup F\}}$  is identified with  $\mathcal{G}_{\{D \cup E \cup F, D' \cup E' \cup F'\}}$ . The union of three disks  $D' \cup E' \cup F'$  cuts up  $W$  into two 3-balls. One of them, say  $B$ , is shown in Figure 10, right. Consider the group of isotopy classes of orientation-preserving homeomorphisms of  $B$  which preserve  $D' \cup E' \cup F'$  and  $(\partial D \cup \partial E \cup \partial F) \cap \partial B$  on the boundary. This group is the dihedral group  $D_6 = \langle \delta, \gamma' \mid \delta^3 = \gamma'^2 = (\gamma'\delta)^2 = 1 \rangle$  of order 6 with generators  $\delta$  and  $\gamma'$  in Figure 10, right. The element  $\gamma$  in Figure 10, left, is different from  $\gamma'$ , since  $\gamma$  exchanges the two 3-balls. But they are related by  $\gamma = \alpha\gamma'$ , where  $\alpha$  is the hyperelliptic involution exchanging the two 3-balls as described in Figure 10, left. Thus we see that the relation  $(\gamma'\delta)^2 = 1$  in  $D_6$  is equivalent to  $(\gamma\delta)^2 = 1$ . Since the elements  $\alpha$ ,  $\gamma$  and  $\delta$  extend to elements of  $\mathcal{G}_{\{D \cup E \cup F, D' \cup E' \cup F'\}}$ , this group can be considered as the extension of  $D_6$  by  $\langle \alpha \mid \alpha^2 = 1 \rangle$  with relations  $\alpha\gamma\alpha = \gamma$  and  $\alpha\delta\alpha = \delta$ . Thus we have the desired presentation of  $\mathcal{G}_{\{D \cup E \cup F\}}$ .  $\square$

Finally, the stabilizer subgroups of an edge are calculated in a similar way.

**Lemma 5.3.** *An edge of  $\mathcal{T}$  corresponds to the pair of end vertices.*

- (i) *Let  $\{D, E\}$  be a primitive pair of  $V$  in  $L(p, 1)$ . Then  $\mathcal{G}_{\{E, D \cup E\}} = \mathcal{G}_{\{E, D\}}$  has a presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \gamma \mid \gamma^2 = 1 \rangle$  if  $p = 2$ , and a presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle$  if  $p \geq 3$ .*
- (ii) *Let  $\{D, E, F\}$  be a primitive triple of  $V$  in  $L(3, 1)$ . Then  $\mathcal{G}_{\{E, D \cup E \cup F\}} = \mathcal{G}_{\{E, D \cup F\}}$  has a presentation  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \gamma \mid \gamma^2 = 1 \rangle$ .*

Combining Lemmas 5.1, 5.2, and 5.3, we obtain the main result.

**Theorem 5.4.** *The genus-2 Goeritz group  $\mathcal{G}$  of a lens space  $L(p, 1)$  with  $p \geq 2$  has the following presentations:*

- (i)  $\langle \beta, \rho, \gamma \mid \rho^4 = \gamma^2 = (\gamma\rho)^2 = \rho^2\beta\rho^2\beta^{-1} = 1 \rangle$  if  $p = 2$ .
- (ii)  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \beta, \delta, \gamma \mid \delta^3 = \gamma^2 = (\gamma\delta)^2 = 1 \rangle$  if  $p = 3$ .
- (iii)  $\langle \alpha \mid \alpha^2 = 1 \rangle \oplus \langle \beta, \gamma, \sigma \mid \gamma^2 = \sigma^2 = 1 \rangle$  if  $p \geq 4$ .

### Acknowledgments

The author wishes to express his gratitude to Darryl McCullough and Yuya Koda for their helpful discussions, valuable advice, and comments. The main part of this work was carried out while the author was visiting the Korea Institute for Advanced Study (KIAS) in summer 2010. The author is grateful to the institute and to Professor Jaigyoung Choe for their warm hospitality and support. The author is also grateful to the referee for providing helpful comments that improved the presentation.

### References

- [Akbas 2008] E. Akbas, “A presentation for the automorphisms of the 3-sphere that preserve a genus two Heegaard splitting”, *Pacific J. Math.* **236**:2 (2008), 201–222. MR 2009d:57029 Zbl 1157.57002
- [Bonahon 1983] F. Bonahon, “Difféotopies des espaces lenticulaires”, *Topology* **22**:3 (1983), 305–314. MR 85d:57008 Zbl 0526.57009
- [Bonahon and Otal 1983] F. Bonahon and J.-P. Otal, “Scindements de Heegaard des espaces lenticulaires”, *Ann. Sci. École Norm. Sup. (4)* **16**:3 (1983), 451–466. MR 85c:57010 Zbl 0545.57002
- [Cho 2008] S. Cho, “Homeomorphisms of the 3-sphere that preserve a Heegaard splitting of genus two”, *Proc. Amer. Math. Soc.* **136**:3 (2008), 1113–1123. MR 2009c:57029 Zbl 1149.57025
- [Cho and Koda 2012] S. Cho and Y. Koda, “Primitive disk complexes for lens spaces”, preprint, 2012. arXiv 1206.6243
- [Goeritz 1933] L. Goeritz, “Die Abbildungen der Brezelfläche und der Volbrezel vom Geschlecht 2”, *Abh. Math. Sem. Univ. Hamburg* **9**:1 (1933), 244–259. Zbl 0007.08102
- [Gordon 1987] C. M. Gordon, “On primitive sets of loops in the boundary of a handlebody”, *Topology Appl.* **27**:3 (1987), 285–299. MR 88k:57013 Zbl 0634.57007
- [Koda 2011] Y. Koda, “Automorphisms of the 3-sphere that preserve spatial graphs and handlebody-knots”, preprint, 2011. arXiv 1106.4777
- [Osborne and Zieschang 1981] R. P. Osborne and H. Zieschang, “Primitives in the free group on two generators”, *Invent. Math.* **63**:1 (1981), 17–24. MR 82i:20042 Zbl 0438.20017
- [Rolfsen 1976] D. Rolfsen, *Knots and links*, Mathematics Lecture Series **7**, Publish or Perish, Berkeley, CA, 1976. MR 58 #24236 Zbl 0339.55004
- [Scharlemann 2004] M. Scharlemann, “Automorphisms of the 3-sphere that preserve a genus two Heegaard splitting”, *Bol. Soc. Mat. Mexicana* (3) **10**:Special Issue (2004), 503–514. MR 2007c:57020 Zbl 1095.57017
- [Serre 1980] J.-P. Serre, *Trees*, Springer, Berlin, 1980. MR 82c:20083 Zbl 0548.20018
- [Waldhausen 1968] F. Waldhausen, “Heegaard-Zerlegungen der 3-Sphäre”, *Topology* **7** (1968), 195–203. MR 37 #3576 Zbl 0157.54501

Received June 29, 2012. Revised January 9, 2013.

SANGBUM CHO  
 DEPARTMENT OF MATHEMATICS EDUCATION  
 HANYANG UNIVERSITY  
 SEOUL 133-791  
 SOUTH KOREA  
 scho@hanyang.ac.kr

## A COMPACT EMBEDDING THEOREM FOR GENERALIZED SOBOLEV SPACES

SENG-KEE CHUA, SCOTT RODNEY AND RICHARD L. WHEEDEN

**We give an elementary proof of a compact embedding theorem in abstract Sobolev spaces. The result is first presented in a general context and later specialized to the case of degenerate Sobolev spaces defined with respect to nonnegative quadratic forms on  $\mathbb{R}^n$ . Although our primary interest concerns degenerate quadratic forms, our result also applies to nondegenerate cases, and we consider several such applications, including the classical Rellich–Kondrachov compact embedding theorem and results for the class of  $s$ -John domains in  $\mathbb{R}^n$ , the latter for weights equal to powers of the distance to the boundary. We also derive a compactness result for Lebesgue spaces on quasimetric spaces unrelated to  $\mathbb{R}^n$  and possibly without any notion of gradient.**

### 1. The general theorem

The main goal of this paper is to generalize the classical Rellich–Kondrachov theorem concerning compact embedding of Sobolev spaces into Lebesgue spaces. Our principal result applies not only to the classical Sobolev spaces on open sets  $\Omega \subset \mathbb{R}^n$  but also allows us to treat the degenerate Sobolev spaces defined in [Sawyer and Wheeden 2010] and to obtain compact embedding of them into various  $L^q(\Omega)$  spaces. These degenerate Sobolev spaces are associated with quadratic forms  $Q(x, \xi) = \xi' Q(x) \xi$ ,  $x \in \Omega$ ,  $\xi \in \mathbb{R}^n$ , which are nonnegative but may vanish identically in  $\xi$  for some values of  $x$ . Such quadratic forms and Sobolev spaces arise naturally in the study of existence and regularity of weak solutions of some second order subelliptic linear/quasilinear partial differential equations; see, for example, [Sawyer and Wheeden 2006; Rodney 2007; 2012; Monticelli et al. 2012; Rios et al. 2013].

The Rellich–Kondrachov theorem is frequently used to study the existence of solutions to elliptic equations, a famous example being subcritical and critical

---

Chua was partially supported by Singapore Ministry of Education Academic Research Fund Tier 1 R-146-000-150-112. Rodney was partially supported by the Natural Sciences and Engineering Research Council, Canada.

*MSC2010:* primary 46B50, 46E35; secondary 35H20.

*Keywords:* compact embedding, Sobolev spaces, degenerate quadratic forms.

Yamabe equations, resulting in the solution of Yamabe's problem; see [Yamabe 1960; Trudinger 1968; Aubin 1976; Schoen 1984]. Further applications lie in proving the existence of weak solutions to Dirichlet problems for elliptic equations with rough boundary data and coefficients; see [Gilbarg and Trudinger 1997]. In a sequel to this paper, we will apply our compact embedding results to study the existence of solutions for some classes of degenerate equations.

In this section, we state and prove our most general compact embedding results. In Sections 2 and 3, we study some applications to classical and degenerate Sobolev spaces, respectively. In Section 4, more general results in quasimetric spaces are studied.

We begin by listing some useful notation. Let  $w$  be a measure on a  $\sigma$ -algebra  $\Sigma$  of subsets of a set  $\Omega$ , with  $\Omega \in \Sigma$ . For  $0 < p \leq \infty$ , let  $L_w^p(\Omega)$  denote the class of real-valued measurable functions  $f$  satisfying  $\|f\|_{L_w^p(\Omega)} < \infty$ , where  $\|f\|_{L_w^p(\Omega)} = (\int_{\Omega} |f|^p dw)^{1/p}$  if  $p < \infty$  and  $\|f\|_{L_w^\infty(\Omega)} = \text{ess sup}_{\Omega} |f|$ , the essential supremum being taken with respect to  $w$ -measure. When dealing with generic functions in  $L_w^p(\Omega)$ , we will not distinguish between functions which are equal a.e.- $w$ . For  $E \in \Sigma$ ,  $w(E)$  denotes the  $w$ -measure of  $E$ , and if  $0 < w(E) < \infty$ ,  $f_{E,w}$  denotes the  $w$ -average of  $f$  over  $E$ :  $f_{E,w} = \int_E f dw / w(E)$ . Throughout the paper, positive constants are denoted by  $C$  or  $c$  and their dependence on important parameters is indicated.

For  $k \in \mathbb{N}$ , let  $\mathcal{X}(\Omega)$  be a normed linear space of measurable  $\mathbb{R}^k$ -valued functions  $\mathbf{g}$  defined on  $\Omega$  with norm  $\|\mathbf{g}\|_{\mathcal{X}(\Omega)}$ . We assume that there is a subset  $\Sigma_0 \subset \Sigma$  such that  $(\mathcal{X}(\Omega), \Sigma_0)$  satisfies the following properties:

- (A) For any  $\mathbf{g} \in \mathcal{X}(\Omega)$  and  $F \in \Sigma_0$ , the function  $\mathbf{g} \chi_F \in \mathcal{X}(\Omega)$ , where  $\chi_F$  denotes the characteristic function of  $F$ .
- (B <sub>$p$</sub> ) There are constants  $C_1, C_2, p$  satisfying  $1 \leq C_1, C_2, p < \infty$  and such that if  $\{F_l\}$  is a finite collection of sets in  $\Sigma_0$  with  $\sum_l \chi_{F_l}(x) \leq C_1$  for all  $x \in \Omega$ , then

$$\sum_l \|\mathbf{g} \chi_{F_l}\|_{\mathcal{X}(\Omega)}^p \leq C_2 \|\mathbf{g}\|_{\mathcal{X}(\Omega)}^p \quad \text{for all } \mathbf{g} \in \mathcal{X}(\Omega).$$

For  $1 \leq N \leq \infty$ , we will often consider the product space  $L_w^N(\Omega) \times \mathcal{X}(\Omega)$ . This is a normed linear space with norm

$$(1-1) \quad \|(f, \mathbf{g})\|_{L_w^N(\Omega) \times \mathcal{X}(\Omega)} = \|f\|_{L_w^N(\Omega)} + \|\mathbf{g}\|_{\mathcal{X}(\Omega)}.$$

A set  $\mathcal{S} \subset L_w^N(\Omega) \times \mathcal{X}(\Omega)$  will be called a *bounded set in  $L_w^N(\Omega) \times \mathcal{X}(\Omega)$*  if

$$\sup_{(f, \mathbf{g}) \in \mathcal{S}} \|(f, \mathbf{g})\|_{L_w^N(\Omega) \times \mathcal{X}(\Omega)} < \infty.$$

Projection maps such as the one defined by

$$(1-2) \quad \pi : (f, \mathbf{g}) \rightarrow f, \quad (f, \mathbf{g}) \in L_w^N(\Omega) \times \mathcal{X}(\Omega),$$

play a role in our results. If  $w(\Omega) < \infty$ ,  $\pi(L_w^N(\Omega) \times \mathcal{X}(\Omega)) \subset L_w^q(\Omega)$  if  $1 \leq q \leq N$ .

**Theorem 1.1.** *Let  $w$  be a finite measure on a  $\sigma$ -algebra  $\Sigma$  of subsets of a set  $\Omega$ , with  $\Omega \in \Sigma$ . Let  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ ,  $\mathcal{X}(\Omega)$  be a normed linear space satisfying properties (A) and  $(B_p)$  relative to a collection  $\Sigma_0 \subset \Sigma$ , and let  $\mathcal{S}$  be a bounded set in  $L_w^N(\Omega) \times \mathcal{X}(\Omega)$ .*

*Suppose that  $\mathcal{S}$  satisfies the following: given  $\epsilon > 0$ , there are a finite number of pairs  $\{E_l, F_l\}_{l=1}^J$  with  $E_l \in \Sigma$  and  $F_l \in \Sigma_0$  (the pairs and  $J$  may depend on  $\epsilon$ ) satisfying these properties:*

- (i)  $w(\Omega \setminus \bigcup_l E_l) < \epsilon$  and  $w(E_l) > 0$ .
- (ii)  $\{F_l\}$  has bounded overlaps independent of  $\epsilon$  with the same overlap constant as in  $(B_p)$ , that is,

$$(1-3) \quad \sum_{l=1}^J \chi_{F_l}(x) \leq C_1, \quad x \in \Omega,$$

for  $C_1$  as in  $(B_p)$ .

- (iii) For every  $(f, \mathbf{g}) \in \mathcal{S}$ , the local Poincaré-type inequality

$$(1-4) \quad \|f - f_{E_l, w}\|_{L_w^p(E_l)} \leq \epsilon \|\mathbf{g}\|_{\mathcal{X}(\Omega)}$$

holds for each  $(E_l, F_l)$ .

Let  $\hat{\mathcal{S}}$  be the set defined by

$$(1-5) \quad \hat{\mathcal{S}} = \{f \in L_w^N(\Omega) : \text{there exists } \{(f^j, \mathbf{g}^j)\}_{j=1}^\infty \subset \mathcal{S} \text{ with } f^j \rightarrow f \text{ a.e.-}w\}.$$

Then  $\hat{\mathcal{S}}$  is compactly embedded in  $L_w^q(\Omega)$  if  $1 \leq q < N$  in the sense that, for every sequence  $\{f_k\} \subset \hat{\mathcal{S}}$ , there is a single subsequence  $\{f_{k_i}\}$  and a function  $f \in L_w^N(\Omega)$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega$  and in  $L_w^q(\Omega)$  norm for  $1 \leq q < N$ .

Before proceeding with the proof of Theorem 1.1, we make several simple observations. First, in the definition of  $\hat{\mathcal{S}}$ , the property that  $f \in L_w^N(\Omega)$  follows by Fatou's lemma since the associated functions  $f^j$  are bounded in  $L_w^N(\Omega)$ , as  $\mathcal{S}$  is bounded in  $L_w^N(\Omega) \times \mathcal{X}(\Omega)$  by hypothesis. Fatou's lemma also shows that  $\hat{\mathcal{S}}$  is a bounded set in  $L_w^N(\Omega)$ . Moreover, since  $N > 1$ , if  $\{f^j\}$  is bounded in  $L_w^N(\Omega)$  and  $f^j \rightarrow f$  a.e.- $w$ , then  $(f^j)_{E, w} \rightarrow f_{E, w}$  for all  $E \in \Sigma$ ; in fact, in this situation, by using Egorov's theorem, we have  $\int_\Omega f^j \varphi dw \rightarrow \int_\Omega f \varphi dw$  for all  $\varphi \in L_w^{N'}(\Omega)$ ,  $1/N + 1/N' = 1$ .

Next, while the hypothesis  $w(E_l) > 0$  in assumption (i) ensures that the averages  $f_{E_l, w}$  in (1-4) are well-defined, it is not needed since we can discard any pair

$E_l, F_l$  with  $w(E_l) = 0$  without affecting the inequality  $w(\Omega \setminus \bigcup E_l) < \epsilon$  or (1-3) and (1-4).

Finally, since  $\hat{\mathcal{S}}$  contains the first component  $f$  of any pair  $(f, \mathbf{g}) \in \mathcal{S}$ , a simple corollary of Theorem 1.1 is that the projection  $\pi$  defined in (1-2) is a compact mapping of  $\mathcal{S}$  into  $L_w^q(\Omega)$ ,  $1 \leq q < N$ , in the sense that, for every sequence  $\{(f_k, \mathbf{g}_k)\} \subset \mathcal{S}$ , there is a subsequence  $\{f_{k_i}\}$  and a function  $f \in L_w^N(\Omega)$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega$  and in  $L_w^q(\Omega)$  norm for  $1 \leq q < N$ .

*Proof of Theorem 1.1.* Let  $\mathcal{S}$  satisfy the hypotheses and suppose  $\{f_k\}_{k \in \mathbb{N}} \subset \hat{\mathcal{S}}$ . For each  $f_k$ , use the definition of  $\hat{\mathcal{S}}$  to choose a sequence  $\{(f_k^j, \mathbf{g}_k^j)\}_j \subset \mathcal{S}$  with  $f_k^j \rightarrow f_k$  a.e.- $w$  as  $j \rightarrow \infty$ . Since  $\mathcal{S}$  is bounded in  $L_w^N(\Omega) \times \mathcal{X}(\Omega)$ , there is  $M \in (0, \infty)$  such that

$$\|(f_k^j, \mathbf{g}_k^j)\|_{L_w^N(\Omega) \times \mathcal{X}(\Omega)} \leq M$$

for all  $k$  and  $j$ . Also, as noted above,  $\{f_k\}$  is bounded in  $L_w^N(\Omega)$  norm; in fact  $\|f_k\|_{L_w^N(\Omega)} \leq M$  for the same constant  $M$  and all  $k$ .

Since  $\{f_k\}$  is bounded in  $L_w^N(\Omega)$ , if  $1 < N < \infty$ , it has a weakly convergent subsequence, while if  $N = \infty$ , it has a subsequence which converges in the weak-star topology. In either case, we relabel the subsequence as  $\{f_k\}$  to preserve the index. Fix  $\epsilon > 0$  and let  $\{E_l, F_l\}_{l=1}^J$  satisfy the hypotheses of the theorem relative to  $\epsilon$ . Setting  $\Omega^\epsilon = \bigcup E_l$ , we have by assumption (i) that

$$(1-6) \quad w(\Omega \setminus \Omega^\epsilon) < \epsilon.$$

Let us show that there is a positive constant  $C$  independent of  $\epsilon$  such that

$$(1-7) \quad \sum_l \|f_k - (f_k)_{E_l, w}\|_{L_w^p(E_l)}^p \leq C\epsilon^p \quad \text{for all } k.$$

Fix  $k$  and let  $\Delta$  denote the expression on the left side of (1-7). Since

$$f_k^j - (f_k^j)_{E_l, w} \rightarrow f_k - (f_k)_{E_l, w}$$

a.e.- $w$  as  $j \rightarrow \infty$ , Fatou's lemma gives

$$\Delta \leq \sum_l \liminf_{j \rightarrow \infty} \|f_k^j - (f_k^j)_{E_l, w}\|_{L_w^p(E_l)}^p.$$

Consequently, by using the Poincaré inequality (1-4) for  $\mathcal{S}$  and superadditivity of  $\liminf$ , we obtain

$$\Delta \leq \liminf_{j \rightarrow \infty} \sum_l \epsilon^p \|g_k^j \chi_{F_l}\|_{\mathcal{X}(\Omega)}^p.$$

By (1-3), the sets  $F_l$  have finite overlaps uniformly in  $\epsilon$ , with the same overlap constant  $C_1$  as in property (B<sub>p</sub>) of  $\mathcal{X}(\Omega)$ . Hence, by applying property (B<sub>p</sub>) to the last expression together with boundedness of  $\mathcal{S}$ , we get

$$\Delta \leq C_2 \epsilon^p \liminf_{j \rightarrow \infty} \|g_k^j\|_{\mathcal{X}(\Omega)}^p \leq C_2 M^p \epsilon^p.$$

This proves (1-7) with  $C = C_2 M^p$ .

Next note that

$$(1-8) \quad \int_{\Omega^\epsilon} |f_m - f_k|^p dw \leq \sum_l \int_{E_l} |f_m - f_k|^p dw \leq 2^{p-1}(I + II),$$

where

$$I := \sum_l \int_{E_l} |f_m - f_k - (f_m - f_k)_{E_l, w}|^p dw, \quad II := \sum_l |(f_m - f_k)_{E_l, w}|^p w(E_l).$$

We estimate  $I$  and  $II$  separately. We have

$$(1-9) \quad I \leq 2^{p-1} \left( \sum_l \|f_m - (f_m)_{E_l, w}\|_{L_w^p(E_l)}^p + \sum_l \|f_k - (f_k)_{E_l, w}\|_{L_w^p(E_l)}^p \right) \\ \leq 2^{p-1}(C\epsilon^p + C\epsilon^p) = 2^p C\epsilon^p,$$

by (1-7). To estimate  $II$ , first note that

$$II = \sum_{l=1}^J |(f_m - f_k)_{E_l, w}|^p w(E_l) = \sum_{l=1}^J \frac{1}{w(E_l)^{p-1}} \left| \int_{\Omega} (f_m - f_k) \chi_{E_l} dw \right|^p.$$

Since  $w(\Omega) < \infty$ , each characteristic function  $\chi_{E_l} \in L_w^{N'}(\Omega)$ ,  $1/N + 1/N' = 1$  (with  $N' = 1$  if  $N = \infty$ ). As  $\{f_k\}$  converges weakly in  $L_w^N(\Omega)$  when  $1 < N < \infty$  or converges in the weak-star sense when  $N = \infty$ , for  $m, k$  sufficiently large depending on  $\epsilon$ , and for all  $1 \leq l \leq J$ ,

$$\frac{1}{w(E_l)^{p-1}} \left| \int_{\Omega} (f_m - f_k) \chi_{E_l} dw \right|^p \leq \frac{\epsilon^p}{J}.$$

Thus  $II \leq \epsilon^p$  for  $m, k$  sufficiently large depending on  $\epsilon$ . Combining this estimate with (1-8) and (1-9) shows that

$$(1-10) \quad \|f_m - f_k\|_{L_w^p(\Omega^\epsilon)} < C\epsilon$$

for  $m, k$  sufficiently large and  $C = C(M, C_2)$ .

Let us now show that  $\{f_k\}$  is a Cauchy sequence in  $L_w^1(\Omega)$ . For  $m, k$  as in (1-10), Hölder's inequality and the fact that  $\|f_k\|_{L_w^N(\Omega)} \leq M$  for all  $k$  yield

$$\|f_m - f_k\|_{L_w^1(\Omega)} \\ \leq \|f_m - f_k\|_{L_w^1(\Omega^\epsilon)} + \|f_m - f_k\|_{L_w^1(\Omega \setminus \Omega^\epsilon)} \\ \leq \|f_m - f_k\|_{L_w^p(\Omega^\epsilon)} w(\Omega^\epsilon)^{1/p'} + \|f_m - f_k\|_{L_w^N(\Omega \setminus \Omega^\epsilon)} w(\Omega \setminus \Omega^\epsilon)^{1/N'} \\ < C\epsilon w(\Omega^\epsilon)^{1/p'} + 2M w(\Omega \setminus \Omega^\epsilon)^{1/N'} \\ < C\epsilon w(\Omega)^{1/p'} + 2M\epsilon^{1/N'} \quad \text{by (1-6).}$$

Since  $N' < \infty$ , it follows that  $\{f_k\}$  is Cauchy in  $L_w^1(\Omega)$ . Hence it has a subsequence (again denoted by  $\{f_k\}$ ) that converges in  $L_w^1(\Omega)$  and pointwise a.e.- $w$  in  $\Omega$  to a function  $f \in L_w^1(\Omega)$ . If  $N = \infty$ ,  $\{f_k\}$  is bounded in  $L_w^\infty(\Omega)$  by hypothesis, so its pointwise limit  $f \in L_w^\infty(\Omega)$ . If  $N < \infty$ , since  $\{f_k\}$  is bounded in  $L_w^N(\Omega)$ , Fatou's Lemma implies that  $f \in L_w^N(\Omega)$ . This completes the proof in case  $q = 1$ .

For general  $q$ , we use the same subsequence  $\{f_k\}$  as above. Thus we only need to show that  $\{f_k\}$  converges in  $L_w^q(\Omega)$  for  $1 < q < N$ . We use Hölder's inequality. Given  $q \in (1, N)$ , choose  $\lambda \in (0, 1)$ , namely,  $\lambda = (1/q - 1/N)/(1 - 1/N)$ . Hence  $\lambda = 1/q$  if  $N = \infty$ . Then

$$(1-11) \quad \|f_m - f_k\|_{L_w^q(\Omega)} \leq \|f_m - f_k\|_{L_w^1(\Omega)}^\lambda \|f_m - f_k\|_{L_w^N(\Omega)}^{1-\lambda}.$$

As before,  $\|f_k\|_{L_w^N(\Omega)} \leq M$ , and therefore

$$\|f_m - f_k\|_{L_w^q(\Omega)}^{1-\lambda} \leq (2M)^{1-\lambda}.$$

Hence, by (1-11),  $\{f_k\}$  is Cauchy in  $L_w^q(\Omega)$  as it is Cauchy in  $L_w^1(\Omega)$ . This completes the proof of Theorem 1.1.  $\square$

A compact embedding result is also proved in [Franchi et al. 1997, Theorem 3.4] by using Poincaré type estimates. However, Theorem 1.1 applies to situations not considered in [Franchi et al. 1997] since it is not restricted to the context of Lipschitz vector fields in  $\mathbb{R}^n$ . Other abstract compact embedding results can be found in [Hajlasz and Koskela 1998, Theorem 4; Hajlasz and Koskela 2000, Theorem 8.1], including a version [Hajlasz and Koskela 1998, Theorem 5] for weighted Sobolev spaces with nonzero continuous weights, and a version [Hajlasz and Koskela 2000] for metric spaces with a single doubling measure. The proof in [Hajlasz and Koskela 1998] assumes prior knowledge of the classical Rellich–Kondrachov compactness theorem (see, for example, [Gilbarg and Trudinger 1997, Theorem 7.22(i)] and below).

By making minor changes in the proof of Theorem 1.1, we can obtain a sufficient condition for a bounded set in  $L_w^N(\Omega)$  to be precompact in  $L_w^q(\Omega)$ ,  $1 \leq q < N$ , without mentioning the sets  $\{F_I\}$ , the space  $\mathcal{X}(\Omega)$ , properties (A) and (B $_p$ ), or conditions (1-3) and (1-4). We state this result in the next theorem. An application is given in Section 4.

**Theorem 1.2.** *Let  $w$  be a finite measure on a  $\sigma$ -algebra  $\Sigma$  of subsets of a set  $\Omega$ , with  $\Omega \in \Sigma$ . Let  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ , and  $\mathcal{P}$  be a bounded subset of  $L_w^N(\Omega)$ . Suppose there is a positive constant  $C$  such that, for every  $\epsilon > 0$ , there are a finite number of sets  $E_I \in \Sigma$  with*

$$(i) \quad w(\Omega \setminus \bigcup_I E_I) < \epsilon \text{ and } w(E_I) > 0;$$



(ii) for every  $f \in \mathcal{P}$ ,

$$(1-12) \quad \sum_l \|f - f_{E_l, w}\|_{L_w^p(E_l)}^p \leq C\epsilon^p.$$

Let

$$\widehat{\mathcal{P}} = \{f \in L_w^N(\Omega) : \text{there exists } \{f^j\} \subset \mathcal{P} \text{ with } f^j \rightarrow f \text{ a.e.-}w\}.$$

Then, for every sequence  $\{f_k\} \subset \widehat{\mathcal{P}}$ , there is a single subsequence  $\{f_{k_i}\}$  and a function  $f \in L_w^N(\Omega)$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega$  and in  $L_w^q(\Omega)$  norm for  $1 \leq q < N$ .

**Remark 1.3.** (1) Given  $\epsilon > 0$ , let  $\{E_l\}$  satisfy hypothesis (i) of Theorem 1.2. Hypothesis (ii) of Theorem 1.2 is clearly true for  $\{E_l\}$  if, for every  $f \in \mathcal{P}$ , there are nonnegative constants  $\{a_l\}$  such that

$$(1-13) \quad \|f - f_{E_l, w}\|_{L_w^p(E_l)} \leq \epsilon a_l$$

and

$$(1-14) \quad \sum a_l^p \leq C$$

with  $C$  independent of  $f, \epsilon$ . The constants  $\{a_l\}$  may vary with  $f$  and  $\epsilon$ .

(2) Theorem 1.1 is a corollary of Theorem 1.2. To see why, suppose that the hypothesis of Theorem 1.1 holds. Define

$$\mathcal{P} = \pi(\mathcal{S}) = \{f : (f, \mathbf{g}) \in \mathcal{S}\}.$$

Let  $\epsilon > 0$  and choose  $\{(E_l, F_l)\}$  as in Theorem 1.1. Given  $f \in \mathcal{P}$ , choose any  $\mathbf{g}$  such that  $(f, \mathbf{g}) \in \mathcal{S}$  and set  $a_l = \|\mathbf{g} \chi_{F_l}\|_{\mathcal{X}(\Omega)}$  for all  $l$ . Then (1-4), (1-3), and property  $(B_p)$  of  $\mathcal{X}(\Omega)$  imply (1-13) and (1-14). The preceding remark shows that the hypothesis of Theorem 1.2 holds. The conclusion of Theorem 1.1 now follows from Theorem 1.2.

*Proof of Theorem 1.2.* Theorem 1.2 can be proved by checking through the proof of Theorem 1.1. In fact, the nature of hypothesis (1-12) allows simplification of the proof. First recall that if  $f^j \rightarrow f$  a.e.- $w$  and  $\{f^j\}$  is bounded in  $L_w^N(\Omega)$ ,  $(f^j)_{E, w} \rightarrow f_{E, w}$  for every  $E \in \Sigma$ . Therefore, by the definition of  $\widehat{\mathcal{P}}$  and Fatou's lemma, the truth of (1-12) for all  $f \in \mathcal{P}$  implies its truth for all  $f \in \widehat{\mathcal{P}}$ . Given a sequence  $\{f_k\}$  in  $\widehat{\mathcal{P}}$ , we follow the proof of Theorem 1.1, but we no longer need to introduce the  $\{f_k^j\}$  or prove (1-7) since (1-7) now follows from the fact that (1-12) holds for  $\widehat{\mathcal{P}}$ . Further details are left to the reader.  $\square$

We close this section by listing an alternate version of Theorem 1.1 that we use in Section 3D when we consider local results.

**Theorem 1.4.** *Let  $w$  be a measure (not necessarily finite) on a  $\sigma$ -algebra  $\Sigma$  of subsets of a set  $\Omega$ , with  $\Omega \in \Sigma$ . Let  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ ,  $\mathcal{X}(\Omega)$  be a normed linear space satisfying properties (A) and  $(B_p)$  relative to a set  $\Sigma_0 \subset \Sigma$ , and let  $\mathcal{S}$  be a collection of pairs  $(f, \mathbf{g})$  such that  $f$  is  $\Sigma$ -measurable and  $\mathbf{g} \in \mathcal{X}(\Omega)$ .*

*Suppose that  $\mathcal{S}$  satisfies the following conditions relative to a fixed set  $\Omega' \in \Sigma$  (in particular  $\Omega' \subset \Omega$ ): for each  $\epsilon = \epsilon_j = 1/j$  with  $j \in \mathbb{N}$ , there are a finite number of pairs  $\{E_l^\epsilon, F_l^\epsilon\}_l$  with  $E_l^\epsilon \in \Sigma$  and  $F_l^\epsilon \in \Sigma_0$  satisfying these conditions:*

- (i)  $w(\Omega' \setminus \bigcup_l E_l^\epsilon) = 0$  and  $0 < w(E_l^\epsilon) < \infty$ .
- (ii)  $\{F_l^\epsilon\}_l$  has bounded overlaps independent of  $\epsilon$  with the same overlap constant as in  $(B_p)$ , that is,

$$\sum_l \chi_{F_l^\epsilon}(x) \leq C_1, \quad x \in \Omega,$$

for  $C_1$  as in  $(B_p)$ .

- (iii) For every  $(f, \mathbf{g}) \in \mathcal{S}$ , the local Poincaré-type inequality

$$\|f - f_{E_l^\epsilon, w}\|_{L_w^p(E_l^\epsilon)} \leq \epsilon \|\mathbf{g}\|_{\mathcal{X}(\Omega)}$$

holds for each  $(E_l^\epsilon, F_l^\epsilon)$ .

Then, for every sequence  $\{(f_k, \mathbf{g}_k)\}$  in  $\mathcal{S}$  with

$$(1-15) \quad \sup_k [\|f_k\|_{L_w^N(\bigcup_{l,j} E_l^{1/j})} + \|\mathbf{g}_k\|_{\mathcal{X}(\Omega)}] < \infty,$$

there is a subsequence  $\{f_{k_i}\}$  of  $\{f_k\}$  and a function  $f \in L_w^N(\Omega')$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega'$  and in  $L_w^q(\Omega')$  norm for  $1 \leq q \leq p$ . If  $p < N$ , then also  $f_{k_i} \rightarrow f$  in  $L_w^q(\Omega')$  norm for  $1 \leq q < N$ .

The principal difference between the assumptions in Theorems 1.1 and 1.4 occurs in hypothesis (i). When we apply Theorem 1.4 in Section 3D, the sets  $\{E_l^\epsilon\}$  will satisfy  $\Omega' \subset \bigcup_l E_l^\epsilon$  for each  $\epsilon$ , and consequently the condition in hypothesis (i) that  $w(\Omega' \setminus \bigcup_l E_l^\epsilon) = 0$  for each  $\epsilon$  will automatically be true. Unlike Theorem 1.1, the value of  $q$  in Theorem 1.4 is always allowed to equal  $p$ . Although  $w(\Omega)$  is not assumed to be finite in Theorem 1.4,  $w(\Omega') < \infty$  is true due to hypothesis (i) and the fact that the number of  $E_l^\epsilon$  is finite for each  $\epsilon$ . As in Theorem 1.1, the hypothesis  $w(E_l^\epsilon) > 0$  is dispensable.

*Proof of Theorem 1.4.* The proof is like that of Theorem 1.1, with minor changes and some simplifications. We work directly with the pairs  $(f_k, \mathbf{g}_k)$  without considering approximations  $(f_k^j, \mathbf{g}_k^j)$ . Due to the form of assumption (i) in Theorem 1.4, neither the set  $\Omega^\epsilon$  nor estimate (1-6) is now needed. Since  $w(\Omega' \setminus \bigcup_l E_l^\epsilon) = 0$  for each  $\epsilon = 1/j$ , we can replace  $\Omega^\epsilon$  by  $\Omega'$  in the proof, obtaining the estimate

$$(1-16) \quad \|f_m - f_k\|_{L_w^p(\Omega')} < C\epsilon$$

as an analogue of (1-10). In deriving (1-16), the weak and weak-star arguments are guaranteed since, by (1-15),

$$\sup_k \|f_k\|_{L_w^N(\cup_{l,j} E_l^{1/j})} < \infty.$$

The main change in the proof comes by observing that the entire argument formerly used to show that  $\{f_k\}$  is Cauchy in  $L_w^1(\Omega)$  is no longer needed. In fact, (1-16) proves that  $\{f_k\}$  is Cauchy in  $L_w^p(\Omega')$ , and therefore it is also Cauchy in  $L_w^q(\Omega')$  if  $1 \leq q \leq p$  since  $w(\Omega') < \infty$ . The first conclusion in Theorem 1.4 then follows. To prove the second one, assuming that  $p, q < N$ , we use an analogue of (1-11) with  $\Omega'$  in place of  $\Omega$  and the same choice of  $\lambda$ , namely,

$$\|f_m - f_k\|_{L_w^q(\Omega')} \leq \|f_m - f_k\|_{L_w^1(\Omega')}^\lambda \|f_m - f_k\|_{L_w^N(\Omega')}^{1-\lambda}.$$

The desired conclusion then follows as before since we have already shown that the first factor on the right side tends to 0. □

## 2. Applications in the nondegenerate case

Roughly speaking, a consequence of Theorem 1.1 is that a set of functions which is bounded in  $L_w^N(\Omega)$  is precompact in  $L_w^q(\Omega)$  for  $1 \leq q < N$  if the gradients of the functions are bounded in an appropriate norm and a *local* Poincaré inequality holds for them. The requirement of boundedness in  $L_w^N(\Omega)$  will be fulfilled if, for example, the functions satisfy a *global* Poincaré or Sobolev estimate with exponent  $N$  on the left side. In order to illustrate this principle more precisely, we first consider the classical gradient operator and functions on  $\mathbb{R}^n$  with the standard Euclidean metric. We include a simple way to see that the Rellich–Kondrachov compactness theorem follows from our results. Our derivation of this fact is different from those in [Adams and Fournier 2003; Gilbarg and Trudinger 1997]; in particular, it avoids using the Arzelá–Ascoli theorem and regularization of functions by convolution. We also list compactness results for the special class of  $s$ -John domains in  $\mathbb{R}^n$ . Hajlasz and Koskela [1998] mention that such results follow from their development without giving specific statements. See also [Hajlasz and Koskela 2000, Theorem 8.1]. We list results for degenerate quadratic forms and vector fields in Section 3.

We begin by proving a compact embedding result for some Sobolev spaces involving two measures. Let  $w$  be a measure on the Borel subsets of a fixed open set  $\Omega \subset \mathbb{R}^n$ , and let  $\mu$  be a measure on the  $\sigma$ -algebra of Lebesgue measurable subsets of  $\Omega$ . We also assume that  $\mu$  is absolutely continuous with respect to Lebesgue measure. If  $1 \leq p < \infty$ , let  $E_\mu^p(\Omega)$  denote the class of locally Lebesgue integrable functions on  $\Omega$  with distributional derivatives in  $L_\mu^p(\Omega)$ . If  $1 \leq N \leq \infty$ , we say that a set  $Y \subset L_w^N(\Omega) \cap E_\mu^p(\Omega)$  (intersection of function spaces instead of normed

spaces of equivalence classes) is *bounded* in  $L_w^N(\Omega) \cap E_\mu^p(\Omega)$  if

$$\sup_{f \in Y} \{ \|f\|_{L_w^N(\Omega)} + \|\nabla f\|_{L_\mu^p(\Omega)} \} < \infty.$$

We use  $D$  to denote a generic open Euclidean ball. The radius and center of  $D$  will be denoted  $r(D)$  and  $x_D$ , and if  $C$  is a positive constant,  $CD$  will denote the ball concentric with  $D$  whose radius is  $Cr(D)$ .

**Theorem 2.1.** *Let  $\tilde{\Omega} \subset \Omega$  be open sets in  $\mathbb{R}^n$ . Let  $w$  be a Borel measure on  $\Omega$  with  $w(\tilde{\Omega}) = w(\Omega) < \infty$  and  $\mu$  be a measure on the Lebesgue measurable sets in  $\Omega$  which is absolutely continuous with respect to Lebesgue measure. Let  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ , and  $\mathcal{S} \subset L_w^N(\Omega) \cap E_\mu^p(\Omega)$ , and suppose that, for all  $\epsilon > 0$ , there exists  $\delta_\epsilon > 0$  such that*

$$(2-1) \quad \|f - f_D\|_{L_w^p(D)} \leq \epsilon \|\nabla f\|_{L_\mu^p(D)} \quad \text{for all } f \in \mathcal{S}$$

*and all Euclidean balls  $D$  with  $r(D) < \delta_\epsilon$  and  $2D \subset \tilde{\Omega}$ . Then, for any sequence  $\{f_k\} \subset \mathcal{S}$  that is bounded in  $L_w^N(\Omega) \cap E_\mu^p(\Omega)$ , there is a subsequence  $\{f_{k_i}\}$  and a function  $f \in L_w^N(\Omega)$  such that  $\{f_{k_i}\} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega$  and in  $L_w^q(\Omega)$  norm for  $1 \leq q < N$ .*

Before proving Theorem 2.1, we give typical examples of  $\tilde{\Omega}$  and  $w$  with  $w(\tilde{\Omega}) = w(\Omega) < \infty$ . For any two nonempty sets  $E_1, E_2 \subset \mathbb{R}^n$ , let

$$(2-2) \quad \rho(E_1, E_2) = \inf\{|x - y| : x \in E_1, y \in E_2\}$$

denote the Euclidean distance between  $E_1$  and  $E_2$ . If  $x \in \mathbb{R}^n$  and  $E$  is a nonempty set, we write  $\rho(x, E)$  instead of  $\rho(\{x\}, E)$ . Let  $\tilde{\Omega}$  be an open subset of  $\Omega$ . If  $\Omega$  is bounded and  $\Omega \setminus \tilde{\Omega}$  has Lebesgue measure 0, the measure  $w$  on  $\Omega$  defined by  $dw = \rho(x, \mathbb{R}^n \setminus \tilde{\Omega})^\alpha dx$  clearly has the desired properties if  $\alpha \geq 0$ . The range of  $\alpha$  can be increased to  $\alpha > -1$  if  $\Omega$  is a Lipschitz domain and  $\Omega \setminus \tilde{\Omega}$  is a finite set. Indeed, if  $\partial\Omega$  is described in local coordinates  $x = (x_1, \dots, x_n)$  by  $x_n = F(x_1, \dots, x_{n-1})$  with  $F$  Lipschitz, the distance from  $x$  to  $\partial\Omega$  is equivalent to  $|x_n - F(x_1, \dots, x_{n-1})|$ , and, consequently, the restriction  $\alpha > -1$  guarantees that  $w$  is finite near  $\partial\Omega$  by using Fubini's theorem; see also [Chua 1995, Remark 3.4(b)]. If  $\Omega$  is bounded and  $\Omega \setminus \tilde{\Omega}$  is finite, but with no restriction on  $\partial\Omega$ , the range can clearly be further increased to  $\alpha > -n$  for the measure  $\rho(x, \Omega \setminus \tilde{\Omega})^\alpha dx$ . Also note that any  $w$  without point masses satisfies  $w(\tilde{\Omega}) = w(\Omega)$  if  $\tilde{\Omega}$  is obtained by deleting a countable subset of  $\Omega$ .

*Proof of Theorem 2.1.* We verify the hypotheses of Theorem 1.1. Let

$$\mathcal{X}(\Omega) = \left\{ \mathbf{g} = (g_1, \dots, g_n) : |\mathbf{g}| = \left( \sum_{i=1}^n g_i^2 \right)^{1/2} \in L_\mu^p(\Omega) \right\}$$

and  $\|g\|_{\mathcal{X}(\Omega)} = \|g\|_{L_w^p(\Omega)}$ . Then

$$\|\nabla f\|_{\mathcal{X}(\Omega)} = \|\nabla f\|_{L_w^p(\Omega)} \quad \text{if } f \in E_w^p(\Omega).$$

If  $f \in E_w^p(\Omega)$ , we may identify  $f$  with the pair  $(f, \nabla f)$  since the distributional gradient  $\nabla f$  is uniquely determined by  $f$  up to a set of Lebesgue measure zero. Then  $L_w^N(\Omega) \cap E_w^p(\Omega)$  can be viewed as a subset of  $L_w^N(\Omega) \times \mathcal{X}(\Omega)$ . In Theorem 1.1, choose  $\mathcal{S}$  to be the particular sequence  $\{f_k\} \subset \mathcal{S}$  in the hypothesis of Theorem 2.1,  $\Sigma$  to be the Lebesgue measurable subsets of  $\Omega$ , and  $\Sigma_0$  to be the collection of balls  $D \subset \Omega$ . Then hypotheses (A) and (B<sub>p</sub>) are valid with  $C_2 = C_1$  for any  $C_1$ . Given  $\epsilon > 0$ , since  $w(\tilde{\Omega}) = w(\Omega) < \infty$ , there is a compact set  $K \subset \tilde{\Omega}$  with  $w(\Omega \setminus K) < \epsilon$ . Let  $0 < \delta'_\epsilon < \rho(K, \mathbb{R}^n \setminus \tilde{\Omega})$  (where  $\rho(K, \mathbb{R}^n \setminus \tilde{\Omega})$  is interpreted as  $\infty$  if  $\tilde{\Omega} = \mathbb{R}^n$ ). Let  $\delta_\epsilon$  be as in (2-1), and fix  $r_\epsilon$  with  $0 < r_\epsilon < \min\{\delta_\epsilon, \delta'_\epsilon\}$ . By considering the triples of balls in a maximal collection of pairwise disjoint balls of radius  $r_\epsilon/6$  centered in  $K$ , we obtain a collection  $\{E_l^\epsilon\}_l$  of balls of radius  $r_\epsilon/2$  which satisfy  $2E_l^\epsilon \subset \tilde{\Omega}$ , have bounded overlaps with overlap constant independent of  $\epsilon$ , and whose union covers  $K$ . Since  $K$  is compact, we may assume the collection is finite. Also,

$$w\left(\Omega \setminus \bigcup_l E_l^\epsilon\right) \leq w(\Omega \setminus K) < \epsilon,$$

and (1-4) holds with  $F_l = E_l = E_l^\epsilon$  by (2-1). Theorem 2.1 now follows from Theorem 1.1 applied to  $\Omega$ .  $\square$

In particular, we obtain the following result when  $w = \mu$  is a Muckenhoupt  $A_p(\mathbb{R}^n)$  weight, that is, when  $d\mu = dw = \eta dx$ , where  $\eta(x)$  satisfies

$$\left(\frac{1}{|D|} \int_D \eta dx\right) \left(\frac{1}{|D|} \int_D \eta^{-1/(p-1)} dx\right)^{p-1} \leq C$$

if  $1 < p < \infty$ , and satisfies  $|D|^{-1} \int_D \eta dx \leq C \operatorname{ess\,inf}_D w$  if  $p = 1$ , for all Euclidean balls  $D$ , with  $C$  independent of  $D$ . As is well known, such a weight also satisfies the classical doubling condition

$$(2-3) \quad w(D_r(x)) \leq C \left(\frac{r}{r'}\right)^\theta w(D_{r'}(x)), \quad 0 < r' < r < \infty,$$

with  $\theta = np - \epsilon$  for some  $\epsilon > 0$  if  $p > 1$ , and with  $\theta = n$  if  $p = 1$ , where  $C$  and  $\epsilon$  are independent of  $r, r', x$ .

We denote by  $W^{1,p,w}(\Omega)$  the weighted Sobolev space defined as all functions in  $L_w^p(\Omega)$  whose distributional gradient is in  $L_w^p(\Omega)$ . Therefore  $W^{1,p,w}(\Omega) = L_w^p(\Omega) \cap E_w^p(\Omega)$ . If  $w(\Omega) < \infty$ , it follows that  $L_w^N(\Omega) \cap E_w^p(\Omega) \subset W^{1,p,w}(\Omega)$  when  $N \geq p$ , and that the opposite containment holds when  $N \leq p$ .

**Theorem 2.2.** *Let  $1 \leq p < \infty$ ,  $w \in A_p(\mathbb{R}^n)$ , and  $\Omega$  be an open set in  $\mathbb{R}^n$  with  $w(\Omega) < \infty$ . If  $1 < N \leq \infty$ , then any bounded subset of  $L_w^N(\Omega) \cap E_w^p(\Omega)$  is*

precompact in  $L_w^q(\Omega)$  if  $1 \leq q < N$ . Consequently, if  $N > p$  and  $\mathcal{S}$  is a subset of  $W^{1,p,w}(\Omega)$  with

$$(2-4) \quad \|f\|_{L_w^N(\Omega)} \leq C(\|f\|_{L_w^p(\Omega)} + \|\nabla f\|_{L_w^p(\Omega)}) \quad \text{for all } f \in \mathcal{S},$$

then any set in  $\mathcal{S}$  that is bounded in  $W^{1,p,w}(\Omega)$  is precompact in  $L_w^q(\Omega)$  for  $1 \leq q < N$ .

If  $\Omega$  is a John domain, then there exists  $N > p$  ( $N$  can be  $\theta p/(\theta - p)$  for some  $\theta > p$  as described after (2-3)) such that  $W^{1,p,w}(\Omega)$  is compactly embedded in  $L_w^q(\Omega)$  for  $1 \leq q < N$ . In particular, the embedding of  $W^{1,p,w}(\Omega)$  into  $L_w^p(\Omega)$  is compact when  $w \in A_p(\mathbb{R}^n)$  and  $\Omega$  is a John domain.

**Remark 2.3.** When  $w = 1$  and  $p < n$ , the choices  $N = np/(n - p)$  and  $\mathcal{S} = W_0^{1,p}(\Omega)$  — the closure in  $W^{1,p}(\Omega)$  of the class of Lipschitz functions with compact support in  $\Omega$  — guarantee (2-4) by the classical Sobolev inequality for functions in  $W_0^{1,p}(\Omega)$  (see, for example, [Gilbarg and Trudinger 1997, Theorem 7.10]). Consequently, the classical Rellich–Kondrachov theorem giving the compact embedding of  $W_0^{1,p}(\Omega)$  in  $L^q(\Omega)$  for  $1 \leq q < np/(n - p)$  follows as a special case of the first part of Theorem 2.2.

*Proof.* We apply Theorem 2.1 with  $w = \mu$ . Fix  $p$  and  $w$  with  $1 \leq p < \infty$  and  $w \in A_p(\mathbb{R}^n)$ . By [Fabes et al. 1982], there is a constant  $C$  such that the weighted Poincaré inequality

$$\|f - f_{D,w}\|_{L_w^p(D)} \leq Cr(D)\|\nabla f\|_{L_w^p(D)}, \quad f \in C^\infty(\Omega),$$

holds for all Euclidean balls  $D \subset \Omega$ . Then since  $C^\infty(\Omega)$  is dense in  $L_w^N(\Omega) \cap E_w^p(\Omega)$  if  $1 \leq N < \infty$  (see, for example, [Turesson 2000]), by fixing any  $\epsilon > 0$  we obtain from Fatou’s lemma that, for all balls  $D \subset \Omega$  with  $Cr(D) \leq \epsilon$ ,

$$\|f - f_{D,w}\|_{L_w^p(D)} \leq \epsilon \|\nabla f\|_{L_w^p(D)} \quad \text{if } f \in L_w^N(\Omega) \cap E_w^p(\Omega).$$

The same holds when  $N = \infty$  since  $L_w^\infty(\Omega) = L^\infty(\Omega) \subset L_w^p(\Omega)$  due to the assumptions  $w \in A_p(\mathbb{R}^n)$  and  $w(\Omega) < \infty$ . With  $1 < N \leq \infty$ , the first statement of the theorem now follows from Theorem 2.1, and the second statement is a corollary of the first one.

Next, let  $\Omega$  be a John domain. Choose  $\theta > p$  such that  $w$  satisfies (2-3) and define  $N = \theta p/(\theta - p)$ . Then  $N > p$  and, by [Chua and Wheeden 2008, Theorem 1.8(b) or Theorem 4.1],

$$\|f - f_{\Omega,w}\|_{L_w^N(\Omega)} \leq C\|\nabla f\|_{L_w^p(\Omega)}, \quad \forall f \in C^\infty(\Omega).$$

Again, the inequality remains true for functions in  $W^{1,p,w}(\Omega)$  by density and Fatou’s lemma. It is now clear that (2-4) holds, and the last part of the theorem follows.  $\square$

Our next example involves domains in  $\mathbb{R}^n$  which are more restricted. For special  $\Omega$ , there are values  $N > 1$  such that

$$(2-5) \quad \|f\|_{L^N(\Omega)} \leq C(\|f\|_{L^1(\Omega)} + \|\nabla f\|_{L^p(\Omega)})$$

for all  $f \in L^1(\Omega) \cap E^p(\Omega)$ . Note that if  $\Omega$  has finite Lebesgue measure, then  $W^{1,p}(\Omega) \subset L^1(\Omega) \cap E^p(\Omega)$ . As we will explain, (2-5) is true for some  $N > 1$  if  $\Omega$  is an  $s$ -John domain in  $\mathbb{R}^n$  and  $1 \leq s < 1 + p/(n-1)$ . Recall that, for  $1 \leq s < \infty$ , a bounded domain  $\Omega \subset \mathbb{R}^n$  is called an  $s$ -John domain with central point  $x' \in \Omega$  if for some constant  $c > 0$  and all  $x \in \Omega$  with  $x \neq x'$ , there is a curve  $\Gamma : [0, l] \rightarrow \Omega$  such that  $\Gamma(0) = x$ ,  $\Gamma(l) = x'$ ,

$$\begin{aligned} |\Gamma(t_1) - \Gamma(t_2)| &\leq t_2 - t_1 && \text{for all } [t_1, t_2] \subset [0, l], \\ \rho(\Gamma(t), \Omega^c) &\geq ct^s && \text{for all } t \in [0, l]. \end{aligned}$$

The terms 1-John domain and John domain are the same. When  $\Omega$  is an  $s$ -John domain for some  $s \in [1, 1 + p/(n-1))$ , it is shown in [Kilpeläinen and Malý 2000; Chua and Wheeden 2008; 2011] that (2-5) holds for all finite  $N$  with

$$(2-6) \quad \frac{1}{N} \geq \frac{s(n-1) - p + 1}{np}$$

and for all  $f \in W^{1,p}(\Omega)$  without any support restrictions. Note that the right side of (2-6) is strictly less than  $1/p$  for such  $s$ , and consequently there are values  $N > p$  which satisfy (2-6). For  $N$  as in (2-6), the global estimate

$$(2-7) \quad \|f - f_\Omega\|_{L^N(\Omega)} \leq C\|\nabla f\|_{L^p(\Omega)}, \quad f_\Omega = \int_\Omega f(x) dx / |\Omega|,$$

is shown to hold if  $f \in \text{Lip}_{\text{loc}}(\Omega)$  [Chua and Wheeden 2011], and then follows for all  $f \in L^1(\Omega) \cap E^p(\Omega)$ ; see the proof of Theorem 2.4 for related comments. Inequality (2-5) is clearly a consequence of (2-7).

More generally, weighted versions of (2-7) hold for  $s$ -John domains and lead to weighted compactness results, as we now show. Let  $1 \leq p < \infty$ , and, for real  $\alpha$  and  $\rho(x, \Omega^c)$  as in (2-2), let  $L^p_{\rho^\alpha dx}(\Omega)$  be the class of Lebesgue measurable  $f$  on  $\Omega$  with

$$\|f\|_{L^p_{\rho^\alpha dx}(\Omega)} = \left( \int_\Omega |f(x)|^p \rho(x, \Omega^c)^\alpha dx \right)^{1/p} < \infty.$$

**Theorem 2.4.** *Suppose that  $1 \leq s < \infty$  and  $\Omega$  is an  $s$ -John domain in  $\mathbb{R}^n$ . Let  $p, a, b$  satisfy  $1 \leq p < \infty$ ,  $a \geq 0$ ,  $b \in \mathbb{R}$ , and  $b - a < p$ .*

(i) *If*

$$(2-8) \quad n + a > s(n - 1 + b) - p + 1,$$

then, for any  $1 \leq q < \infty$  such that

$$(2-9) \quad \frac{1}{q} > \max \left\{ \frac{1}{p} - \frac{1}{n}, \frac{s(n-1+b)-p+1}{(n+a)p} \right\},$$

$L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega)$  is compactly embedded in  $L^q_{\rho^a dx}(\Omega)$ .

(ii) If  $p > 1$  and

$$(2-10) \quad n + ap > s(n-1+b) - p + 1 \geq n + a,$$

then, for any  $1 \leq q < \infty$  such that

$$(2-11) \quad \frac{a}{q} > \max \left\{ \frac{b}{p} - 1, \frac{s(n-1+b)-p-n+1}{p} \right\},$$

$L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega)$  is compactly embedded in  $L^q_{\rho^a dx}(\Omega)$ .

**Remark 2.5.** (1) If  $a = b = 0$ , (2-8) is the same as  $s < 1 + p/(n-1)$ . If  $a = 0$ , (2-10) never holds.

(2) The requirement that  $b-a < p$  follows from (2-8) and (2-9) by considering the cases  $n-1+b \geq 0$  and  $n-1+b < 0$  separately. Hence  $b-a < p$  automatically holds in Theorem 2.4(i), but it is an assumption in (ii). Also, (2-10) and (2-11) imply that  $q < p$ , and consequently that  $p > 1$ .

(3) Conditions (2-8) and (2-9) imply there exists  $N \in (p, \infty)$  with

$$(2-12) \quad \frac{1}{q} > \frac{1}{N} > \max \left\{ \frac{1}{p} - \frac{1}{n}, \frac{s(n-1+b)-p+1}{(n+a)p} \right\}.$$

Conversely, (2-8) holds if there exists  $N \in (p, \infty)$  such that (2-12) holds.

(4) Assumption (2-11) ensures that there exists  $N \in (q, \infty)$  such that (2-11) holds with  $q$  replaced by  $N$ .

*Proof of Theorem 2.4.* This result is also a consequence of Theorem 2.1, but we deduce it from Theorem 1.1 by using arguments like those in the proofs of Theorems 2.1 and 2.2. Fix  $a, b, p, q$  as in the hypothesis and denote  $\rho(x) = \rho(x, \Omega^c)$ . Choose  $w = \rho^a dx$  and note that  $w(\Omega) < \infty$  since  $a \geq 0$  and  $\Omega$  is now bounded. Define

$$\mathcal{X}(\Omega) = \{ \mathbf{g} = (g_1, \dots, g_n) : |\mathbf{g}| \in L^p_{\rho^b dx}(\Omega) \}$$

and  $\|\mathbf{g}\|_{\mathcal{X}(\Omega)} = \|\mathbf{g}\|_{L^p_{\rho^b dx}(\Omega)}$ . Fix  $\epsilon > 0$  and choose a compact set  $K \subset \Omega$  with

$$|\Omega \setminus K|_{\rho^a dx} := \int_{\Omega \setminus K} \rho^a dx < \epsilon.$$

Also choose  $\delta'_\epsilon$  with  $0 < \delta'_\epsilon < \rho(K, \Omega^c)$ , where  $\rho(K, \Omega^c)$  is the Euclidean distance between  $K$  and  $\Omega^c$ .



If  $D$  is a Euclidean ball with center  $x_D \in K$  and  $r(D) < \frac{1}{2}\delta'_\epsilon$ , then  $2D \subset \Omega$  and  $\rho(x)$  is essentially constant on  $D$ ; in fact, for such  $D$ ,

$$\frac{1}{2}\rho(x_D) \leq \rho(x) \leq \frac{3}{2}\rho(x_D), \quad x \in D.$$

We claim that, for such  $D$ , the simple unweighted Poincaré estimate

$$\|f - f_D\|_{L^p(D)} \leq Cr(D)\|\nabla f\|_{L^p(D)}, \quad f \in \text{Lip}_{\text{loc}}(\Omega),$$

where  $f_D = \int_D f$ , implies that for  $f \in \text{Lip}_{\text{loc}}(\Omega)$ ,

$$(2-13) \quad \|f - f_{D, \rho^a dx}\|_{L^p_{\rho^a dx}(D)} \\ \leq \tilde{C}(r(D))^{(a-b)/p} + \text{diam}(\Omega)^{(a-b)/p} r(D)\|\nabla f\|_{L^p_{\rho^b dx}(D)},$$

where  $f_{D, \rho^a dx} = \int_D f \rho^a dx / \int_D \rho^a dx$  and  $\tilde{C}$  depends on  $C, a, b$  but is independent of  $D, f$ . To show this, first note that, for such  $D$ , since  $\rho \sim \rho(x_D)$  on  $D$ , the simple Poincaré estimate immediately gives

$$\|f - f_D\|_{L^p_{\rho^a dx}(D)} \leq \tilde{C}\rho(x_D)^{(a-b)/p} r(D)\|\nabla f\|_{L^p_{\rho^b dx}(D)}, \quad f \in \text{Lip}_{\text{loc}}(\Omega),$$

and then a similar estimate with  $f_D$  replaced by  $f_{D, \rho^a dx}$  follows by standard arguments. Clearly (2-13) will now follow if we show that

$$\rho(x_D)^{(a-b)/p} \leq r(D)^{(a-b)/p} + \text{diam}(\Omega)^{(a-b)/p} \quad \text{for such } D.$$

However, this is clear since  $r(D) \leq \rho(x_D) \leq \text{diam}(\Omega)$  for  $D$  as above, and (2-13) is proved.

We can now apply the weighted density result of [Hajlasz 1993; Hajlasz and Koskela 1998] to conclude that (2-13) holds for all  $f \in L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega)$  and all balls  $D$  with  $x_D \in K$  and  $r(D) < \frac{1}{2}\delta'_\epsilon$ .

Recall that  $(a-b)/p + 1 > 0$ . Thus there exists  $r_\epsilon$  with  $0 < r_\epsilon < \frac{1}{2}\delta'_\epsilon$  and

$$\tilde{C}(r_\epsilon)^{(a-b)/p} + \text{diam}(\Omega)^{(a-b)/p} r_\epsilon < \epsilon.$$

Let  $\Sigma$  and  $\Sigma_0$  be as in the proof of Theorem 2.1, and let  $\{E_l\}_l = \{F_l\}_l$  be the triples of balls in a maximal collection of pairwise disjoint balls centered in  $K$  with radius  $\frac{1}{3}r_\epsilon$ . Then (2-13) and the choice of  $r_\epsilon$  give the desired version of (1-4), namely

$$\|f - f_{D, \rho^a dx}\|_{L^p_{\rho^a dx}(D)} \leq \epsilon \|\nabla f\|_{L^p_{\rho^b dx}(D)}$$

for  $D = E_l$  and  $f \in L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega)$ . Next, use the last two parts of Remark 2.5 to choose  $N \in (q, \infty)$  such that either (2-9) or (2-11) holds with  $q$

replaced by  $N$ . Every  $f \in L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega)$  then satisfies the global Poincaré estimate

$$(2-14) \quad \|f - f_{\Omega, \rho^a dx}\|_{L^N_{\rho^a dx}(\Omega)} \leq C \|\nabla f\|_{L^p_{\rho^b dx}(\Omega)},$$

$$f \in L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega),$$

where

$$f_{\Omega, \rho^a dx} = \int_{\Omega} f \rho^a dx / \int_{\Omega} \rho^a dx.$$

In fact, under the hypothesis of Theorem 2.4, this is proved for example in [Chua and Wheeden 2011] for  $f \in \text{Lip}_{\text{loc}}(\Omega) \cap L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega)$ , and then follows for all  $f \in L^1_{\rho^a dx}(\Omega) \cap E^p_{\rho^b dx}(\Omega)$  by the density result of [Hajłasz 1993; Hajłasz and Koskela 1998] and Fatou's lemma. By (2-14),

$$\|f\|_{L^N_{\rho^a dx}(\Omega)} \leq C \|f\|_{L^1_{\rho^a dx}(\Omega)} + C \|\nabla f\|_{L^p_{\rho^b dx}(\Omega)}$$

for the same class of  $f$ . The remaining details of the proof are left to the reader.  $\square$

In passing, we mention that the role played by the distance function  $\rho(x, \Omega^c)$  in Theorem 2.4 can instead be played by

$$\rho_0(x) = \inf\{|x - y| : y \in \Omega_0\}, \quad x \in \Omega,$$

for certain  $\Omega_0 \subset \Omega^c$ ; see [Chua and Wheeden 2011, Theorem 1.6] for a description of such  $\Omega_0$  and the required Poincaré estimate, and note that the density result in [Hajłasz and Koskela 1998] holds for positive continuous weights.

### 3. Applications in the degenerate case

In this section,  $\Omega$  will denote a fixed open set in  $\mathbb{R}^n$ , possibly unbounded. For  $(x, \xi) \in \Omega \times \mathbb{R}^n$ , we consider a nonnegative quadratic form  $\xi' Q(x) \xi$  which may degenerate, that is, which may vanish for some  $\xi \neq 0$ . Such quadratic forms occur naturally in the context of subelliptic equations and give rise to degenerate Sobolev spaces as discussed below. Our goal is to apply Theorem 1.1 to obtain compact embedding of these degenerate spaces into Lebesgue spaces related to the gain in integrability provided by Poincaré–Sobolev inequalities. The framework that we use contains the subelliptic one developed in [Sawyer and Wheeden 2006; 2010], where regularity theory for weak solutions of linear subelliptic equations of second order in divergence form is studied.

**3A. Standing assumptions.** We now list some notation and assumptions that will be in force everywhere in Section 3, even when not explicitly mentioned.

**Definition 3.1.** A function  $d$  is called a finite symmetric quasimetric (or simply a quasimetric) on  $\Omega$  if  $d : \Omega \times \Omega \rightarrow [0, \infty)$  and there is a constant  $\kappa \geq 1$  such that, for all  $x, y, z \in \Omega$ ,

$$(3-1) \quad \begin{aligned} d(x, y) &= d(y, x), \\ d(x, y) &= 0 \iff x = y, \\ d(x, y) &\leq \kappa[d(x, z) + d(z, y)]. \end{aligned}$$

If  $d$  is a quasimetric on  $\Omega$ , we refer to the pair  $(\Omega, d)$  as a quasimetric space. In some applications,  $d$  is closely related to  $Q(x)$ . For example,  $d$  is sometimes chosen to be the Carnot–Carathéodory control metric related to  $Q$ ; cf. [Sawyer and Wheeden 2006].

Given  $x \in \Omega$ ,  $r > 0$ , and a quasimetric  $d$ , the subset of  $\Omega$  defined by

$$B_r(x) = \{y \in \Omega : d(x, y) < r\}$$

will be called the quasimetric  $d$ -ball centered at  $x$  of radius  $r$ . Note that every  $d$ -ball  $B = B_r(x)$  satisfies  $B \subset \Omega$  by definition.

It is sometimes possible, and desirable in case the boundary of  $\Omega$  is rough, to be able to work only with  $d$ -balls that are deep inside  $\Omega$  in the sense that their Euclidean closures  $\bar{B}$  lie in  $\Omega$ . See Remark 3.6(ii) for comments about being able to use such balls.

Recall that  $D_s(x)$  denotes the ordinary Euclidean ball of radius  $s$  centered at  $x$ . We always assume that  $d$  is related to the standard Euclidean metric as follows:

$$(3-2) \quad \forall x \in \Omega \text{ and } r > 0, \exists s = s(x, r) > 0 \text{ such that } D_s(x) \subset B_r(x).$$

**Remark 3.2.** Condition (3-2) is clearly true if  $d$ -balls are open, and it is weaker than the well-known condition of C. Fefferman and Phong stating that for each compact  $K \subset \Omega$ , there are constants  $\beta, r_0 > 0$  such that  $D_{r\beta}(x) \subset B_r(x)$  for all  $x \in K$  and  $0 < r < r_0$ .

Throughout Section 3,  $Q(x)$  denotes a fixed Lebesgue measurable  $n \times n$  non-negative symmetric matrix on  $\Omega$  and we assume that every  $d$ -ball  $B$  centered in  $\Omega$  is Lebesgue measurable. We deal with three locally finite measures  $w, \nu, \mu$  on the Lebesgue measurable subsets of  $\Omega$ , each with a particular role. In Section 3C, where only global results are developed, we assume  $w(\Omega) < \infty$ , but this assumption is not required for the local results of Section 3D. The measure  $\mu$  is assumed to be absolutely continuous with respect to Lebesgue measure; the comment following (3-4) explains why this assumption is natural. In Section 3, we sometimes assume that  $w$  is absolutely continuous with respect to  $\nu$ , but we drop this assumption completely in the Appendix.

We do not require the existence of a doubling measure for the collection of  $d$ -balls, but we always assume that  $(\Omega, d)$  satisfies the weaker local geometric doubling property given in the next definition; see [Hytönen and Martikainen 2012] for a global version.

**Definition 3.3.** A quasimetric space  $(\Omega, d)$  satisfies the *local geometric doubling condition* if for every compact  $K \subset \Omega$ , there exists  $\delta' = \delta'(K) > 0$  such that, for all  $x \in K$  and all  $0 < r' < r < \delta'$ , the number of disjoint  $d$ -balls of radius  $r'$  contained in  $B_r(x)$  is at most a constant  $\mathcal{C}_{r/r'}$  depending on  $r/r'$  but not on  $K$ .

**3B. The degenerate Sobolev spaces  $W_{v,\mu}^{1,p}(\Omega, Q)$  and  $W_{v,\mu,0}^{1,p}(\Omega, Q)$ .** We will define weighted degenerate Sobolev spaces by using an approach like the one in [Sawyer and Wheeden 2010] or [Monticelli et al. 2012] for the unweighted case. We first define an appropriate space of vectors, including vectors which will eventually play the role of gradients, where size is measured relative to the nonnegative quadratic form

$$Q(x, \xi) = \xi' Q(x) \xi, \quad (x, \xi) \in \Omega \times \mathbb{R}^n.$$

For  $1 \leq p < \infty$ , consider the collection of measurable  $\mathbb{R}^n$ -valued functions  $\vec{g}(x) = (g_1(x), \dots, g_n(x))$  satisfying

$$(3-3) \quad \begin{aligned} \|\vec{g}\|_{\mathcal{L}_\mu^p(\Omega, Q)} &= \left\{ \int_\Omega Q(x, \vec{g}(x))^{p/2} d\mu \right\}^{1/p} \\ &= \left\{ \int_\Omega |\sqrt{Q(x)} \vec{g}(x)|^p d\mu \right\}^{1/p} \\ &< \infty. \end{aligned}$$

We identify any two functions  $\vec{g}, \vec{h}$  in the collection for which  $\|\vec{g} - \vec{h}\|_{\mathcal{L}_\mu^p(\Omega, Q)} = 0$ . Then (3-3) defines a norm on the resulting space of equivalence classes. The form-weighted space  $\mathcal{L}_\mu^p(\Omega, Q)$  is defined to be the collection of these equivalence classes, with norm (3-3). By using methods similar to those in [Sawyer and Wheeden 2010], it follows that  $\mathcal{L}_\mu^2(\Omega, Q)$  is a Hilbert space and  $\mathcal{L}_\mu^p(\Omega, Q)$  is a Banach space for  $1 \leq p < \infty$ .

Now consider the (possibly infinite) norm on  $\text{Lip}_{\text{loc}}(\Omega)$  defined by

$$(3-4) \quad \|f\|_{W_{v,\mu}^{1,p}(\Omega, Q)} = \|f\|_{L_v^p(\Omega)} + \|\nabla f\|_{\mathcal{L}_\mu^p(\Omega, Q)}.$$

We comment here that our standing assumption that  $\mu(Z) = 0$  when  $Z$  has Lebesgue measure 0 assures that  $\|\nabla f\|_{\mathcal{L}_\mu^p(\Omega, Q)}$  is well-defined if  $f \in \text{Lip}_{\text{loc}}(\Omega)$ ; in fact, for such  $f$ , the Rademacher–Stepanov theorem implies that  $\nabla f$  exists a.e. in  $\Omega$  with respect to Lebesgue measure.

**Definition 3.4.** Let  $1 \leq p < \infty$ .

- (1) The degenerate Sobolev space  $W_{v,\mu}^{1,p}(\Omega, Q)$  is the completion under the norm (3-4) of the set

$$\text{Lip}_{Q,p}(\Omega) = \text{Lip}_{Q,p,v,\mu}(\Omega) = \{f \in \text{Lip}_{\text{loc}}(\Omega) : \|f\|_{W_{v,\mu}^{1,p}(\Omega, Q)} < \infty\}.$$

- (2) The degenerate Sobolev space  $W_{v,\mu,0}^{1,p}(\Omega, Q)$  is the completion under the norm (3-4) of the set  $\text{Lip}_{Q,p,0}(\Omega) = \text{Lip}_0(\Omega) \cap \text{Lip}_{Q,p}(\Omega)$ , where  $\text{Lip}_0(\Omega)$  denotes the collection of Lipschitz functions with compact support in  $\Omega$ . If  $Q \in L_{\text{loc}}^{p/2}(\Omega)$ ,  $\text{Lip}_{Q,p,0}(\Omega) = \text{Lip}_0(\Omega)$  since  $v$  and  $\mu$  are locally finite.

We now make some comments about  $W_{v,\mu}^{1,p}(\Omega, Q)$ , most of which have analogues for  $W_{v,\mu,0}^{1,p}(\Omega, Q)$ . By definition,  $W_{v,\mu}^{1,p}(\Omega, Q)$  is the Banach space of equivalence classes of Cauchy sequences of  $\text{Lip}_{Q,p}(\Omega)$  functions with respect to the norm (3-4). Given a Cauchy sequence  $\{f_j\}$  of  $\text{Lip}_{Q,p}(\Omega)$  functions, denote its equivalence class by  $[\{f_j\}]$ . If  $\{v_j\} \in [\{f_j\}]$ , then  $\{v_j\}$  is a Cauchy sequence in  $L_v^p(\Omega)$  and  $\{\nabla v_j\}$  is a Cauchy sequence in  $\mathcal{L}_\mu^p(\Omega, Q)$ . Hence there is a pair  $(f, \vec{g}) \in L_v^p(\Omega) \times \mathcal{L}_\mu^p(\Omega, Q)$  so that

$$\|v_j - f\|_{L_v^p(\Omega)} \rightarrow 0 \quad \text{and} \quad \|\nabla v_j - \vec{g}\|_{\mathcal{L}_\mu^p(\Omega, Q)} \rightarrow 0$$

as  $j \rightarrow \infty$ . The pair  $(f, \vec{g})$  is uniquely determined by the equivalence class  $[\{f_j\}]$ , that is, it is independent of a particular  $\{v_j\} \in [\{f_j\}]$ . We say that  $(f, \vec{g})$  is *represented* by  $\{v_j\}$ . We obtain a Banach space isomorphism  $\mathcal{F}$  from  $W_{v,\mu}^{1,p}(\Omega, Q)$  onto a closed subspace  $\mathcal{W}_{v,\mu}^{1,p}(\Omega, Q)$  of  $L_v^p(\Omega) \times \mathcal{L}_\mu^p(\Omega, Q)$  by setting

$$(3-5) \quad \mathcal{F}([\{f_j\}]) = (f, \vec{g}).$$

We often do not distinguish between  $W_{v,\mu}^{1,p}(\Omega, Q)$  and  $\mathcal{W}_{v,\mu}^{1,p}(\Omega, Q)$ . Similarly,  $\mathcal{W}_{v,\mu,0}^{1,p}(\Omega, Q)$  denotes the image of  $W_{v,\mu,0}^{1,p}(\Omega, Q)$  under  $\mathcal{F}$ , but we often consider these spaces to be the same.

It is important to think of a typical element of  $\mathcal{W}_{v,\mu}^{1,p}(\Omega, Q)$ , or  $W_{v,\mu}^{1,p}(\Omega, Q)$ , as a pair  $(f, \vec{g})$  as above, and not simply as the first component  $f$ . In fact, if  $(f, \vec{g}) \in \mathcal{W}_{v,\mu}^{1,p}(\Omega, Q)$ , the vector  $\vec{g}$  may not be uniquely determined by  $f$ ; see [Fabes et al. 1982, Section 2.1] for a well-known example.

If  $f \in \text{Lip}_{Q,p}(\Omega)$ , the pair  $(f, \nabla f)$  may be viewed as an element of  $W_{v,\mu}^{1,p}(\Omega, Q)$  by identifying it with the equivalence class  $[\{f\}]$  corresponding to the sequence each of whose entries is  $f$ . When viewed as a class,  $(f, \nabla f)$  generally contains pairs whose first components are not Lipschitz functions; for example, if  $f \in \text{Lip}_{Q,p}(\Omega)$  and  $F$  is any function with  $F = f$  a.e.- $v$ , then  $(f, \nabla f) = (F, \nabla F)$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ . However, in what follows, when we consider a pair  $(f, \nabla f)$  with  $f \in \text{Lip}_{Q,p}(\Omega)$ , we do *not* adopt this point of view. Instead we identify an  $f \in \text{Lip}_{Q,p}(\Omega)$  with the single pair  $(f, \nabla f)$  whose first component is  $f$  (defined everywhere in  $\Omega$ )

and whose second component is  $\nabla f$ , which exists a.e. with respect to Lebesgue measure by the Rademacher–Stepanov theorem. This convention lets us avoid assuming that  $w$  is absolutely continuous with respect to  $\nu$ , written  $w \ll \nu$ , in Poincaré–Sobolev estimates for  $\text{Lip}_{Q,p}(\Omega)$  functions. We reserve the notation  $\mathcal{H}$  for subsets of  $\text{Lip}_{Q,p}(\Omega)$  viewed in this way.

On the other hand,  $\mathcal{W}$  denotes various subsets of  $W_{\nu,\mu}^{1,p}(\Omega, Q)$  with elements viewed as equivalence classes. When our hypotheses are phrased in terms of such  $\mathcal{W}$ , we assume that  $w \ll \nu$  in order to avoid technical difficulties associated with sets of measure 0; see the comment after (3-18). In the Appendix, we drop the assumption  $w \ll \nu$  altogether.

We abuse the notation (3-4) by writing

$$(3-6) \quad \|(f, \nabla f)\|_{W_{\nu,\mu}^{1,p}(\Omega, Q)} = \|f\|_{L_{\nu}^p(\Omega)} + \|\nabla f\|_{\mathcal{L}_{\mu}^p(\Omega, Q)}, \quad f \in \text{Lip}_{Q,p}(\Omega),$$

and we extend this to generic  $(f, \vec{g}) \in W_{\nu,\mu}^{1,p}(\Omega, Q)$  by writing

(3-7)

$$\|(f, \vec{g})\|_{W_{\nu,\mu}^{1,p}(E, Q)} = \|f\|_{L_{\nu}^p(E)} + \|\vec{g}\|_{\mathcal{L}_{\mu}^p(E, Q)} \quad \text{for any measurable } E \subset \Omega.$$

**3C. Global compactness results for degenerate spaces.** In this section, we state and prove compactness results which apply to the entire set  $\Omega$ . Results which are more local are given in Section 3D.

In order to apply Theorem 1.1 in this setting, we use the following version of Poincaré’s inequality for  $d$ -balls.

**Definition 3.5.** Let  $1 \leq p < \infty$ , let  $\text{Lip}_{Q,p}(\Omega)$  be as in Definition 3.4, and let  $\mathcal{H} \subset \text{Lip}_{Q,p}(\Omega)$ . We say that the *Poincaré property of order  $p$  holds for  $\mathcal{H}$*  if there is a constant  $c_0 \geq 1$  such that for every  $\epsilon > 0$  and every compact set  $K \subset \Omega$ , there exists  $\delta = \delta(\epsilon, K) > 0$  such that, for all  $f \in \mathcal{H}$  and every  $d$ -ball  $B_r(y)$  with  $y \in K$  and  $0 < r < \delta$ ,

$$(3-8) \quad \left( \int_{B_r(y)} |f - f_{B_r(y),w}|^p dw \right)^{1/p} \leq \epsilon \|(f, \nabla f)\|_{W_{\nu,\mu}^{1,p}(B_{c_0 r}(y), Q)}.$$

**Remark 3.6.** (i) Inequality (3-8) is not of standard Poincaré form. A more typical form is

$$(3-9) \quad \left( \frac{1}{w(B_r(y))} \int_{B_r(y)} |f - f_{B_r(y),w}|^p dw \right)^{1/p} \leq Cr \left( \frac{1}{\mu(B_{c_0 r}(y))} \int_{B_{c_0 r}(y)} |\sqrt{Q} \nabla f|^p d\mu \right)^{1/p}$$

for some  $c_0 \geq 1$ . In [Sawyer and Wheeden 2006; 2010; Rodney 2007; 2012], the unweighted version of (3-9) with  $p = 2$  is used. Let  $\rho(x, \partial\Omega)$  and  $\rho(E, \partial\Omega)$  be as in (2-2). In [Sawyer and Wheeden 2010], the unweighted form of (3-9) with  $p = 2$

is assumed for all  $f \in \text{Lip}_{Q,2}(\Omega)$  and all  $B_r(y)$  with  $y \in \Omega$  and  $0 < r < \delta_0 \rho(y, \partial\Omega)$  for some  $\delta_0 \in (0, 1)$  independent of  $y, r$ . If  $K$  is a compact set in  $\Omega$ , this version would then hold for all  $B_r(y)$  with  $y \in K$  and  $0 < r < \delta_0 \rho(K, \partial\Omega)$ . For general  $p, w$ , and  $\mu$ , if for every compact  $K \subset \Omega$ , (3-9) is valid for all  $B_r(y)$  with  $y \in K$  and  $0 < r < \delta_0 \rho(K, \partial\Omega)$ , then (3-8) follows easily, provided

$$(3-10) \quad \lim_{r \rightarrow 0} \left\{ \sup_{y \in K} r^p \frac{w(B_r(y))}{\mu(B_{c_0 r}(y))} \right\} = 0$$

for every compact  $K \subset \Omega$ . Note that (3-10) automatically holds if  $w = \mu$ .

If both (3-9) and (3-10) hold, then (3-8) is true for any choice of  $v$ . In this situation, one can pick  $v = w$  in order to avoid technicalities encountered below when  $w$  is not absolutely continuous with respect to  $v$ .

(ii) Especially when  $\partial\Omega$  is rough, it is simplest to deal only with  $d$ -balls  $B$  which stay away from  $\partial\Omega$ , that is, which satisfy

$$(3-11) \quad \bar{B} \subset \Omega.$$

We can always assume this for the balls in (3-8) if the converse of (3-2) is also true, namely, if

$$(3-12) \quad \forall x \in \Omega \text{ and } r > 0, \exists s = s(r, x) > 0 \text{ such that } B_s(x) \subset D_r(x).$$

To see why, let us first show that given a compact set  $K$  and an open set  $G$  with  $K \subset G \subset \Omega$ , there exists  $t > 0$  so that  $\overline{B_t(y)} \subset G$  for all  $y \in K$ . Indeed, for such  $K$  and  $G$ , let  $t' = \frac{1}{2} \rho(K, G^c)$ . By (3-12), for each  $x \in K$ , there exists  $r(x) > 0$  such that  $B_{r(x)}(x) \subset D_{t'}(x)$ . Further, by (3-2), there exists  $s(x) > 0$  such that  $D_{s(x)}(x) \subset B_{r(x)/(2\kappa)}(x)$ , where  $\kappa$  is as in (3-1). Since  $K$  is compact, we may choose finite collections  $\{B_{r_i/(2\kappa)}(x_i)\}$  and  $\{D_{s_i}(x_i)\}$  with  $x_i \in K$ ,  $r_i = r(x_i)$ ,  $s_i = s(x_i)$ , and  $K \subset \bigcup D_{s_i}(x_i) \subset \bigcup B_{r_i/(2\kappa)}(x_i)$ . Now set  $t = \min\{r_i/(2\kappa)\}$ . Let  $y \in K$  and choose  $i$  such that  $y \in B_{r_i/(2\kappa)}(x_i)$ . By (3-1),  $B_t(y) \subset B_{r_i}(x_i)$  and, consequently,  $B_t(y) \subset D_{t'}(x_i)$ . Since  $\overline{D_{t'}(x_i)} \subset G$ , we obtain  $\overline{B_t(y)} \subset G$  for every  $y \in K$ , as desired. In particular,  $\overline{B_t(y)} \subset \Omega$  for all  $y \in K$ . Since the validity of (3-8) for some  $\delta = \delta(\epsilon, K)$  implies its validity for  $\min\{\delta, t\}$ , it follows that we may assume (3-11) for every  $B_r(y)$  in (3-8) when (3-12) holds. Similarly, since the constant  $c_0$  in (3-8) is independent of  $K$ , we may also assume that every  $B_{c_0 r}(y)$  in (3-8) has closure in  $\Omega$ .

(iii) We can often slightly weaken the assumption in Definition 3.5 that  $K$  is an arbitrary compact set in  $\Omega$ . For example, in our results where  $w(\Omega) < \infty$ , it is generally enough to assume that for each  $\epsilon > 0$ , there is a particular compact  $K$  with  $w(\Omega \setminus K) < \epsilon$  such that (3-8) holds. However, in Section 3D, where we do not assume  $w(\Omega) < \infty$ , it is convenient to keep the hypothesis that  $K$  is arbitrary.

Given a set  $\mathcal{H} \subset \text{Lip}_{Q,p}(\Omega)$ , define

$$(3-13) \quad \widehat{\mathcal{H}} = \{f : \text{there exists } \{f^j\} \subset \mathcal{H} \text{ with } f^j \rightarrow f \text{ a.e.-}w\}.$$

It will be useful later to note that if  $\mathcal{H}$  is bounded in  $L_w^N(\Omega)$  for some  $N$ , then  $\widehat{\mathcal{H}}$  is also bounded in  $L_w^N(\Omega)$  by Fatou's lemma; in particular, every  $f \in \widehat{\mathcal{H}}$  then belongs to  $L_w^N(\Omega)$ . See (3-15) for a relationship between  $\widehat{\mathcal{H}}$  and the closure of  $\mathcal{H}$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$  in case  $w \ll v$ .

We now state our simplest global result. Its proof is given after Corollary 3.11.

**Theorem 3.7.** *Let the assumptions of Section 3A hold,  $w(\Omega) < \infty$ ,  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ , and  $\mathcal{H} \subset \text{Lip}_{Q,p}(\Omega)$ . Suppose that the Poincaré property of order  $p$  in Definition 3.5 holds for  $\mathcal{H}$  and that*

$$(3-14) \quad \sup_{f \in \mathcal{H}} \{\|f\|_{L_w^N(\Omega)} + \|f\|_{L_v^p(\Omega)} + \|\nabla f\|_{\mathcal{L}_\mu^p(\Omega, Q)}\} < \infty.$$

*Then any sequence  $\{f_k\} \subset \widehat{\mathcal{H}}$  has a subsequence that converges in  $L_w^q(\Omega)$  norm for every  $1 \leq q < N$  to a function belonging to  $L_w^N(\Omega)$ .*

Let  $\mathcal{H} \subset \text{Lip}_{Q,p}(\Omega)$  and  $\widehat{\mathcal{H}}$  be as in (3-13). We reserve the notation  $\overline{\mathcal{H}}$  for the closure of  $\mathcal{H}$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ , that is, for the closure of the collection

$$\{(f, \nabla f) : f \in \mathcal{H}\}$$

with respect to the norm (3-6). Elements of  $\overline{\mathcal{H}}$  are viewed as equivalence classes. If  $w \ll v$ ,

$$(3-15) \quad \{f : \text{there exists } \vec{g} \text{ such that } (f, \vec{g}) \in \overline{\mathcal{H}}\} \subset \widehat{\mathcal{H}}.$$

Indeed, if  $(f, \vec{g}) \in \overline{\mathcal{H}}$ , there is a sequence  $\{f^j\} \subset \mathcal{H}$  such that  $(f^j, \nabla f^j) \rightarrow (f, \vec{g})$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$  norm, and consequently  $f^j \rightarrow f$  in  $L_v^p(\Omega)$ . By using a subsequence, we may assume that  $f^j \rightarrow f$  pointwise a.e.- $v$ , and hence, by absolute continuity, that  $f^j \rightarrow f$  pointwise a.e.- $w$ . This proves (3-15). In fact, it can be verified by using Egorov's theorem that

$$(3-16) \quad \{f : \text{there exists } \{(f^j, \vec{g}^j)\} \subset \overline{\mathcal{H}} \text{ with } f^j \rightarrow f \text{ a.e.-}w\} \subset \widehat{\mathcal{H}}.$$

Theorem 3.7 and (3-15) immediately imply the following corollary.

**Corollary 3.8.** *Let the assumptions of Section 3A hold,  $w(\Omega) < \infty$ , and  $w \ll v$ . Let  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ ,  $\mathcal{H} \subset \text{Lip}_{Q,p}(\Omega)$ , and  $\overline{\mathcal{H}}$  be the closure of  $\mathcal{H}$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ . Suppose that the Poincaré property of order  $p$  in Definition 3.5 holds for  $\mathcal{H}$  and that*

$$(3-17) \quad \sup_{f \in \mathcal{H}} \{\|f\|_{L_w^N(\Omega)} + \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, Q)}\} < \infty.$$



Then any sequence  $\{f_k\}$  in

$$\{f : \text{there exists } \vec{g} \text{ such that } (f, \vec{g}) \in \overline{\mathcal{H}}\}$$

has a subsequence that converges in  $L_w^q(\Omega)$  norm for  $1 \leq q < N$  to a function that belongs to  $L_w^N(\Omega)$ .

**Remark 3.9.** Corollary 3.8 may be thought of as an analogue in the degenerate setting of the Rellich–Kondrachov theorem since it contains this classical result as a special case. To see why, set  $Q(x) = \text{Id}$  and  $w = v = \mu$  to be Lebesgue measure. Then, given a bounded sequence  $\{(f_k, \vec{g}_k)\} \subset W_0^{1,p}(\Omega) = W_{dx,dx,0}^{1,p}(\Omega, Q)$ , we may choose  $\{f_k^j\} \subset \text{Lip}_0(\Omega)$  with  $(f_k^j, \nabla f_k^j) \rightarrow (f_k, \vec{g}_k)$  in  $W^{1,p}(\Omega)$  norm. Thus, setting  $\mathcal{H} = \{f_k^j\}_{k \in \mathbb{N}, j > J_k}$  where each  $J_k$  is chosen sufficiently large to preserve boundedness, the classical Sobolev inequality gives (3-17) with  $N = np/(n-p)$  for  $1 \leq p < n$ . The Rellich–Kondrachov theorem now follows from Corollary 3.8.

We next mention analogues of these results when  $\mathcal{H}$  is replaced by a set

$$\mathfrak{W} \subset W_{v,\mu}^{1,p}(\Omega, Q)$$

with elements viewed as equivalence classes, assuming that  $w \ll v$ . We then modify Definition 3.5 by replacing (3-8) with the analogous estimate

$$(3-18) \quad \left( \int_{B_r(y)} |f - f_{B_r(y),w}|^p dw \right)^{1/p} \leq \epsilon \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(B_{c_0 r}(y), Q)} \quad \text{if } (f, \vec{g}) \in \mathfrak{W}.$$

The assumption  $w \ll v$  guarantees that the left side of (3-18) does not change when the first component of a pair is arbitrarily altered in a set of  $v$ -measure zero.

If Poincaré’s inequality is known to hold for subsets of Lipschitz functions in the form (3-8), it can often be extended by approximation to the similar form (3-18) for subsets of  $W_{v,\mu}^{1,p}(\Omega, Q)$ . Indeed, let us show without using weak convergence that if  $w \ll v$  and the Radon–Nikodym derivative  $dw/dv \in L_v^{p'}(\Omega)$ ,  $1/p + 1/p' = 1$ , then (3-18) holds with  $\mathfrak{W} = W_{v,\mu}^{1,p}(\Omega, Q)$  if (3-8) holds with  $\mathcal{H} = \text{Lip}_{Q,p}(\Omega)$ . This follows easily from Fatou’s lemma since if  $(f, \vec{g}) \in W_{v,\mu}^{1,p}(\Omega, Q)$  and we choose  $\{f_j\} \subset \text{Lip}_{Q,p}(\Omega)$  with  $(f_j, \nabla f_j) \rightarrow (f, \vec{g})$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ , then, for any ball  $B$ , since  $f_j \rightarrow f$  in  $L_v^p(\Omega)$ , we have

$$(f_j)_{B,w} = \frac{1}{w(B)} \int_B f_j \frac{dw}{dv} dv \rightarrow \frac{1}{w(B)} \int_B f \frac{dw}{dv} dv = f_{B,w}.$$

Of course we may also assume that  $f_j \rightarrow f$  a.e.- $w$  by selecting a subsequence of  $\{f_j\}$  which converges to  $f$  a.e.- $v$ . The same argument shows that if (3-18) holds for all pairs in any set  $\mathfrak{W} \subset W_{v,\mu}^{1,p}(\Omega, Q)$ , then it also holds for pairs in the closure  $\overline{\mathfrak{W}}$  of  $\mathfrak{W}$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ . Moreover, if all balls  $B$  in question satisfy  $\overline{B} \subset \Omega$  (cf. (3-11)), the assumption can clearly be weakened to  $dw/dv \in L_{v,\text{loc}}^{p'}(\Omega)$ . As we observed in

Remark 3.6(ii), the balls in (3-8) can be assumed to satisfy (3-11) provided (3-12) is true.

Analogues of Theorem 3.7 and Corollary 3.8 for a set  $\mathcal{W} \subset W_{v,\mu}^{1,p}(\Omega, Q)$  are given in the next result, which also includes the Rellich–Kondrachov theorem as a special case.

**Theorem 3.10.** *Let the assumptions of Section 3A hold,  $w(\Omega) < \infty$ , and  $w \ll v$ . Let  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ , and  $\mathcal{W} \subset W_{v,\mu}^{1,p}(\Omega, Q)$ . Suppose that the Poincaré property in Definition 3.5 holds, but in the modified form given in (3-18), and that*

$$(3-19) \quad \sup_{(f,\vec{g}) \in \mathcal{W}} \{ \|f\|_{L_w^N(\Omega)} + \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(\Omega, Q)} \} < \infty.$$

Let

$$\widehat{\mathcal{W}} = \{f : \text{there exists } \{(f^j, \vec{g}^j)\} \subset \mathcal{W} \text{ with } f^j \rightarrow f \text{ a.e.} -w\}.$$

Then any sequence in  $\widehat{\mathcal{W}}$  has a subsequence that converges in  $L_w^q(\Omega)$  norm for every  $1 \leq q < N$  to a function belonging to  $L_w^N(\Omega)$ . In particular, if  $\overline{\mathcal{W}}$  denotes the closure of  $\mathcal{W}$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ , the same is true for any sequence in

$$\{f : \text{there exists } \vec{g} \text{ such that } (f, \vec{g}) \in \overline{\mathcal{W}}\}.$$

As a corollary, we obtain a result for arbitrary sequences  $\{(f_k, \vec{g}_k)\}$  which are bounded in  $W_{v,\mu}^{1,p}(\Omega, Q)$  and whose first components  $\{f_k\}$  are bounded in  $L_w^N(\Omega)$ .

**Corollary 3.11.** *Let the assumptions of Section 3A hold,  $w(\Omega) < \infty$ ,  $w \ll v$ ,  $1 \leq p < \infty$ , and  $1 < N \leq \infty$ . Suppose that the Poincaré property in Definition 3.5 holds for all of  $W_{v,\mu}^{1,p}(\Omega, Q)$ , that is, Definition 3.5 holds with (3-8) replaced by (3-18) for  $\mathcal{W} = W_{v,\mu}^{1,p}(\Omega, Q)$ . Then if  $\{(f_k, \vec{g}_k)\}$  is any sequence in  $W_{v,\mu}^{1,p}(\Omega, Q)$  such that*

$$\sup_k [\|f_k\|_{L_w^N(\Omega)} + \|(f_k, \vec{g}_k)\|_{W_{v,\mu}^{1,p}(\Omega, Q)}] < \infty,$$

there is a subsequence of  $\{f_k\}$  that converges in  $L_w^q(\Omega)$  norm for  $1 \leq q < N$  to a function belonging to  $L_w^N(\Omega)$ . If in addition  $dw/dv \in L_v^{p'}(\Omega)$ ,  $1/p + 1/p' = 1$ , the conclusion remains valid if the Poincaré property holds just for  $\text{Lip}_{Q,p}(\Omega)$ .

In fact, the first conclusion in Corollary 3.11 follows by applying Theorem 3.10 with  $\mathcal{W}$  chosen to be the specific sequence  $\{(f_k, \vec{g}_k)\}_k$  in question, and the second statement follows from the first one and our observation above that (3-18) holds with  $\mathcal{W} = W_{v,\mu}^{1,p}(\Omega, Q)$  if  $dw/dv \in L_v^{p'}(\Omega)$ ,  $1/p + 1/p' = 1$ , and if (3-8) holds with  $\mathcal{H} = \text{Lip}_{Q,p}(\Omega)$ .

*Proofs of Theorems 3.7 and 3.10.* We will concentrate on the proof of Theorem 3.7. The proof of Theorem 3.10 is similar and omitted. We begin with a useful covering lemma.

**Lemma 3.12.** *Let the assumptions of Section 3A hold and let  $w(\Omega) < \infty$ . Fix  $p \in [1, \infty)$  and a set  $\mathcal{H} \subset \text{Lip}_{\mathcal{Q},p}(\Omega)$ . Suppose the Poincaré property of order  $p$  in Definition 3.5 holds for  $\mathcal{H}$ , and let  $\kappa$  be as in (3-1) and  $c_0$  be as in (3-8). Then, for every  $\epsilon > 0$ , there are positive constants  $r = r(\epsilon, \kappa, c_0)$ ,  $M = M(\kappa, c_0)$ , and a finite collection  $\{B_r(y_k)\}_k$  of  $d$ -balls, such that*

$$(3-20) \quad w(\Omega \setminus \bigcup_k B_r(y_k)) < \epsilon,$$

$$(3-21) \quad \sum_k \chi_{B_{c_0 r}(y_k)}(x) \leq M \quad \text{for all } x \in \Omega,$$

$$(3-22) \quad \|f - f_{B_r(y_k),w}\|_{L_w^p(B_r(y_k))} \leq \epsilon \| (f, \nabla f) \|_{W_{v,\mu}^{1,p}(B_{c_0 r}(y_k), \mathcal{Q})}$$

for all  $f \in \mathcal{H}$  and all  $k$ . Note that  $M$  is independent of  $\epsilon$ .

*Proof.* We first recall the “swallowing” property of  $d$ -balls: there is a constant  $\gamma \geq 1$  depending only on  $\kappa$  such that if  $x, y \in \Omega$ ,  $0 < r_1 \leq r_2 < \infty$  and  $B_{r_1}(x) \cap B_{r_2}(y) \neq \emptyset$ , then

$$(3-23) \quad B_{r_1}(x) \subset B_{\gamma r_2}(y).$$

Indeed, by [Chua and Wheeden 2008, Observation 2.1],  $\gamma$  can be chosen to be  $\kappa + 2\kappa^2$ .

Fix  $\epsilon > 0$ . Since  $w(\Omega) < \infty$ , there is a compact set  $K \subset \Omega$  with  $w(\Omega \setminus K) < \epsilon$ . Let  $\delta' = \delta'(\epsilon)$  be as in Definition 3.3 for  $K$ , and let  $\delta = \delta(\epsilon)$  be as in (3-8). Fix  $r$  with  $0 < r < \min\{\delta, \delta'/(c_0\gamma)\}$  where  $c_0$  is as in (3-8). For each  $x \in K$ , use (3-2) to pick  $s(x, r) > 0$  so that  $D_{s(x,r)}(x) \subset B_{r/\gamma}(x)$ . Since  $K$  is compact, there are finitely many points  $\{x_j\}$  in  $K$  such that  $K \subset \bigcup_j B_{r/\gamma}(x_j)$ . Choose a maximal pairwise disjoint subcollection  $\{B_{r/\gamma}(y_k)\}$  of  $\{B_{r/\gamma}(x_j)\}$ . We show that the collection  $\{B_r(y_k)\}$  satisfies (3-20)–(3-22).

To verify (3-20), it is enough to show that  $K \subset \bigcup_k B_r(y_k)$ . Let  $y \in K$ . Then  $y \in B_{r/\gamma}(x_j)$  for some  $x_j$ . If  $x_j = y_k$  for some  $y_k$  then  $y \in B_r(y_k)$ . If  $x_j \neq y_k$  for all  $y_k$ , there exists  $y_l$  such that  $B_{r/\gamma}(y_l) \cap B_{r/\gamma}(x_j) \neq \emptyset$ . Then  $B_{r/\gamma}(x_j) \subset B_r(y_l)$  by (3-23), and so  $y \in B_r(y_l)$ . In either case, we obtain  $y \in \bigcup_k B_r(y_k)$  as desired.

To verify (3-21), suppose that  $\{k_i\}_{i=1}^L$  satisfies  $\bigcap_{i=1}^L B_{c_0 r}(y_{k_i}) \neq \emptyset$ . Then, by (3-23),  $B_{c_0 r}(y_{k_i}) \subset B_{c_0 \gamma r}(y_{k_1})$  for  $1 \leq i \leq L$ . Since  $\gamma, c_0 \geq 1$ , we have

$$B_{r/\gamma}(y_k) \subset B_{c_0 r}(y_k)$$

for all  $k$ , and consequently

$$\bigcup B_{r/\gamma}(y_{k_i}) \subset \bigcup B_{c_0 r}(y_{k_i}) \subset B_{c_0 \gamma r}(y_{k_1}).$$

By construction,  $\{B_{r/\gamma}(y_k)\}$  is pairwise disjoint in  $k$ . Since  $0 < r/\gamma < c_0 \gamma r < \delta'$ , the corresponding constant  $\mathcal{C}$  in the definition of geometric doubling depends only on  $(c_0 \gamma r)/(r/\gamma) = c_0 \gamma^2$ , that is,  $\mathcal{C}$  depends only on  $\kappa$  and  $c_0$ . Choosing  $M$  to be

this constant, we obtain that  $L \leq M$  as desired. The same argument shows that the collection  $\{B_{c_0 r}(y_k)\}$  has the stronger bounded intercept property with the same bound  $M$ , that is, any ball in the collection intersects at most  $M - 1$  others.

Finally, let us verify (3-22). Recall that  $0 < r < \delta$  by construction. Hence (3-8) implies that for each  $k$  and all  $f \in \mathcal{H}$ ,

$$(3-24) \quad \|f - f_{B_r(y_k), w}\|_{L_w^p(B_r(y_k))} \leq \epsilon \|(f, \nabla f)\|_{W_{v, \mu}^{1, p}(B_{c_0 r}(y_k), Q)}. \quad \square$$

We deduce the proof of Theorem 3.7 from Theorem 1.1 by choosing  $\mathcal{X}(\Omega) = L_v^p(\Omega) \times \mathcal{L}_\mu^p(\Omega, Q)$  and considering the product space

$$\mathfrak{B}_{N, \mathcal{X}(\Omega)} = L_w^N(\Omega) \times (L_v^p(\Omega) \times \mathcal{L}_\mu^p(\Omega, Q)).$$

We always choose  $\Sigma$  to be the Lebesgue measurable subsets of  $\Omega$  and

$$\Sigma_0 = \{B_r(x) : r > 0, x \in \Omega\}.$$

Note that  $\mathcal{X}(\Omega)$  and  $\mathfrak{B}_{N, \mathcal{X}(\Omega)}$  are normed linear spaces (even Banach spaces), and the norm in  $\mathfrak{B}_{N, \mathcal{X}(\Omega)}$  is

$$(3-25) \quad \|(h, (f, \vec{g}))\|_{\mathfrak{B}_{N, \mathcal{X}(\Omega)}} = \|h\|_{L_w^N(\Omega)} + \|f\|_{L_v^p(\Omega)} + \|\vec{g}\|_{\mathcal{L}_\mu^p(\Omega, Q)}.$$

The roles of  $\mathbf{g}$  and  $(f, \mathbf{g})$  in Section 1 are now played by  $(f, \vec{g})$  and  $(h, (f, \vec{g}))$  respectively.

Let us verify properties (A) and (B<sub>p</sub>) in Section 1 with  $\mathcal{X}(\Omega)$  and  $\Sigma_0$  chosen as above. To verify (A), fix  $B \in \Sigma_0$  and  $(f, \vec{g}) \in \mathcal{X}(\Omega)$ . Clearly  $f\chi_B \in L_v^p(\Omega)$  since  $f \in L_v^p(\Omega)$ . Also,

$$\begin{aligned} \int_{\Omega} ((\vec{g}\chi_B)' Q(\vec{g}\chi_B))^{p/2} d\mu &= \int_B (\vec{g}' Q(x) \vec{g})^{p/2} d\mu \\ &\leq \int_{\Omega} (\vec{g}' Q(x) \vec{g})^{p/2} d\mu < \infty. \end{aligned}$$

Thus  $(f, \vec{g})\chi_B \in \mathcal{X}(\Omega)$  and property (A) is proved.

To verify (B<sub>p</sub>), let  $\{B_l\}$  be a finite collection of  $d$ -balls satisfying  $\sum_l \chi_{B_l}(x) \leq C_1$  for all  $x \in \Omega$ . Then if  $(f, \vec{g}) \in \mathcal{X}(\Omega)$ ,

$$\begin{aligned} \sum_l \|(f, \vec{g})\chi_{B_l}\|_{\mathcal{X}(\Omega)}^p &= \sum_l (\|f\chi_{B_l}\|_{L_v^p(\Omega)} + \|\vec{g}\chi_{B_l}\|_{\mathcal{L}_\mu^p(\Omega, Q)})^p \\ &\leq 2^{p-1} \sum_l (\|f\chi_{B_l}\|_{L_v^p(\Omega)}^p + \|\vec{g}\chi_{B_l}\|_{\mathcal{L}_\mu^p(\Omega, Q)}^p) \\ &= 2^{p-1} \int_{\Omega} |f|^p \left( \sum_l \chi_{B_l} \right) dv + \int_{\Omega} (\vec{g}' Q \vec{g})^{p/2} \left( \sum_l \chi_{B_l} \right) d\mu \\ &\leq 2^{p-1} C_1 (\|f\|_{L_v^p(\Omega)}^p + \|\vec{g}\|_{\mathcal{L}_\mu^p(\Omega, Q)}^p) \leq 2^p C_1 \|(f, \vec{g})\|_{\mathcal{X}(\Omega)}^p. \end{aligned}$$

This verifies  $(B_p)$  with  $C_2$  chosen to be  $2^p C_1$ .

The proof of Theorem 3.7 is now very simple. Let  $\mathcal{H}$  satisfy its hypotheses and choose  $\mathcal{S}$  in Theorem 1.1 to be the set

$$\mathcal{S} = \{(f, (f, \nabla f)) : f \in \mathcal{H}\}.$$

Note that  $\mathcal{S}$  is a bounded subset of  $\mathcal{B}_{N, \mathcal{H}}(\Omega)$  by hypothesis (3-14). Next, in order to choose the pairs  $\{E_l, F_l\}_l$  and verify conditions (i)–(iii) of Theorem 1.1 (see (1-3) and (1-4)), we appeal to Lemma 3.12. Given  $\epsilon > 0$ , let  $\{E_l, F_l\}_l = \{B_r(y_k), B_{c_0 r}(y_k)\}_k$  where  $\{y_k\}$  and  $r$  are as in Lemma 3.12. Then  $E_l, F_l \in \Sigma_0$ , and conditions (i)–(iii) of Theorem 1.1 are guaranteed by Lemma 3.12. Finally, by noting that the set  $\widehat{\mathcal{H}}$  defined in (3-13) is the same as the set  $\widehat{\mathcal{S}}$  defined in (1-5), the conclusion of Theorem 3.7 follows from Theorem 1.1.  $\square$

For special domains  $\Omega$  and special choices of  $N$ , the boundedness assumption (3-14) (or (3-17)) can be weakened to

$$(3-26) \quad \sup_{f \in \mathcal{H}} \{\|f\|_{L_v^p(\Omega)} + \|\nabla f\|_{\mathcal{L}_\mu^p(\Omega, \mathcal{Q})}\} = \sup_{f \in \mathcal{H}} \|(f, \nabla f)\|_{W_{v, \mu}^{1,p}(\Omega, \mathcal{Q})} < \infty.$$

This is clearly the case for any  $\Omega$  and  $N$  for which there exists a global Sobolev–Poincaré estimate that bounds  $\|f\|_{L_w^N(\Omega)}$  by

$$\|(f, \nabla f)\|_{W_{v, \mu}^{1,p}(\Omega, \mathcal{Q})}$$

for all  $f \in \mathcal{H}$ . We now formalize this situation assuming that  $w \ll v$ . In the Appendix, we consider a case when  $w \ll v$  fails.

The form of the global Sobolev–Poincaré estimate we will use is given in the next definition. It guarantees that (3-14) and (3-26) are the same when  $N = p\sigma$ .

**Definition 3.13.** Let  $1 \leq p < \infty$  and  $\mathcal{H} \subset \text{Lip}_{\mathcal{Q}, p}(\Omega)$ . Then the *global Sobolev property of order  $p$*  holds for  $\mathcal{H}$  if there are constants  $C > 0$  and  $\sigma > 1$  such that

$$(3-27) \quad \|f\|_{L_w^{p\sigma}(\Omega)} \leq C \|(f, \nabla f)\|_{W_{v, \mu}^{1,p}(\Omega, \mathcal{Q})} \quad \text{for all } f \in \mathcal{H}.$$

If  $w \ll v$ , (3-27) extends to  $(f, \vec{g}) \in \overline{\mathcal{H}}$ . In fact, let  $(f, \vec{g}) \in \overline{\mathcal{H}}$  and choose  $\{f_j\} \subset \mathcal{H}$  with  $(f_j, \nabla f_j) \rightarrow (f, \vec{g})$  in  $W_{v, \mu}^{1,p}(\Omega, \mathcal{Q})$ . Then  $f_j \rightarrow f$  in  $L_v^p(\Omega)$  norm, and by choosing a subsequence we may assume that  $f_j \rightarrow f$  a.e.- $v$ . Hence  $f_j \rightarrow f$  a.e.- $w$  because  $w \ll v$ . Since each  $f_j$  satisfies (3-27), it follows that

$$(3-28) \quad \|f\|_{L_w^{p\sigma}(\Omega)} \leq C \|(f, \vec{g})\|_{W_{v, \mu}^{1,p}(\Omega, \mathcal{Q})} \quad \text{if } (f, \vec{g}) \in \overline{\mathcal{H}}.$$

Under the same assumptions, namely, that Definition 3.13 holds for a set

$$\mathcal{H} \subset \text{Lip}_{\mathcal{Q}, p}(\Omega)$$

and that  $w \ll v$ , the same sequence  $\{f_j\}$  as above is also bounded in  $L_w^{p\sigma}(\Omega)$  norm and so satisfies  $(f_j)_{E,w} \rightarrow f_{E,w}$  for measurable  $E$  by the same weak convergence argument given after the statement of Theorem 1.1. Hence the Poincaré estimate in Definition 3.5 also extends to  $\bar{\mathcal{H}}$  in the same form as (3-18), with  $\mathcal{W}$  replaced by  $\bar{\mathcal{H}}$ , that is,

$$(3-29) \quad \left( \int_{B_r(y)} |f - f_{B_r(y),w}|^p dw \right)^{1/p} \leq \epsilon \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(B_{c_0 r}(y), Q)} \text{ if } (f, \vec{g}) \in \bar{\mathcal{H}}.$$

Hence we immediately obtain the next result by choosing  $\mathcal{W} = \bar{\mathcal{H}}$  and  $N = p\sigma$  in Theorem 3.10.

**Theorem 3.14.** *Let the assumptions of Section 3A hold,  $w(\Omega) < \infty$ , and  $w \ll v$ . Fix  $p \in [1, \infty)$  and a set  $\mathcal{H} \subset \text{Lip}_{Q,p}(\Omega)$ . Suppose the Poincaré and global Sobolev properties of order  $p$  in Definitions 3.5 and 3.13 hold for  $\mathcal{H}$ , and let  $\sigma$  be as in (3-27). If  $\{(f_k, \vec{g}_k)\}$  is a sequence in  $\bar{\mathcal{H}}$  with*

$$(3-30) \quad \sup_k \|(f_k, \vec{g}_k)\|_{W_{v,\mu}^{1,p}(\Omega, Q)} < \infty,$$

*then  $\{f_k\}$  has a subsequence which converges in  $L_w^q(\Omega)$  for  $1 \leq q < p\sigma$ , and the limit of the subsequence belongs to  $L_w^{p\sigma}(\Omega)$ .*

A result for the entire space  $W_{v,\mu}^{1,p}(\Omega, Q)$  follows by choosing  $\mathcal{H} = \text{Lip}_{Q,p}(\Omega)$  in Theorem 3.14 or Corollary 3.8:

**Corollary 3.15.** *Suppose the hypotheses of Theorem 3.14 hold with  $\mathcal{H} = \text{Lip}_{Q,p}(\Omega)$ . If  $\{(f_k, \vec{g}_k)\} \subset W_{v,\mu}^{1,p}(\Omega, Q)$  and (3-30) is true,  $\{f_k\}$  has a subsequence which converges in  $L_w^q(\Omega)$  for  $1 \leq q < p\sigma$ , and the limit of the subsequence belongs to  $L_w^{p\sigma}(\Omega)$ .*

See the Appendix for analogues of Theorem 3.14 and Corollary 3.15 without the assumption  $w \ll v$ .

**3D. Local compactness results for degenerate spaces.** In this section, for general bounded open sets  $\Omega'$  with  $\bar{\Omega}' \subset \Omega$ , we study compact embedding of subsets of  $W_{v,\mu}^{1,p}(\Omega, Q)$  into  $L_w^q(\Omega')$  without assuming a global Sobolev estimate for  $\Omega$  or  $\Omega'$  and without assuming  $w(\Omega) < \infty$ . For some applications, see the comment at the end of the section.

The theorems below assume a much weaker condition than the global Sobolev estimate (3-27), namely, the following local estimate.

**Definition 3.16.** Let  $1 \leq p < \infty$ . We say that the *local Sobolev property of order  $p$*  holds if, for some fixed constant  $\sigma > 1$  and every compact set  $K \subset \Omega$ , there is a constant  $r_1 > 0$  such that, for all  $d$ -balls  $B = B_r(y)$  with  $y \in K$  and  $0 < r < r_1$ ,

$$(3-31) \quad \|f\|_{L_w^{p\sigma}(B)} \leq C(B) \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, Q)} \text{ if } f \in \text{Lip}_0(\Omega) \text{ with } \text{supp } f \subset B,$$

where  $C(B)$  is a positive constant independent of  $f$ .

**Remark 3.17.** (i) A more standard assumption than (3-31) is a normalized inequality that includes a factor  $r$  in the gradient term on the right side:

$$(3-32) \quad \left( \frac{1}{w(B_r(y))} \int_{B_r(y)} |f|^{p\sigma} dw \right)^{1/(p\sigma)} \\ \leq C \left( \frac{1}{v(B_r(y))} \int_{B_r(y)} |f|^p dv \right)^{1/p} + Cr \left( \frac{1}{\mu(B_r(y))} \int_{B_r(y)} |\sqrt{Q} \nabla f|^p d\mu \right)^{1/p},$$

with  $C$  independent of  $r, y$ ; see, for example, [Sawyer and Wheeden 2006; Rodney 2007; 2012] in the unweighted case with  $p = 2$ . Clearly (3-32) is a stronger requirement than (3-31).

(ii) In the classical  $n$ -dimensional elliptic case for linear second order equations in divergence form,  $Q$  satisfies  $c|\xi|^2 \leq Q(x, \xi) \leq C|\xi|^2$  for some fixed constants  $c, C > 0$  and  $d$  is the standard Euclidean metric  $d(x, y) = |x - y|$ . For  $1 \leq p < n$  and  $\sigma = n/(n - p)$ , (3-31) then holds with  $dw = dv = d\mu = dx$  since the corresponding version of (3-32) is true with  $|\sqrt{Q} \nabla f|$  replaced by  $|\nabla f|$ .

We also use a notion of Lipschitz cutoff functions on  $d$ -balls:

**Definition 3.18.** For  $s \geq 1$ , we say that the *cutoff property of order  $s$*  holds for  $\mu$  if, for each compact  $K \subset \Omega$ , there exists  $\delta = \delta(K) > 0$  such that, for every  $d$ -ball  $B_r(y)$  with  $y \in K$  and  $0 < r < \delta$ , there is a function  $\phi \in \text{Lip}_0(\Omega)$  and a constant  $\gamma = \gamma(y, r) \in (0, r)$  satisfying

- (i)  $0 \leq \phi \leq 1$  in  $\Omega$ ,
- (ii)  $\text{supp } \phi \subset B_r(y)$  and  $\phi = 1$  in  $B_\gamma(y)$ ,
- (iii)  $\nabla \phi \in \mathcal{L}_\mu^s(\Omega, Q)$ .

Since  $\mu$  is always assumed to be locally finite, the strongest form of Definition 3.18, namely, the version with  $s = \infty$ , automatically holds if  $Q$  is locally bounded in  $\Omega$  and (3-12) is true; recall that we always assume (3-2). To see why, fix a compact set  $K \subset \Omega$  and consider  $B_r(y)$  with  $y \in K$  and  $r < 1$ . Use (3-2) to choose open Euclidean balls  $D', D$  with common center  $y$  such that  $\bar{D}' \subset D \subset B_r(y)$  ( $\subset \Omega$  by definition). Construct a smooth function  $\phi$  in  $\Omega$  with support in  $D$  such that  $0 \leq \phi \leq 1$  and  $\phi = 1$  on  $D'$ . By (3-12), there is  $\gamma > 0$  such that  $B_\gamma(y) \subset D'$ . Then  $\phi$  satisfies Definition 3.18(i)–(iii) with  $s = \infty$ ; for (iii), we use the fact that  $\nabla \phi$  has compact support in  $\Omega$  together with local boundedness of  $Q$  and local finiteness of  $\mu$ .

To compensate for the lack of a global Sobolev estimate, given  $\mathcal{H} \subset \text{Lip}_{Q,p}(\Omega)$ , we assume in conjunction with the cutoff property of some order  $s \geq p\sigma'$  that, for every compact set  $K \subset \Omega$ , there exists  $\delta = \delta(K) > 0$  such that, for every  $d$ -ball  $B$

with center in  $K$  and radius less than  $\delta$ , there is a constant  $C_1(B)$  such that

$$(3-33) \quad \|f\|_{L_{\mu}^{p t'}(B)} \leq C_1(B) \|(f, \nabla f)\|_{W_{v, \mu}^{1, p}(\Omega, Q)} \quad \text{if } f \in \mathcal{H},$$

where  $t = s/p$  and  $1/t + 1/t' = 1$ . Note that  $1 \leq t' \leq \sigma$  since  $s \geq p\sigma'$ .

**Remark 3.19.** Inequality (3-33) is different in nature from (3-31) even if  $t' = \sigma$  and  $w = \mu$  since there is a restriction on supports in (3-31) but not in (3-33). However, (3-33) implies (3-31) when  $s = p\sigma'$ ,  $w = \mu$ , and  $\mathcal{H}$  contains all functions in  $\text{Lip}_0(\Omega)$  with support in any ball. On the other hand, (3-33) is often automatic if  $\mu = v$ . For example, as mentioned earlier, if  $Q$  is locally bounded and (3-12) is true, the cutoff property holds with  $s = \infty$ , giving  $t = \infty$  and  $t' = 1$ . In this case, when  $\mu = v$ , the left side of (3-33) is clearly smaller than the right side (in fact smaller than  $\|f\|_{L_v^p(\Omega)}$ ).

We can now state our main local result.

**Theorem 3.20.** *Let the assumptions of Section 3A and condition (3-12) hold, and let  $w \ll v$ . Fix  $p \in [1, \infty)$  and suppose the Poincaré property of order  $p$  in Definition 3.5 holds for a fixed set  $\mathcal{H} \subset \text{Lip}_{Q, p}(\Omega)$  and the local Sobolev property of order  $p$  in Definition 3.16 holds. Assume the cutoff property of some order  $s \geq p\sigma'$  is true for  $\mu$ , with  $\sigma$  as in (3-31), and that (3-33) holds for  $\mathcal{H}$  with  $t = s/p$ . Then, for every  $\{(f_k, \vec{g}_k)\} \subset \bar{\mathcal{H}}$  that is bounded in  $W_{v, \mu}^{1, p}(\Omega, Q)$  norm, there is a subsequence  $\{f_{k_i}\}$  of  $\{f_k\}$  and an  $f \in L_{w, \text{loc}}^{p\sigma}(\Omega)$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega$  and in  $L_w^q(\Omega')$  norm for all  $1 \leq q < p\sigma$  and every bounded open  $\Omega'$  with  $\bar{\Omega}' \subset \Omega$ .*

See the Appendix for a version of Theorem 3.20 without assuming  $w \ll v$ .

Recall that  $\bar{\mathcal{H}} = W_{v, \mu}^{1, p}(\Omega, Q)$  if  $\mathcal{H} = \text{Lip}_{Q, p}(\Omega)$ . In the important case when  $Q \in L_{\text{loc}}^q(\Omega)$ , Theorem 3.20 and Remark 3.19 immediately imply the next result.

**Corollary 3.21.** *Let  $Q$  be locally bounded in  $\Omega$  and suppose that (3-12) holds. Fix  $p \in [1, \infty)$ , and with  $w = v = \mu$ , assume the Poincaré property of order  $p$  holds for  $\text{Lip}_{Q, p}(\Omega)$  and the local Sobolev property of order  $p$  holds. Then, for every bounded sequence  $\{(f_k, \vec{g}_k)\} \subset W_{w, w}^{1, p}(\Omega, Q)$ , there is a subsequence  $\{f_{k_i}\}$  of  $\{f_k\}$  and a function  $f \in L_{w, \text{loc}}^{p\sigma}(\Omega)$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega$  and in  $L_w^q(\Omega')$  norm,  $1 \leq q < p\sigma$ , for every bounded open  $\Omega'$  with  $\bar{\Omega}' \subset \Omega$ .*

*Proof of Theorem 3.20.* We begin by using the cutoff property in Definition 3.18 to construct a partition of unity relative to  $d$ -balls and compact subsets of  $\Omega$ .

**Lemma 3.22.** *Fix  $\Omega$  and  $s \geq 1$ , and suppose the cutoff property of order  $s$  holds for  $\mu$ . If  $K$  is a compact subset of  $\Omega$  and  $r > 0$ , there is a finite collection of  $d$ -balls  $\{B_r(y_j)\}$  with  $y_j \in K$  together with functions  $\{\psi_j\}$  in  $\text{Lip}_0(\Omega)$  such that  $\text{supp } \psi_j \subset B_r(y_j)$  and*



- (a)  $K \subset \bigcup_j B_r(y_j)$ ,
- (b)  $0 \leq \psi_j \leq 1$  in  $\Omega$  for each  $j$  and  $\sum_j \psi_j(x) = 1$  for all  $x \in K$ ,
- (c)  $\nabla \psi_j \in \mathcal{L}_\mu^s(\Omega, Q)$  for each  $j$ .

*Proof.* The argument is an adaptation of one in [Rudin 1987] for the usual Euclidean case. The authors thank D. D. Monticelli for related discussions. Fix  $r > 0$  and a compact set  $K \subset \Omega$ , and set  $\beta = \min\{\delta/2, r\}$  for  $\delta = \delta(K)$  as in Definition 3.18. Since  $\beta < \delta$ , Definition 3.18 implies that, for each  $y \in K$ , there exist  $\gamma(y) \in (0, \beta)$  and  $\phi_y(x) \in \text{Lip}_0(\Omega)$  such that  $0 \leq \phi_y \leq 1$  in  $\Omega$ ,  $\text{supp } \phi_y \subset B_\beta(y)$ ,  $\phi_y = 1$  in  $B_{\gamma(y)}(y)$  and  $\nabla \phi_y \in \mathcal{L}_\mu^s(\Omega, Q)$ . The collection  $\{B_{\gamma(y)}(y)\}_{y \in K}$  covers  $K$ , so by (3-2) and the compactness of  $K$ , there is a finite subcollection  $\{B_{\gamma(y_j)}(y_j)\}_{j=1}^m$  whose union covers  $K$ . Part (a) follows since  $\gamma(y_j) < r$ . Next let  $\phi_j(x) = \phi_{y_j}(x)$  and define  $\{\psi_j\}_{j=1}^m$  as follows: set  $\psi_1 = \phi_1$  and

$$\psi_j = (1 - \phi_1) \cdots (1 - \phi_{j-1}) \phi_j$$

for  $j = 2, \dots, m$ . Then each  $\psi_j \in \text{Lip}_0(\Omega)$ , and  $\text{supp } \phi_j \subset B_r(y_j)$  since  $\beta < r$ . Also,  $0 \leq \psi_j \leq 1$  in  $\Omega$  and

$$\sum_{j=1}^m \psi_j(x) = 1 - \prod_{j=1}^m (1 - \phi_j(x)), \quad x \in \Omega.$$

If  $x \in K$ ,  $x \in B_{\gamma(y_j)}(y_j)$  for some  $j$ . Hence some  $\phi_j(x) = 1$  and consequently  $\sum_j \psi_j(x) = 1$ . This proves part (b). Lastly, we use Leibniz's product rule to compute  $\nabla \psi_j$  and then apply Minkowski's inequality  $j$  times to obtain part (c) from the fact that  $\nabla \phi_j \in \mathcal{L}_\mu^s(\Omega, Q)$ .  $\square$

The next lemma shows how the local Sobolev estimate (3-31) and Lemma 3.22 lead to a local analogue of the global Sobolev estimate (3-27).

**Lemma 3.23.** *Let  $\Omega'$  be a bounded open set with  $\overline{\Omega'} \subset \Omega$ . Suppose that both Definition 3.16 and the cutoff property for  $\mu$  of some order  $s \geq p\sigma'$  hold, and also that (3-33) holds with  $t = s/p$  for a fixed set  $\mathcal{H} \subset \text{Lip}_{\text{loc}}(\Omega)$ . Then there is a finite constant  $C(\Omega')$  such that*

$$(3-34) \quad \|f\|_{L_w^{p\sigma'}(\Omega')} \leq C(\Omega') \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, Q)} \quad \text{if } f \in \mathcal{H}.$$

*Proof.* Let  $r_1$  be as in Definition 3.16 relative to the compact set  $\overline{\Omega'} \subset \Omega$ , and let  $\delta$  be as in (3-33). Use Lemma 3.22 to cover  $\overline{\Omega'}$  by the union of a finite number of  $d$ -balls  $\{B_j\}$  each of radius smaller than  $\min\{r_1, \delta\}$ . Associated with this cover is a collection  $\{\psi_j\} \subset \text{Lip}_0(\Omega)$  with  $\text{supp } \psi_j \subset B_j$ ,  $\sum_j \psi_j = 1$  in  $\Omega'$ , and

$$\nabla \psi_j \in \mathcal{L}_\mu^s(\Omega, Q).$$

If  $f \in \mathcal{H}$ , then

$$(3-35) \quad \|f\|_{L_w^{p\sigma}(\Omega')} = \left\| f \sum_j \psi_j \right\|_{L_w^{p\sigma}(\Omega')} \leq \sum_j \|\psi_j f\|_{L_w^{p\sigma}(B_j)}.$$

Since  $\psi_j f \in \text{Lip}_0(\Omega)$  and  $\text{supp}(\psi_j f) \subset B_j$ , the estimate (3-31) and the product rule give

$$(3-36) \quad \begin{aligned} \|\psi_j f\|_{L_w^{p\sigma}(B_j)} &\leq C(B_j) \|(\psi_j f, \nabla(\psi_j f))\|_{W_{v,\mu}^{1,p}(B_j, \mathcal{Q})} \\ &= C(B_j) (\|\psi_j f\|_{L_v^p(B_j)} + \|\sqrt{Q}\nabla(\psi_j f)\|_{L_\mu^p(B_j)}) \\ &\leq C(B_j) (\|\psi_j f\|_{L_v^p(B_j)} + \|\psi_j \sqrt{Q}\nabla f\|_{L_\mu^p(B_j)} + \|f \sqrt{Q}\nabla \psi_j\|_{L_\mu^p(B_j)}) \\ &\leq C(B_j) (\|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, \mathcal{Q})} + \|f \sqrt{Q}\nabla \psi_j\|_{L_\mu^p(B_j)}), \end{aligned}$$

where we have used  $|\psi_j| \leq 1$ . We estimate the second term on the right of (3-36) by using (3-33). Recall that  $t = s/p \geq \sigma'$  and  $1/t + 1/t' = 1$ . Let

$$\bar{C} = \max_j \|\sqrt{Q}\nabla \psi_j\|_{L_\mu^s(B_j)}.$$

By Hölder's inequality and (3-33),

$$(3-37) \quad \begin{aligned} \|f \sqrt{Q}\nabla \psi_j\|_{L_\mu^p(B_j)} &\leq \|f\|_{L_\mu^{p t'}(B_j)} \|\sqrt{Q}\nabla \psi_j\|_{L_\mu^s(B_j)} \\ &\leq \bar{C} C_1(B_j) \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, \mathcal{Q})}. \end{aligned}$$

Combining this with (3-36) gives

$$\|\psi_j f\|_{L_w^{p\sigma}(B_j)} \leq C(B_j) (1 + \bar{C} C_1(B_j)) \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, \mathcal{Q})}.$$

By (3-35), for any  $f \in \mathcal{H}$ ,

$$\begin{aligned} \|f\|_{L_w^{p\sigma}(\Omega')} &\leq \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, \mathcal{Q})} \sum_j C(B_j) (1 + \bar{C} C_1(B_j)) \\ &= C(\Omega') \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, \mathcal{Q})}. \end{aligned} \quad \square$$

Theorem 3.20 follows from Lemma 3.23 and Theorem 1.4. We sketch the proof, omitting some familiar details. By choosing a sequence of compact sets increasing to  $\Omega$  and using a diagonalization argument, it is enough to prove the conclusion for a fixed open  $\Omega'$  with compact closure  $\bar{\Omega}'$  in  $\Omega$ . Fix such an  $\Omega'$  and select a bounded open  $\Omega''$  with  $\bar{\Omega}' \subset \Omega'' \subset \bar{\Omega}'' \subset \Omega$ . For  $\mathcal{H}$  as in Theorem 3.20, apply Lemma 3.23 to the set  $\Omega''$  to obtain

$$(3-38) \quad \|f\|_{L_w^{p\sigma}(\Omega'')} \leq C(\Omega'') \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, \mathcal{Q})}, \quad f \in \mathcal{H}.$$

By assumption,  $w \ll v$ , so (3-38) extends to  $\bar{\mathcal{H}}$  in the form

$$(3-39) \quad \|f\|_{L_w^{p\sigma}(\Omega'')} \leq C(\Omega'') \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(\Omega, Q)}, \quad (f, \vec{g}) \in \bar{\mathcal{H}}.$$

Let  $\epsilon > 0$ . By hypothesis,  $\mathcal{H}$  satisfies the Poincaré estimate (3-8) for balls  $B_r(y)$  with  $y \in \bar{\Omega}'$  and  $r < \delta(\epsilon, \Omega')$ . Since the Euclidean distance between  $\bar{\Omega}'$  and  $\partial\Omega''$  is positive and we have assumed (3-12), we may also assume by Remark 3.6(ii) that all such balls lie in the larger set  $\Omega''$ . Next we claim that (3-8) extends to  $\bar{\mathcal{H}}$ , that is,

$$(3-40) \quad \left( \int_{B_r(y)} |f - f_{B_r(y),w}|^p dw \right)^{1/p} \leq \epsilon \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(B_{c_0r}(y), Q)} \quad \text{if } (f, \vec{g}) \in \bar{\mathcal{H}},$$

for the same class of balls  $B_r(y)$ . In fact, if  $(f, \vec{g}) \in \bar{\mathcal{H}}$  and  $\{f^j\} \subset \mathcal{H}$  satisfies  $(f^j, \nabla f^j) \rightarrow (f, \vec{g})$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$  norm, then there is a subsequence, still denoted  $\{f^j\}$ , with  $f^j \rightarrow f$  a.e.- $v$  in  $\Omega$ , and so with  $f^j \rightarrow f$  a.e.- $w$  in  $\Omega$  since  $w \ll v$ . By (3-38),  $\{f^j\}$  is bounded in  $L_w^{p\sigma}(\Omega'')$ . Hence, since the balls in (3-40) satisfy  $B_r(y) \subset \Omega''$ , we obtain  $f_{B_r(y),w}^j \rightarrow f_{B_r(y),w}$  by our usual weak convergence argument, and (3-40) follows by Fatou's lemma from its analogue (3-8) for the  $(f^j, \nabla f^j)$ .

Now let  $\{(f_k, \vec{g}_k)\} \subset \bar{\mathcal{H}}$  be bounded in  $W_{v,\mu}^{1,p}(\Omega, Q)$  norm and apply Theorem 1.4 with  $\mathcal{X}(\Omega) = L_v^p(\Omega) \times \mathcal{L}_\mu^p(\Omega, Q)$  to the set  $\mathcal{S}$  defined by

$$\mathcal{S} = \{(f_k, (f_k, \vec{g}_k))\}_k,$$

and with  $\{(E_l^\epsilon, F_l^\epsilon)\}_l$  chosen to be a finite number of pairs  $\{(B_r(y_l), B_{c_0r}(y_l))\}_l$  as in (3-40), but now with  $r$  fixed depending on  $\epsilon$ , and with  $\Omega' \subset \bigcup_l B_r(y_l)$ . Such a finite choice exists by (3-2) and the Heine–Borel theorem since  $\bar{\Omega}'$  is compact; cf. the proof of Lemma 3.12. Since  $\Omega'$  is completely covered by  $\bigcup_l E_l^\epsilon$ , assumption (i) of Theorem 1.4 is fulfilled. Moreover, the collection  $\{F_l^\epsilon\}$  has bounded overlaps uniformly in  $\epsilon$  by the geometric doubling argument used to prove Lemma 3.12.

Finally, (1-15) follows from (3-39) applied to the bounded sequence  $\{(f_k, \vec{g}_k)\}$  since  $\bigcup_{l,\epsilon} E_l^\epsilon \subset \Omega''$ . Thus Theorem 1.4 implies that there is a subsequence  $\{f_{k_i}\}$  of  $\{f_k\}$  and a function  $f \in L_w^{p\sigma}(\Omega')$  such that  $f_{k_i} \rightarrow f$  a.e.- $w$  in  $\Omega'$  and in  $L_w^q(\Omega')$  norm,  $1 \leq q < p\sigma$ . This completes the proof of Theorem 3.20.  $\square$

For functions which are compactly supported in a fixed bounded open  $\Omega'$  with  $\bar{\Omega}' \subset \Omega$ , the proof of Theorem 3.20 can be modified to yield compact embedding into  $L_w^q(\Omega')$  for the same  $\Omega'$  without assuming (3-12). Of course we always require (3-2). Given such  $\Omega'$  and a set  $\mathcal{H} \subset \text{Lip}_{Q,p,0}(\Omega')$ , we may view  $\mathcal{H}$  as a subset of  $\text{Lip}_{Q,p,0}(\Omega)$  simply by extending functions in  $\mathcal{H}$  to all of  $\Omega$  as 0 in  $\Omega \setminus \Omega'$ . In this way, the proof of Theorem 3.20 works without (3-12). For example, choosing  $\mathcal{H} = \text{Lip}_{Q,p,0}(\Omega')$ , we obtain:

**Theorem 3.24.** *Let the assumptions of Section 3A hold and  $w \ll v$ . Let  $\Omega'$  be a bounded open set with  $\bar{\Omega}' \subset \Omega$ . Fix  $p \in [1, \infty)$  and suppose the Poincaré property of order  $p$  in Definition 3.5 holds for  $\text{Lip}_{Q,p,0}(\Omega')$ , with  $\text{Lip}_{Q,p,0}(\Omega')$  viewed as a subset of  $\text{Lip}_{Q,p,0}(\Omega)$  using extension by 0, and suppose the local Sobolev property of order  $p$  in Definition 3.16 holds. Assume the cutoff property of some order  $s \geq p\sigma'$  is true for  $\mu$ , with  $\sigma$  as in (3-31), and that (3-33) holds for  $\text{Lip}_{Q,p,0}(\Omega')$  with  $t = s/p$ . Then, for every sequence  $\{(f_k, \vec{g}_k)\} \subset W_{v,\mu,0}^{1,p}(\Omega', Q)$  which is bounded in  $W_{v,\mu}^{1,p}(\Omega', Q)$  norm, there is a subsequence  $\{f_{k_i}\}$  of  $\{f_k\}$  and a function  $f \in L_w^{p\sigma}(\Omega')$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega'$  and in  $L_w^q(\Omega')$  norm,  $1 \leq q < p\sigma$ .*

The full force of the local Sobolev estimate in Definition 3.16 is not needed to prove Theorem 3.24. In fact, it is enough to assume that (3-31) holds only for balls centered in the fixed compact set  $\bar{\Omega}'$ .

The proof of Theorem 3.24 is like that of Theorem 3.20, working with the set  $\Omega'$  that occurs in the hypotheses of Theorem 3.24. However, now (3-34) in the conclusion of Lemma 3.23 (with  $\mathcal{H} = \text{Lip}_{Q,p,0}(\Omega')$ ) remains valid if  $\Omega'$  is replaced on the left side by  $\Omega$  since every  $f \in \text{Lip}_{Q,p,0}(\Omega')$  vanishes on  $\Omega \setminus \Omega'$ . The resulting estimate serves as a replacement for (3-38), so it is not necessary to demand that the  $E_j^\epsilon$  are subsets of a compact set  $\bar{\Omega}'' \subset \Omega$ . Hence (3-12) is no longer required. Finally, the Poincaré estimate extends as usual to  $W_{v,\mu,0}^{1,p}(\Omega', Q)$  (the closure of  $\text{Lip}_{Q,p,0}(\Omega')$ ), and due to support considerations, the  $E_j^\epsilon$  can be restricted to subsets of  $\Omega'$  by replacing  $E_j^\epsilon$  by  $E_j^\epsilon \cap \Omega'$ ; this guarantees  $w(E_j^\epsilon) < \infty$  since  $w$  is locally finite by hypothesis.

Recalling the comments immediately after Definition 3.18 and in Remark 3.19, we obtain a useful special case of Theorem 3.24:

**Corollary 3.25.** *Let the assumptions of Section 3A hold,  $\Omega$  and  $Q$  be bounded,  $w = v = \mu$ , and (3-12) be true. Let  $\Omega'$  be an open set with  $\bar{\Omega}' \subset \Omega$ . Fix  $p \in [1, \infty)$  and suppose the Poincaré property of order  $p$  in Definition 3.5 holds for  $\text{Lip}_{Q,p,0}(\Omega')$  and the local Sobolev property of order  $p$  in Definition 3.16 holds. Then, for every  $\{(f_k, \vec{g}_k)\} \subset W_{v,\mu,0}^{1,p}(\Omega', Q)$  which is bounded in  $W_{v,\mu}^{1,p}(\Omega, Q)$  norm, there is a subsequence  $\{f_{k_i}\}$  of  $\{f_k\}$  and a function  $f \in L_w^{p\sigma}(\Omega')$  such that  $f_{k_i} \rightarrow f$  pointwise a.e.- $w$  in  $\Omega'$  and in  $L_w^q(\Omega')$  norm,  $1 \leq q < p\sigma$ .*

In the case where  $p = 2$  and all measures are Lebesgue measure, Corollary 3.25 is used in [Rodney 2007; 2012] to show the existence of weak solutions to Dirichlet problems for some linear subelliptic equations. It is also used in [Rodney 2010] to derive the following global Sobolev inequality from the local estimate (3-32), where  $\Omega'$  is open and  $\bar{\Omega}' \subset \Omega$ :

$$(3-41) \quad \|f\|_{L^{2\sigma}(\Omega')} \leq C \left( \int_{\Omega'} |\sqrt{Q}\nabla f|^2 dx \right)^{1/2}.$$

#### 4. Precompact subsets of $L^N$ in a quasimetric space

In this section, we consider the situation of an open set  $\Omega$  in a topological space  $X$  when  $X$  is also endowed with a quasimetric  $d$ . As there is no easy way to define Sobolev spaces on general quasimetric spaces, this section concentrates on establishing a simple criterion not directly related to Sobolev spaces ensuring that bounded subsets of  $L_w^N(\Omega)$  are precompact in  $L_w^q(\Omega)$  when  $1 \leq q < N \leq \infty$ .

We begin by further describing the setting for our result. The topology on  $X$  is expressed in terms of a fixed collection  $\mathcal{T}$  of subsets of  $X$  which may not be related to the quasimetric  $d$ . Thus when we say that a set  $\mathcal{O} \subset X$  is *open*, we mean that  $\mathcal{O} \in \mathcal{T}$ . Given an open  $\Omega$ , we assume the following:

- (i)  $\forall x \in X$  and  $r > 0$ , the  $d$ -ball  $B_r(x) = \{y \in X : d(x, y) < r\}$  is a Borel set.
- (ii)  $\forall x \in X$  and  $r > 0$ , there is an open set  $\mathcal{O}$  such that  $x \in \mathcal{O} \subset B_r(x)$ .
- (iii) If  $X \neq \Omega$ , then  $\forall x \in \Omega$ ,  $d(x, \Omega^c) = \inf\{d(x, y) : y \in \Omega^c\} > 0$ .

Property (ii) serves as a substitute for (3-2).

Unlike the situation in Section 3,  $d$ -balls centered in  $\Omega$  may not be subsets of  $\Omega$  unless  $X = \Omega$ . However, we note the following fact.

**Remark 4.1.** Properties (ii) and (iii) guarantee that for any compact set  $K \subset \Omega$ , there exists  $\varepsilon(K) > 0$  such that  $B_r(x) \subset \Omega$  if  $x \in K$  and  $r < \varepsilon(K)$ . In fact, first note that for any  $x \in \Omega$ , (iii) implies that the  $d$ -ball  $B(x)$  with center  $x$  and radius  $r_x = d(x, \Omega^c)/(2\kappa)$  lies in  $\Omega$ . If  $K$  is a compact set in  $\Omega$ , (ii) shows that  $K$  can be covered by a finite number of such balls  $\{B(x_i)\}$ . With  $\varepsilon(K)$  chosen to be a suitably small multiple (depending on  $\kappa$ ) of  $\min\{r_{x_i}\}$ , the remark then follows easily from the swallowing property of  $d$ -balls.

Further, we assume that  $(\Omega, d)$  satisfies the local geometric doubling condition in Definition 3.3, that is, for each compact set  $K \subset \Omega$ , there exists  $\delta'(K) > 0$  such that, for all  $x \in K$  and all  $0 < r' < r < \delta'(K)$ , the number of disjoint  $d$ -balls of common radius  $r'$  contained in  $B_r(x)$  is at most a constant  $\mathcal{C}_{r/r'}$  depending on  $r/r'$  but not on  $K$ . We will choose  $\delta'(K) \leq \varepsilon(K)$ .

With this framework in force, we now state the main result of the section.

**Theorem 4.2.** *Let  $\Omega \subset X$  be as above, and let  $w$  be a finite Borel measure on  $\Omega$  such that, given any  $\epsilon > 0$ , there is a compact set  $K \subset \Omega$  with  $w(\Omega \setminus K) < \epsilon$ . Let  $1 \leq p < \infty$  and  $1 < N \leq \infty$ , and suppose  $\mathcal{S} \subset L_w^N(\Omega)$  has the property that, for any compact set  $K \subset \Omega$ , there exists  $\delta_K > 0$  such that*

(4-1)

$$\|f - f_{B,w}\|_{L_w^p(B)} \leq b(f, B) \quad \text{if } f \in \mathcal{S} \text{ and } B = B_r(x), x \in K, 0 < r < \delta_K,$$

where  $b(f, B)$  is a nonnegative ball set function. Furthermore, suppose there is a constant  $c_0 \geq 1$  such that for every  $\epsilon > 0$  and every compact set  $K \subset \Omega$ , there exists

$\tilde{\delta}_{\epsilon, K} > 0$  such that

$$(4-2) \quad \sum_{B \in \mathcal{F}} b(f, B)^p \leq \epsilon^p \quad \text{for all } f \in \mathcal{S}$$

for every finite family  $\mathcal{F} = \{B\}$  of  $d$ -balls centered in  $K$  with common radius less than  $\tilde{\delta}_{\epsilon, K}$  for which  $\{c_0 B\}$  is a pairwise disjoint family of subsets of  $\Omega$ . Then any sequence in  $\mathcal{S}$  that is bounded in  $L_w^N(\Omega)$  has a subsequence that converges in  $L_w^q(\Omega)$  for  $1 \leq q < N$  to a function in  $L_w^N(\Omega)$ .

*Proof.* Let  $\epsilon > 0$  and choose a compact set  $K \subset \Omega$  with  $w(\Omega \setminus K) < \epsilon$ . Next, for  $c_0 \geq 1$ , as in the proof of Lemma 3.12, there is a positive constant  $r = r(\epsilon, K, c_0) < \min\{\delta_K, \tilde{\delta}_{\epsilon, K}, \delta'(K), \varepsilon(K)/(\gamma c_0)\}$  (see (4-1), (4-2), Definition 3.3 and Remark 4.1), where  $\gamma = \kappa + 2\kappa^2$  with  $\kappa$  as in (3-1), and a finite family  $\{B_r(y_k)\}_k$  of  $d$ -balls centered in  $K$  satisfying  $K \subset \bigcup_k B_r(y_k)$  and whose dilates  $\{B_{c_0 r}(y_k)\}_k$  lie in  $\Omega$  and have the bounded intercept property (with intercept constant  $M$  independent of  $\epsilon$ ). Since  $\{B_{c_0 r}(y_k)\}_k$  has bounded intercepts with bound  $M$ , it can be written as the union of at most  $M$  families of disjoint  $d$ -balls; see, for example, the proof of [Chua and Wheeden 2008, Lemma 2.5]. By (4-2), we conclude that

$$\sum_k b(f, B_r(y_k))^p \leq M\epsilon^p.$$

Theorem 4.2 follows then immediately from Theorem 1.2; see also Remark 1.3(1).  $\square$

As an application of Theorem 4.2 we present a version of [Hajlasz and Koskela 2000, Theorem 8.1] in the case  $p \geq 1$ . Our version improves the one in [Hajlasz and Koskela 2000] by allowing two different measures and by relaxing the assumptions made about embedding and doubling. Furthermore, while the analogue in [Hajlasz and Koskela 2000] of our (4-3) uses only the  $L_w^1(B)$  norm on the left side, it automatically self-improves to the  $L_w^p(B)$  norm due to the doubling assumption, with a further fixed enlargement of the ball  $c_0 B$  on the right side; see, for example, [Hajlasz and Koskela 2000, Theorem 5.1].

**Corollary 4.3.** *Let  $X, d, \Omega, w$  be as above, and let  $\mu$  be a Borel measure on  $\Omega$ . Fix  $1 \leq p < \infty$ ,  $1 < N \leq \infty$ , and  $c_0 \geq 1$ . Consider a sequence of pairs*

$$\{(f_i, g_i)\} \subset L_w^N(\Omega) \times L_\mu^p(\Omega)$$

such that, for any compact set  $K \subset \Omega$ , there exists  $\bar{\delta}_K > 0$  with

$$(4-3) \quad \|f_i - (f_i)_{B,w}\|_{L_w^p(B)} \leq a_*(B) \|g_i\|_{L_\mu^p(c_0 B)}$$

for all  $i$  and all  $d$ -balls  $B$  centered in  $K$  with  $c_0 B \subset \Omega$  and  $r(B) < \bar{\delta}_K$ , where

$a_*(B)$  is a nonnegative ball set function satisfying

$$(4-4) \quad \limsup_{r \rightarrow 0} \{ \sup_{y \in K} a_*(B_r(y)) \} = 0.$$

Then if  $\{f_i\}$  and  $\{g_i\}$  are bounded in  $L_w^N(\Omega)$  and  $L_\mu^p(\Omega)$ , respectively,  $\{f_i\}$  has a subsequence converging in  $L_w^q(\Omega)$  for  $1 \leq q < N$  to a function belonging to  $L_w^N(\Omega)$ .

*Proof.* Given  $\epsilon > 0$  and compact set  $K \subset \Omega$ , use (4-4) to choose  $r_0 > 0$  such that  $a_*(B_r) < \epsilon/\beta$  for any  $d$ -ball  $B_r$  centered in  $K$  with  $r < r_0$ , where  $\beta = \sup_i \|g_i\|_{L_\mu^p(\Omega)} < \infty$ . In Theorem 4.2, choose  $\mathcal{S} = \{f_i\}$ ,  $\delta_K = \bar{\delta}_K$ ,  $b(f_i, B) = a_*(B)\|g_i\|_{L_\mu^p(c_0B)}$ , and

$$\tilde{\delta}_{\epsilon, K} = \min\{\bar{\delta}_K, \delta'(K), r_0, \epsilon(K)/c_0\}.$$

If  $B$  is a  $d$ -ball with center in  $K$  and  $r(B) < \tilde{\delta}_{\epsilon, K}$ , then  $c_0B \subset \Omega$ . Hence

$$\sum_{B \in \mathcal{F}} (a_*(B)\|g_i\|_{L_\mu^p(c_0B)})^p \leq \epsilon^p \|g_i\|_{L_\mu^p(\Omega)}^p / \beta^p \leq \epsilon^p$$

for every  $\mathcal{F}$  as in Theorem 4.2. The conclusion now follows from Theorem 4.2.  $\square$

**Remark 4.4.** (1) The  $g_i$  in (4-3) are usually the modulus of a fixed derivative of the corresponding  $f_i$ , such as  $|\nabla f_i|$  when  $X$  is a Riemannian manifold. More generally,  $g_i$  may be the upper gradient of  $f_i$  (see [Heinonen 2001] for the definition).

(2) Theorem 4.2 can also be used to obtain an extension of Theorem 2.4 to  $s$ -John domains in quasimetric spaces; see [Chua and Wheeden 2011, Theorem 1.6].

### Appendix

We briefly consider analogues of Theorem 3.14, Corollary 3.15, and Theorem 3.20 without assuming  $w \ll v$ , but adding the assumption that  $\mathcal{H}$  is linear. In this case, (3-27) can be extended by continuity to obtain a bounded linear map from  $\bar{\mathcal{H}}$  into  $L_w^{p\sigma}(\Omega)$ . Here, as always,  $\bar{\mathcal{H}}$  denotes the closure of  $\{(f, \nabla f) : f \in \mathcal{H}\}$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ . However, when  $w \ll v$  fails, there is no natural way to obtain the extension for every  $(f, \vec{g}) \in \bar{\mathcal{H}}$  keeping the same  $f$  on the left side. In fact, let  $(f, \vec{g}) \in \bar{\mathcal{H}}$  and choose  $\{f_j\} \subset \mathcal{H}$  with  $(f_j, \nabla f_j) \rightarrow (f, \vec{g})$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ . Linearity of  $\mathcal{H}$  allows us to apply (3-27) to differences of the  $f_j$  and conclude that  $\{f_j\}$  is a Cauchy sequence in  $L_w^{p\sigma}(\Omega)$ . Therefore  $f_j \rightarrow f^*$  in  $L_w^{p\sigma}(\Omega)$  for some  $f^* \in L_w^{p\sigma}(\Omega)$ , and

$$\|f^*\|_{L_w^{p\sigma}(\Omega)} \leq C \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(\Omega, Q)} \quad \text{if } (f, \vec{g}) \in \bar{\mathcal{H}}.$$

The function  $f^*$  is determined by  $(f, \vec{g})$ , that is,  $f^*$  is independent of the particular sequence  $\{f_j\} \subset \mathcal{H}$  above. Indeed, if  $\{\tilde{f}_j\}$  is another sequence in  $\mathcal{H}$  with  $(\tilde{f}_j, \nabla \tilde{f}_j) \rightarrow (f, \vec{g})$  in  $W_{v,\mu}^{1,p}(\Omega, Q)$ , and if  $\tilde{f}_j \rightarrow \tilde{f}^*$  in  $L_w^{p\sigma}(\Omega)$ , then, by (3-27) and linearity of  $\mathcal{H}$ ,

$$\|\tilde{f}_j - f_j\|_{L_w^{p\sigma}(\Omega)} \leq C \|(\tilde{f}_j - f_j, \nabla \tilde{f}_j - \nabla f_j)\|_{W_{v,\mu}^{1,p}(\Omega, Q)} \rightarrow 0.$$

Consequently  $\|\tilde{f}^* - f^*\|_{L_w^{p\sigma}(\Omega)} = 0$ . Thus  $(f, \vec{g})$  determines  $f^*$  uniquely as an element of  $L_w^{p\sigma}(\Omega)$ . Define a mapping

$$(A-1) \quad T : \bar{\mathcal{H}} \rightarrow L_w^{p\sigma}(\Omega) \quad \text{by setting } T(f, \vec{g}) = f^*.$$

Note that  $\bar{\mathcal{H}}$  is a linear set in  $W_{v,\mu}^{1,p}(\Omega, Q)$  since  $\mathcal{H}$  is linear, and that  $T$  is a bounded linear map from  $\bar{\mathcal{H}}$  into  $L_w^{p\sigma}(\Omega)$ . Also note that  $T$  satisfies  $T(f, \nabla f) = f$  when restricted to those  $(f, \nabla f)$  with  $f \in \mathcal{H}$ . Furthermore, if  $w \ll v$ , then  $T(f, \vec{g}) = f$  for all  $(f, \vec{g}) \in \bar{\mathcal{H}}$ , that is,  $f^* = f$  a.e.- $w$  for all  $(f, \vec{g}) \in \bar{\mathcal{H}}$ . This follows since  $f_j \rightarrow f$  in  $L_v^p(\Omega)$  norm and  $f_j \rightarrow f^*$  in  $L_w^{p\sigma}(\Omega)$  norm. In this appendix, where it is not assumed that  $w \ll v$ ,  $f^*$  plays a main role. One can find a function  $h$  such that  $h = f^*$  a.e.- $w$  and  $h = f$  a.e.- $v$ , but as this fact is not needed, we omit its proof.

An analogue of Theorem 3.14 is given in the next result.

**Theorem A.1.** *Let all the assumptions of Theorem 3.14 hold except that now the set  $\mathcal{H}$  is linear and we do not assume  $w \ll v$ . Then the map  $T : \bar{\mathcal{H}} \rightarrow L_w^q(\Omega)$  defined in (A-1) is compact if  $1 \leq q < p\sigma$ . Equivalently, if  $\{(f_k, \vec{g}_k)\}$  is a sequence in  $\bar{\mathcal{H}}$  with  $\sup_k \|(f_k, \vec{g}_k)\|_{W_{v,\mu}^{1,p}(\Omega, Q)} < \infty$ , then  $\{f_k^*\}$  has a subsequence which converges in  $L_w^q(\Omega)$  for  $1 \leq q < p\sigma$ , where  $f_k^* = T(f_k, \vec{g}_k)$ . Moreover, the limit of the subsequence belongs to  $L_w^{p\sigma}(\Omega)$ .*

*Proof.* Let  $\mathcal{H}$  satisfy the hypothesis of the theorem and let  $\{(f_k, \vec{g}_k)\} \subset \bar{\mathcal{H}}$  be bounded in  $W_{v,\mu}^{1,p}(\Omega, Q)$ . For each  $k$ , choose  $h_k \in \mathcal{H}$  such that

$$(A-2) \quad \|(f_k, \vec{g}_k) - (h_k, \nabla h_k)\|_{W_{v,\mu}^{1,p}(\Omega, Q)} \leq 2^{-k}.$$

Set  $\mathcal{H}_1 = \{h_k\}_k \subset \mathcal{H}$ . Then  $\{(h_k, \nabla h_k) : h_k \in \mathcal{H}_1\}$  is bounded in  $W_{v,\mu}^{1,p}(\Omega, Q)$ . Furthermore, (3-27) implies a version of (3-14), namely,

$$\sup_{f \in \mathcal{H}_1} \{\|f\|_{L_w^{p\sigma}(\Omega)} + \|(f, \nabla f)\|_{W_{v,\mu}^{1,p}(\Omega, Q)}\} < \infty.$$

Theorem 3.7 now applies to  $\mathcal{H}_1$  with  $N = p\sigma$  and gives that any sequence in  $\widehat{\mathcal{H}}_1$  has a subsequence which converges in  $L_w^q(\Omega)$  norm for  $1 \leq q < p\sigma$  to a function belonging to  $L_w^{p\sigma}(\Omega)$ . The sequence  $\{h_k\}$  lies in  $\widehat{\mathcal{H}}_1$ , as is easily seen by considering, for each fixed  $k$ , the constant sequence  $\{f^j\}$  defined by  $f^j = h_k$  for



all  $j$ . We conclude that  $\{h_k\}$  has a subsequence  $\{h_{k_l}\}$  converging in  $L_w^q(\Omega)$  norm for  $1 \leq q < p\sigma$  to a function  $h \in L_w^{p\sigma}(\Omega)$ . By linearity and boundedness of  $T$  from  $\mathcal{H}$  to  $L_w^{p\sigma}(\Omega)$  together with (A-2), we have (writing  $f_k^* = T(f_k, \vec{g}_k)$ )

$$\|f_k^* - h_k\|_{L_w^{p\sigma}(\Omega)} = \|T(f_k, \vec{g}_k) - T(h_k, \nabla h_k)\|_{L_w^{p\sigma}(\Omega)} \leq C2^{-k} \rightarrow 0.$$

Restricting  $k$  to  $\{k_l\}$  and using  $w(\Omega) < \infty$ , we conclude that  $\{f_{k_l}^*\}$  also converges to  $h$  in  $L_w^q(\Omega)$  for  $1 \leq q < p\sigma$ , which completes the proof.  $\square$

Setting  $\mathcal{H} = \text{Lip}_{Q,p}(\Omega)$  in Theorem A.1 gives an analogue of Corollary 3.15.

**Corollary A.2.** *Let the hypotheses of Theorem A.1 hold for  $\mathcal{H} = \text{Lip}_{Q,p}(\Omega)$ . Then the map  $T$  defined by (A-1) is a compact map of  $W_{v,\mu}^{1,p}(\Omega, Q)$  into  $L_w^q(\Omega)$  for  $1 \leq q < p\sigma$ , that is, if  $\{(f_k, \vec{g}_k)\} \subset W_{v,\mu}^{1,p}(\Omega, Q)$  and  $\sup_k \|(f_k, \vec{g}_k)\|_{W_{v,\mu}^{1,p}(\Omega, Q)} < \infty$ , then  $\{f_k^*\}$  has a subsequence which converges in  $L_w^q(\Omega)$  for  $1 \leq q < p\sigma$ , where  $f_k^* = T(f_k, \vec{g}_k)$ . Moreover, the limit of the subsequence belongs to  $L_w^{p\sigma}(\Omega)$ .*

Theorem 3.20 also has an analogue without assuming  $w \ll v$  provided  $\mathcal{H}$  is linear, and in this instance (3-27) is not required: the subsequence  $\{f_{k_l}\}$  of  $\{f_k\}$  in the conclusion is then replaced by a subsequence of  $\{f_k^*\}$ , where  $f_k^*$  is constructed as above but now using bounded open  $\Omega'$  whose closures increase to  $\Omega$ . Now  $f^*$  arises when (3-38) is extended to  $\bar{\mathcal{H}}$ , namely, instead of (3-39), we obtain

$$\|f^*\|_{L_w^{p\sigma}(\Omega')} \leq C(\Omega'') \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(\Omega, Q)} \quad \text{if } (f, \vec{g}) \in \bar{\mathcal{H}}$$

where  $f^*$  is constructed for a pair  $(f, \vec{g}) \in \bar{\mathcal{H}}$  by using linearity of  $\mathcal{H}$  and (3-38) for a particular  $(\Omega', \Omega'')$ . It is easy to see that  $f^* \in L_{w,\text{loc}}^{p\sigma}(\Omega)$  by letting  $\Omega' \nearrow \Omega$ . The Poincaré inequality analogous to (3-40) is

$$\left( \int_{B_r(y)} |f^* - f_{B_r(y),w}^*|^p dw \right)^{1/p} \leq \epsilon \|(f, \vec{g})\|_{W_{v,\mu}^{1,p}(B_{\epsilon_0 r}(y), Q)} \quad \text{if } (f, \vec{g}) \in \bar{\mathcal{H}},$$

obtained by extending (3-8) from  $\mathcal{H}$  to  $\bar{\mathcal{H}}$ . Further details are omitted.

### References

- [Adams and Fournier 2003] R. A. Adams and J. J. F. Fournier, *Sobolev spaces*, 2nd ed., Pure and Applied Mathematics **140**, Academic Press, New York, 2003. MR 2009e:46025 Zbl 1098.46001
- [Aubin 1976] T. Aubin, “Équations différentielles non linéaires et problème de Yamabe concernant la courbure scalaire”, *J. Math. Pures Appl.* (9) **55**:3 (1976), 269–296. MR 55 #4288 Zbl 0336.53033
- [Chua 1995] S.-K. Chua, “Weighted Sobolev interpolation inequalities on certain domains”, *J. London Math. Soc.* (2) **51**:3 (1995), 532–544. MR 96d:46033 Zbl 0845.26008
- [Chua and Wheeden 2008] S.-K. Chua and R. L. Wheeden, “Self-improving properties of inequalities of Poincaré type on measure spaces and applications”, *J. Funct. Anal.* **255**:11 (2008), 2977–3007. MR 2010f:46050 Zbl 1172.46020

- [Chua and Wheeden 2011] S.-K. Chua and R. L. Wheeden, “Self-improving properties of inequalities of Poincaré type on  $s$ -John domains”, *Pacific J. Math.* **250**:1 (2011), 67–108. MR 2012c:46068 Zbl 1214.26014
- [Fabes et al. 1982] E. B. Fabes, C. E. Kenig, and R. P. Serapioni, “The local regularity of solutions of degenerate elliptic equations”, *Comm. Partial Differential Equations* **7**:1 (1982), 77–116. MR 84i:35070 Zbl 0498.35042
- [Franchi et al. 1997] B. Franchi, R. Serapioni, and F. Serra Cassano, “Approximation and imbedding theorems for weighted Sobolev spaces associated with Lipschitz continuous vector fields”, *Boll. Un. Mat. Ital. B (7)* **11**:1 (1997), 83–117. MR 98c:46062 Zbl 0952.49010
- [Gilbarg and Trudinger 1997] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, 3rd ed., Grundlehren der mathematischen Wissenschaften **224**, Springer, Berlin, 1997. MR 2001k:35004 Zbl 1042.35002
- [Hajłasz 1993] P. Hajłasz, “Note on Meyers–Serrin’s theorem”, *Exposition. Math.* **11**:4 (1993), 377–379. MR 94e:46060 Zbl 0799.46042
- [Hajłasz and Koskela 1998] P. Hajłasz and P. Koskela, “Isoperimetric inequalities and imbedding theorems in irregular domains”, *J. London Math. Soc. (2)* **58**:2 (1998), 425–450. MR 99m:46079 Zbl 0922.46034
- [Hajłasz and Koskela 2000] P. Hajłasz and P. Koskela, *Sobolev met Poincaré*, Mem. Amer. Math. Soc. **688**, American Mathematical Society, 2000. MR 2000j:46063 Zbl 0952.46022
- [Heinonen 2001] J. Heinonen, *Lectures on analysis on metric spaces*, Springer, New York, 2001. MR 2002c:30028 Zbl 0985.46008
- [Hytönen and Martikainen 2012] T. Hytönen and H. Martikainen, “Non-homogeneous  $Tb$  theorem and random dyadic cubes on metric measure spaces”, *J. Geom. Anal.* **22**:4 (2012), 1071–1107. MR 2965363 Zbl 06124339
- [Kilpeläinen and Malý 2000] T. Kilpeläinen and J. Malý, “Sobolev inequalities on sets with irregular boundaries”, *Z. Anal. Anwendungen* **19**:2 (2000), 369–380. MR 2001g:46075 Zbl 0959.46020
- [Monticelli et al. 2012] D. D. Monticelli, S. Rodney, and R. L. Wheeden, “Boundedness of weak solutions of degenerate quasilinear equations with rough coefficients”, *Differential Integral Equations* **25**:1-2 (2012), 143–200. MR 2906551 Zbl 1249.35117
- [Rios et al. 2013] C. Rios, E. T. Sawyer, and R. L. Wheeden, “Hypoellipticity for infinitely degenerate quasilinear equations and the Dirichlet problem”, *J. d’Analyse Math.* **119** (2013), 1–62. MR 3043146 Zbl 06186919
- [Rodney 2007] S. W. Rodney, *Existence of weak solutions to subelliptic partial differential equations in divergence form and the necessity of the Sobolev and Poincaré inequalities*, Ph.D. thesis, McMaster University, 2007, available at <http://search.proquest.com/docview/304819194>. MR 2711279
- [Rodney 2010] S. Rodney, “A degenerate Sobolev inequality for a large open set in a homogeneous space”, *Trans. Amer. Math. Soc.* **362**:2 (2010), 673–685. MR 2011f:35055 Zbl 1190.35010
- [Rodney 2012] S. Rodney, “Existence of weak solutions of linear subelliptic Dirichlet problems with rough coefficients”, *Canad. J. Math.* **64**:6 (2012), 1395–1414. MR 2994671 Zbl 06111146
- [Rudin 1987] W. Rudin, *Real and complex analysis*, 3rd ed., McGraw-Hill, New York, 1987. MR 88k:00002 Zbl 0925.00005
- [Sawyer and Wheeden 2006] E. T. Sawyer and R. L. Wheeden, *Hölder continuity of weak solutions to subelliptic equations with rough coefficients*, Mem. Amer. Math. Soc. **847**, American Mathematical Society, 2006. MR 2007f:35037 Zbl 1096.35031

- [Sawyer and Wheeden 2010] E. T. Sawyer and R. L. Wheeden, “Degenerate Sobolev spaces and regularity of subelliptic equations”, *Trans. Amer. Math. Soc.* **362**:4 (2010), 1869–1906. MR 2010m:35077 Zbl 1191.35085
- [Schoen 1984] R. Schoen, “Conformal deformation of a Riemannian metric to constant scalar curvature”, *J. Differential Geom.* **20**:2 (1984), 479–495. MR 86i:58137 Zbl 0576.53028
- [Trudinger 1968] N. S. Trudinger, “Remarks concerning the conformal deformation of Riemannian structures on compact manifolds”, *Ann. Scuola Norm. Sup. Pisa* (3) **22** (1968), 265–274. MR 39 #2093 Zbl 0159.23801
- [Turesson 2000] B. O. Turesson, *Nonlinear potential theory and weighted Sobolev spaces*, Lecture Notes in Mathematics **1736**, Springer, Berlin, 2000. MR 2002f:31027 Zbl 0949.31006
- [Yamabe 1960] H. Yamabe, “On a deformation of Riemannian structures on compact manifolds”, *Osaka Math. J.* **12** (1960), 21–37. MR 23 #A2847 Zbl 0096.37201

Received July 25, 2012.

SENG-KEE CHUA  
DEPARTMENT OF MATHEMATICS  
NATIONAL UNIVERSITY OF SINGAPORE  
10, LOWER KENT RIDGE ROAD  
SINGAPORE 119076  
SINGAPORE  
matcsk@nus.edu.sg

SCOTT RODNEY  
DEPARTMENT OF MATHEMATICS, PHYSICS, AND GEOLOGY  
CAPE BRETON UNIVERSITY  
P.O. BOX 5300, 1250 GRAND LAKE ROAD  
SYDNEY, NS B1P 6L2  
CANADA  
scott\_rodney@cbu.ca

RICHARD L. WHEEDEN  
DEPARTMENT OF MATHEMATICS  
RUTGERS UNIVERSITY  
110 FRELINGHUYSEN ROAD  
PISCATAWAY, NJ 08854  
UNITED STATES  
wheeden@math.rutgers.edu



## PARTIAL INTEGRABILITY OF ALMOST COMPLEX STRUCTURES AND THE EXISTENCE OF SOLUTIONS FOR QUASILINEAR CAUCHY–RIEMANN EQUATIONS

CHONG-KYU HAN AND JONG-DO PARK

We study the local solvability of the system of quasilinear Cauchy–Riemann equations for  $d$  unknown functions in  $n$  complex variables, which is a system of elliptic type and overdetermined if  $n \geq 2$ . We consider an associated almost complex structure on  $\mathbb{C}^{n+d}$  and its partial integrability and prove by using the Newlander–Nirenberg theorem and its algebraic generalizations that the existence of a pseudoholomorphic function on the zero set is equivalent to the local solvability of the original quasilinear system. We discuss an algorithm for finding pseudoholomorphic functions on the zero set and then present examples.

### Introduction

A classical method for solving partial differential equations (PDE) of first order is the *method of characteristics* which originated from [Monge 1803]. One finds curves along which a PDE becomes a system of ordinary differential equations and constructs a solution whose graph, or 1-jet graph, is a union of those curves. Consider a quasilinear equation

$$(0-1) \quad \sum_{\lambda=1}^n a^\lambda(x, u) \frac{\partial u}{\partial x^\lambda} = b(x, u)$$

for a real-valued function  $u$  in  $n$  real variables  $x = (x^1, \dots, x^n)$ , where  $a^\lambda$  and  $b$  are smooth ( $C^\infty$ ) and  $a^\lambda$ ,  $\lambda = 1, \dots, n$ , are not all zero. The characteristic vector field of (0-1) is a smooth vector field

$$X = \sum_{\lambda=1}^n a^\lambda \frac{\partial}{\partial x^\lambda} + b \frac{\partial}{\partial u}$$

---

The authors were partially supported by National Research Foundation (NRF) of Republic of Korea, with grant numbers 2009-0070971 and 2010-0011841, respectively.

*MSC2010*: primary 32W05, 35N10; secondary 32Q60, 35J60.

*Keywords*: overdetermined system, elliptic PDE system, almost complex structure, nonlinear Cauchy–Riemann equations.

on  $\mathbb{R}^{n+1} = \{(x, u)\}$  and a smooth real-valued function  $\phi(x, u)$  is a first integral of  $X$  if  $X\phi = 0$ . Then, by the implicit function theorem, any first integral  $\phi$  with  $\phi_u \neq 0$  gives an implicit solution  $\phi(x, u) = 0$  to (0-1). The same method works for systems. Consider

$$(0-2) \quad \sum_{\lambda=1}^n a_j^\lambda(x, u) \frac{\partial u}{\partial x^\lambda} = b_j(x, u), \quad j = 1, \dots, p, \quad p \leq n,$$

for a real-valued function  $u$  in  $n$  real variables  $x = (x^1, \dots, x^n)$ . We assume the matrix  $(a_j^\lambda)$  has maximal rank  $p$ . If  $p \geq 2$ , then (0-2) is overdetermined; therefore, there are no solutions generically. To discuss the existence of solutions let

$$(0-3) \quad X_j = \sum_{\lambda=1}^n a_j^\lambda \frac{\partial}{\partial x^\lambda} + b_j \frac{\partial}{\partial u}, \quad j = 1, \dots, p,$$

be vector fields on  $\mathbb{R}^{n+1} = \{(x, u)\}$ . For a smooth function  $u(x)$ , a normal vector to the graph  $S = \{(x, u(x)) \in \mathbb{R}^{n+1}\}$  is  $(\nabla u, -1) = (\partial u / \partial x^1, \dots, \partial u / \partial x^n, -1)$ . Then (0-2) is equivalent to  $X_j \cdot (\nabla u, -1) = 0$ , which implies that  $X_j$  is tangent to the graph  $S$  at every point. A smooth real-valued function  $F$  is said to have invariant zero-level with respect to vector fields  $X_1, \dots, X_p$  if  $(X_j F)(x) = 0$  for all  $j = 1, \dots, p$  and for all  $x$  with  $F(x) = 0$ . We have:

**Theorem 0.1.** *Let  $(x_0, u_0) \in \mathbb{R}^n \times \mathbb{R}$ . On a neighborhood of  $(x_0, u_0)$  there exists a solution  $u(x)$  of (0-2) with  $u(x_0) = u_0$  if and only if there is a function  $F(x, u)$  with  $\partial F / \partial u \neq 0$  and  $F(x_0, u_0) = 0$  that has invariant zero-level with respect to the set of vector fields (0-3).*

*Proof.* Suppose that  $u = f(x)$ ,  $x = (x^1, \dots, x^n)$ , is a solution of (0-2). Let  $F(x, u) := f(x) - u$ . Then we see that  $F_u \neq 0$  and that, for each  $j = 1, \dots, p$ ,

$$X_j F = 0 \quad \text{on } \{F = 0\}.$$

Conversely, suppose that  $F(x, u)$  is a function with  $F_u \neq 0$ ,  $F(x_0, u_0) = 0$ , that has invariant zero-level. Differentiating the implicit function  $F(x, u) = 0$  with respect to  $x^\lambda$  using the chain rule, we have

$$(0-4) \quad \frac{\partial F}{\partial x^\lambda} = -\frac{\partial F}{\partial u} \frac{\partial u}{\partial x^\lambda} \quad \text{for each } \lambda = 1, \dots, n.$$

On the other hand, since  $F$  has invariant zero-level we have

$$(0-5) \quad X_j F = \sum_{\lambda=1}^n a_j^\lambda \frac{\partial F}{\partial x^\lambda} + b_j \frac{\partial F}{\partial u} = 0 \quad \text{on } \{F = 0\}.$$

Substituting (0-4) for  $\partial F / \partial x^\lambda$  in (0-5) we have

$$-\frac{\partial F}{\partial u} \left( \sum_{\lambda=1}^n a_j^\lambda \frac{\partial u}{\partial x^\lambda} - b_j \right) = 0 \quad \text{on } \{F = 0\};$$

that is,  $u = f(x)$  satisfies (0-2). □

For more details on (0-2) we refer the readers to [Han and Park 2013]. In this paper we study the complex analogue of Theorem 0.1. Let  $z = (z^1, \dots, z^n)$  for  $z^j = x^j + \sqrt{-1}y^j$  be complex variables and let

$$\frac{\partial}{\partial \bar{z}^j} := \frac{1}{2} \left( \frac{\partial}{\partial x^j} + \sqrt{-1} \frac{\partial}{\partial y^j} \right) \quad \text{and} \quad \frac{\partial}{\partial z^j} := \frac{1}{2} \left( \frac{\partial}{\partial x^j} - \sqrt{-1} \frac{\partial}{\partial y^j} \right).$$

Consider a system of PDE for a complex-valued unknown function  $w = w(z, \bar{z})$ :

$$(0-6) \quad \frac{\partial w}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k(z, \bar{z}, w, \bar{w}) \frac{\partial w}{\partial z^k} = B_j(z, \bar{z}, w, \bar{w}), \quad j = 1, \dots, n,$$

where  $A_j^k$  and  $B_j$  are complex-valued  $C^\infty$  functions that are defined on a neighborhood of the origin of  $\mathbb{C}^{n+1} = \{(z, w)\}$  and  $A_j^k$  are sufficiently small. If  $n \geq 2$ , then (0-6) is overdetermined. We shall call (0-6) *quasilinear Cauchy–Riemann equations*. We prove the local solvability in the  $C^\infty$  category by purely formal arguments based on the Newlander–Nirenberg theorem and its algebraic generalizations. We observed that a function  $\zeta(z, \bar{z}, w, \bar{w}) = 0$  is an implicit solution to (0-6) if and only if  $\zeta$  is pseudoholomorphic on the zero set (see Definition 1.6) with respect to an almost complex structure  $J$  on  $\mathbb{C}^{n+1} = \{(z, w)\}$  determined by the coefficients  $A_j^k$  and  $B_j$  (Theorem 3.3). Another observation is that a function  $\zeta(z, \bar{z}, w, \bar{w})$  with  $d\zeta \wedge d\bar{\zeta} \neq 0$  is pseudoholomorphic on the zero set if and only if the zero locus  $\zeta = 0$  is a  $J$ -invariant submanifold of  $(\mathbb{C}^{n+1}, J)$  (Theorem 2.5). To check the partial integrability of the almost complex structure we make use of Theorem 1.3, which is due to L. Nirenberg and F. Trèves.

Section 4 is a generalization of our results of Section 3 to the cases of multiple unknown functions  $w = (w^1, \dots, w^d)$ :

$$(0-7) \quad \frac{\partial w^\alpha}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k(z, \bar{z}, w, \bar{w}) \frac{\partial w^\alpha}{\partial z^k} = B_j^\alpha(z, \bar{z}, w, \bar{w})$$

for each  $j = 1, \dots, n$  ( $n \geq 1$ ) and  $\alpha = 1, \dots, d$ , where  $A_j^k$  and  $B_j^\alpha$  are  $C^\infty$  functions defined on a neighborhood of the origin of  $\mathbb{C}^{n+d}$  and  $A_j^k$  are sufficiently small.

We discuss in Section 5 the determined case  $n = 1$ . In Section 6 we present examples of  $n = 2$  including the equations for the pseudoanalytic functions, which will be introduced in Section 7. In the last section of this paper we briefly survey

the history of the perturbed Cauchy–Riemann equations and overdetermined PDE systems.

Finally we mention the regularity of solutions to (0-6) or (0-7): it is well-known (see [Gilbarg and Trudinger 1998]) that a linear elliptic partial differential equation is hypoelliptic; that is, any distribution solution is  $C^\infty$  whenever all the coefficients of the differential operators are  $C^\infty$ . In general, a nonlinear differential equation is said to be elliptic if its linearization is an elliptic differential operator (see [Taylor 1997]). It was proved in [Douglis and Nirenberg 1955] that nonlinear elliptic systems are hypoelliptic by Schauder-type a priori estimates. Since the linearization of each quasilinear equation in (0-6) or (0-7) is elliptic, any  $C^{1,\alpha}$  solution to (0-6) or (0-7) is always  $C^\infty$ .

### 1. $J$ -holomorphic functions on almost complex manifolds

Let  $M$  be a smooth ( $C^\infty$ ) manifold of dimension  $2m$ ,  $m \geq 2$ , with smooth almost complex structure  $J$ . Then the complexified tangent bundle  $\mathbb{C}TM$  has the decomposition

$$\mathbb{C}TM = T^{1,0}(M) \oplus T^{0,1}(M),$$

where  $T^{1,0}(M)$  ( $T^{0,1}(M)$ , respectively) is the subbundle of rank- $m$  of eigenvectors of  $J$  associated with the eigenvalue  $i$  ( $-i$ , respectively). The dual decomposition of the complexified cotangent bundle  $\mathbb{C}T^*M$  is

$$\mathbb{C}T^*M = (T^*M)^{1,0} \oplus (T^*M)^{0,1}.$$

We can find real vector fields  $X_j$ ,  $j = 1, \dots, m$ , such that

$$X_1, JX_1, \dots, X_m, JX_m$$

spans the real tangent bundle  $TM$ . Let

$$Z_j = \frac{1}{2}(X_j - iJX_j) \quad \text{and} \quad \bar{Z}_j = \frac{1}{2}(X_j + iJX_j)$$

for each  $j = 1, \dots, m$ . Then  $\{Z_1, \dots, Z_m\}$  spans  $T^{1,0}(M)$  and  $\{\bar{Z}_1, \dots, \bar{Z}_m\}$  spans  $T^{0,1}(M)$ . Let  $\{\theta^1, \dots, \theta^m\}$  be a set of independent 1-forms that annihilates  $T^{0,1}(M)$  and thus  $\{\bar{\theta}^1, \dots, \bar{\theta}^m\}$  annihilates  $T^{1,0}(M)$ . Then the subbundles  $(T^*M)^{1,0}$  and  $(T^*M)^{0,1}$  of the complexified cotangent bundle are the linear spans of  $\{\theta^1, \dots, \theta^m\}$  and  $\{\bar{\theta}^1, \dots, \bar{\theta}^m\}$ , respectively.

A complex-valued function  $\zeta$  is called  $J$ -holomorphic (or pseudoholomorphic) if

$$(1-1) \quad \bar{Z}_j \zeta = 0, \quad j = 1, \dots, m.$$

Equation (1-1) is an overdetermined system of linear PDE, and thus, in general, there are no solutions other than constants.  $J$ -holomorphic functions  $\zeta^1, \dots, \zeta^q$



are said to be *independent* if

$$d\zeta^1 \wedge \dots \wedge d\zeta^q \neq 0.$$

Equation (1-1) is equivalent to saying that  $d\zeta$  is a section of  $(T^*M)^{1,0}$ , so that there exist at most  $m$  independent  $J$ -holomorphic functions.  $J$  is said to be *integrable* if

$$(1-2) \quad [T^{1,0}(M), T^{1,0}(M)] \subset T^{1,0}(M),$$

which means that the bracket of any two sections of  $T^{1,0}(M)$  is again a section of  $T^{1,0}(M)$ . For the theory of general integrable structures, we refer the readers to [Berhanu et al. 2008].

We consider the exterior algebra of differential forms with complex coefficients:

$$\Omega^* = \Omega^0 \oplus \Omega^1 \oplus \dots \oplus \Omega^{2m},$$

where  $\Omega^0$  is the ring of smooth complex-valued functions and  $\Omega^r$  ( $r = 1, \dots, 2m$ ) is the module over  $\Omega^0$  of complex-valued smooth  $r$ -forms on  $M$ .

**Definition 1.1.** A subalgebra  $\mathcal{F}$  of  $\Omega^*$  is called an *algebraic ideal* if the following conditions hold:

- (i)  $\mathcal{F} \wedge \Omega^* \subset \mathcal{F}$ ,
- (ii) if  $\phi = \sum_{r=0}^{2m} \phi_r \in \mathcal{F}$ , where  $\phi_r \in \Omega^r$ , then each component  $\phi_r$  belongs to  $\mathcal{F}$  (homogeneity condition).

The homogeneity condition implies that  $\mathcal{F}$  is a two-sided ideal; that is,  $\Omega^* \wedge \mathcal{F} \subset \mathcal{F}$ .

In this paper we consider ideals generated by finitely many complex-valued functions and finitely many 1-forms. Let  $\rho = (\rho^1, \dots, \rho^d)$  and  $\phi = (\phi^1, \dots, \phi^q)$  be a system of functions and 1-forms, respectively. We denote by  $\mathcal{F}(\rho, \phi)$ , or simply by  $(\rho, \phi)$ , the algebraic ideal generated by  $\rho$  and  $\phi$ , which is the set of all elements of  $\Omega^*$  of the form

$$\sum_{\alpha=1}^d \rho^\alpha \omega^\alpha + \sum_{k=1}^q \phi^k \wedge \psi^k \quad \text{for some } \omega^\alpha, \psi^k \in \Omega^*.$$

For two elements  $\alpha$  and  $\beta$  of  $\Omega^*$ ,

$$\alpha \equiv \beta \pmod{(\rho, \phi)}$$

means that  $\alpha - \beta \in \mathcal{F}(\rho, \phi)$ .

The integrability condition (1-2) can be written as

$$(1-3) \quad [Z_j, Z_k] \in \Gamma(T^{1,0}(M)) \quad \text{for all } j, k = 1, \dots, m,$$

where  $\Gamma$  denotes the set of all smooth sections. Equation (1-3) is equivalent to

$$d\theta^\ell \equiv 0 \pmod{(\theta)} \quad \text{for all } \ell = 1, \dots, m,$$

where  $\theta = (\theta^1, \dots, \theta^m)$ .

**Theorem 1.2** [Newlander and Nirenberg 1957]. *Let  $(M^{2m}, J)$  be a  $C^\infty$  almost complex manifold. If  $J$  is integrable then there exist  $m$  independent  $J$ -holomorphic functions.*

The converse is also true, which is rather trivial. Now we fix notations: for any subbundle  $I \subset (T^*M)^{1,0}$  we denote by  $\underline{I}$  the module over  $\Omega^0$  of smooth sections of  $I$  and by  $(I)$  the algebraic ideal of  $\Omega^*$  generated by the smooth sections of  $I$ . By using Theorem 1.2 and the Frobenius theorem the following was proved in [Trèves 1981]:

**Theorem 1.3.** *Suppose that  $T'$  is a subbundle of  $(T^*M)^{1,0}$  of rank  $q$  ( $q < m$ ) and that  $T'$  is closed; that is,  $d\underline{T}' \subset (T')$ . Then there exist  $q$  independent  $J$ -holomorphic functions  $\zeta^1, \dots, \zeta^q$  whose differentials  $d\zeta^1, \dots, d\zeta^q$  span  $T'$ .*

The problem of determining conditions for the existence of  $J$ -holomorphic functions on almost complex manifolds has been examined in [Mushkarov 1981; 1986] by studying the involutivity of the Nijenhuis bundle. Criteria for the existence of  $J$ -holomorphic mappings into other almost complex manifolds are given in [Kruglikov 1998] in terms of Nijenhuis tensors and their generalizations. The following theorem is found in [Han and Kim 2012].

**Theorem 1.4.** *Let  $M^{2m}$  ( $m \geq 2$ ) be a  $C^\infty$  manifold with  $C^\infty$  almost complex structure  $J$ . Let  $(T^*M)^{1,0}$  be the bundle of  $(1, 0)$ -forms. Then there exist a sequence of subbundles  $(T^*M)^{1,0} := I^{(0)} \supset I^{(1)} \supset I^{(2)} \supset \dots$  and a nonnegative integer  $\nu$  such that for  $k = 0, 1, 2, \dots$ :*

- (i)  $I^{(k+1)} \subsetneq I^{(k)}$ , if  $k < \nu$ ,
- (ii)  $I^{(k+1)} = I^{(k)}$ , if  $k \geq \nu$ ,
- (iii)  $d\underline{I}^{(k+1)} \equiv 0 \pmod{I^{(k)}}$ ,

*under a generic assumption in each step of the construction of the sequence. Moreover, a function  $f$  is  $J$ -holomorphic if and only if  $df \in I^{(\nu)}$ ; thus the number of independent  $J$ -holomorphic functions is equal to the rank of  $I^{(\nu)}$ .*

**Definition 1.5.** The integer  $\nu$  of Theorem 1.4 is called the *type* of the almost complex structure  $J$ . We also say that the Pfaffian system  $(\theta^1, \dots, \theta^m)$  has derived length  $\nu$ .

*Proof of Theorem 1.4.* We shall find the largest closed subbundle of  $(T^*M)^{1,0}$  starting with  $I = I^{(0)} = (T^*M)^{1,0}$ : the exterior derivative  $d : \underline{I} \rightarrow \Omega^2$  is not a module homomorphism, but composition with the projection

$$\underline{I} \xrightarrow{d} \Omega^2 \xrightarrow{\pi} \Omega^2/(I)$$

is an  $\Omega^0$ -module homomorphism. Let  $\delta = \pi \circ d$ . Consider the submodule  $\underline{I}^{(1)} := \ker \delta$  of  $\underline{I}$ . We assume that  $\underline{I}^{(1)}$  has constant rank on  $M$ , and hence defines a subbundle  $I^{(1)}$  of  $(T^*M)^{1,0}$ . We have a short exact sequence of  $\Omega^0$ -modules

$$0 \rightarrow \underline{I}^{(1)} \rightarrow \underline{I} \xrightarrow{\delta} d\underline{I}/(I) \rightarrow 0.$$

The subbundle  $I^{(1)}$  is called the first derived system of  $(T^*M)^{1,0}$ . Assuming that  $\underline{I}^{(k-1)}$  has constant rank, we define inductively the  $k$ -th derived system  $I^{(k)}$  by

$$0 \rightarrow \underline{I}^{(k)} \rightarrow \underline{I}^{(k-1)} \xrightarrow{\delta} d\underline{I}^{(k-1)}/(I^{(k-1)}) \rightarrow 0.$$

Let  $\nu$  be the smallest integer with  $I^{(\nu)} = I^{(\nu+1)}$ . Then we have the sequence of subbundles

$$(1-4) \quad (T^*M)^{1,0} := I := I^{(0)} \supset I^{(1)} \supset \dots \supset I^{(\nu-1)} \supset I^{(\nu)}.$$

Notice that  $d\underline{I}^{(\nu)} \subset (I^{(\nu)})$ ; that is,  $I^{(\nu)}$  is closed. Assume that  $\underline{I}^{(\nu)}$  has constant rank  $q$ . Then by Theorem 1.3 there exist independent  $J$ -holomorphic functions  $\zeta^1, \dots, \zeta^q$ , which completes the proof of Theorem 1.4.  $\square$

The idea of Theorem 1.4 came from the theory of first integrals for Pfaffian systems due to E. Cartan and R. Gardner [Gardner 1967], which is a real version of Theorem 1.4. A generalized notion of the first integral has been used in [Ahn and Han 2012; Han and Park 2013]. Our standard reference for the theory of Pfaffian systems is [Bryant et al. 1991]. In this paper we need a notion of  $J$ -holomorphicity on the zero set that we define as follows:

**Definition 1.6.** A system of complex-valued functions  $\zeta = (\zeta^1, \dots, \zeta^d)$  is said to be  *$J$ -holomorphic on the zero set* if, for each  $\alpha = 1, \dots, d$ , we have  $(\bar{Z}_j \zeta^\alpha)(x) = 0$ ,  $j = 1, \dots, m$ , for all  $x$  with  $\zeta(x) = 0$ , or, equivalently, if

$$(1-5) \quad d\zeta^\alpha \equiv 0 \pmod{(\zeta, \bar{\zeta}, \theta)}.$$

Assuming further that  $\theta^j$  are dual to  $Z_k$ , that is,

$$\theta^j(Z_k) = \delta_k^j,$$

we define  $\partial f$  and  $\bar{\partial} f$  for any complex-valued function  $f$  by

$$\partial f := \sum_{j=1}^m (Z_j f) \theta^j, \quad \bar{\partial} f := \sum_{j=1}^m (\bar{Z}_j f) \bar{\theta}^j.$$

Then we have

$$df = \partial f + \bar{\partial} f.$$

We may write (1-5) as

$$\bar{\partial}\zeta^\alpha \equiv 0 \pmod{(\zeta, \bar{\zeta})} \quad \text{for each } \alpha = 1, \dots, d.$$

## 2. $J$ -invariant submanifolds

A submanifold  $N \subset M$  is said to be  $J$ -invariant if  $JT_x N = T_x N$  at every point  $x \in N$ .  $J$ -invariant submanifolds are even dimensional. In this section we shall discuss the properties of a system of real-valued functions

$$\rho = (\rho^1, \dots, \rho^{2d})$$

that defines a  $J$ -invariant submanifold  $N$ . The system  $\rho$  shall be called nondegenerate if

$$d\rho^1 \wedge \dots \wedge d\rho^{2d} \neq 0.$$

Given a set of finitely many differential 1-forms  $\{\phi^1, \phi^2, \dots\}$ , we shall mean by the rank at  $x \in M$  the number of independent 1-forms at  $x$ .

Now we consider a submanifold  $N^{2n}$  of  $(M^{2m}, J)$  locally defined as the common zero set of a nondegenerate set of real-valued functions  $\rho^1, \dots, \rho^{2d}$  with  $d = m - n$ .

**Proposition 2.1.** *Suppose that  $(\rho^1, \dots, \rho^{2d})$  is a nondegenerate set of real-valued functions on a neighborhood of a point  $x$  of  $(M^{2m}, J)$  with  $d \leq m$ . Then we have*

$$d \leq \text{rank}(\partial\rho^1, \dots, \partial\rho^{2d}) \leq 2d.$$

*Proof.* Consider

$$\begin{aligned} (2-1) \quad d\rho^1 \wedge \dots \wedge d\rho^{2d} &= (\partial\rho^1 + \bar{\partial}\rho^1) \wedge \dots \wedge (\partial\rho^{2d} + \bar{\partial}\rho^{2d}) \\ &= (\partial\rho^1 \wedge \dots \wedge \partial\rho^{2d}) + \text{mixed terms} + (\bar{\partial}\rho^1 \wedge \dots \wedge \bar{\partial}\rho^{2d}), \end{aligned}$$

where ‘‘mixed terms’’ means those terms that contain both  $\partial\rho^\alpha$ 's and  $\bar{\partial}\rho^\alpha$ 's. If  $\text{rank}(\partial\rho^1, \dots, \partial\rho^{2d}) \leq d - 1$  then each term in the last line of (2-1) contains either  $\partial\rho^\alpha$ 's more than  $d$  times or  $\bar{\partial}\rho^\alpha$ 's more than  $d$  times. Hence, each term of the last line of (2-1) is zero at  $x$ , which contradicts the nondegeneracy condition.  $\square$

**Proposition 2.2.** *Let  $u$  be a  $C^\infty$  complex-valued function on  $M$  and  $X \in TM$ . Then*

$$\partial u(X) = \frac{1}{2}\{du(X) - \sqrt{-1}du(JX)\} \quad \text{and} \quad \bar{\partial}u(X) = \frac{1}{2}\{du(X) + \sqrt{-1}du(JX)\}.$$

*Proof.* Since  $\partial u$  annihilates any  $(0, 1)$ -vector, we have

$$\begin{aligned} \partial u(X) &= \partial u(X^{1,0} + X^{0,1}) = \partial u(X^{1,0}) = du(X^{1,0}) \\ &= \frac{1}{2}du\{X - \sqrt{-1}JX\} = \frac{1}{2}\{du(X) - \sqrt{-1}du(JX)\}. \end{aligned}$$

We prove the second equality similarly.  $\square$

**Theorem 2.3.** *Let  $N^{2n}$  be a submanifold of  $(M^{2m}, J)$  given as a common zero set of a nondegenerate system of real-valued functions  $\rho^1, \dots, \rho^{2d}$  with  $d = m - n$ . Let  $T^{1,0}N = \{X - \sqrt{-1}JX : X \in TN \cap JTN\}$  and  $T^{0,1}N = \{X + \sqrt{-1}JX : X \in TN \cap JTN\}$ . Then the following are equivalent:*

- (i)  $N$  is  $J$ -invariant.
- (ii)  $T_x^{1,0}N$  and  $T_x^{0,1}N$  have complex dimension  $n$  for each  $x \in N$ .
- (iii)  $\text{rank}(\partial\rho^1, \dots, \partial\rho^{2d})(x) = d$  for each  $x \in N$ .

*Proof.* (i)  $\Rightarrow$  (ii): Suppose that  $N$  is  $J$ -invariant. Then it is easy to see that there exist linearly independent real vector fields

$$X_1, JX_1, \dots, X_n, JX_n$$

that are tangent to  $N$ . Thus  $2n$  complex vectors  $X_k^{1,0} := \frac{1}{2}(X_k - \sqrt{-1}JX_k)$  and  $X_k^{0,1} := \frac{1}{2}(X_k + \sqrt{-1}JX_k)$ ,  $k = 1, \dots, n$ , are linearly independent and tangent to  $N$ , which implies (ii).

(ii)  $\Rightarrow$  (iii): Suppose that, for each  $x \in N$ ,  $T_x^{1,0}N$  has complex dimension  $n$ . Since

$$\begin{aligned} T^{1,0}N &= \{Z \in T^{1,0}M : d\rho^\alpha(Z) = \partial\rho^\alpha(Z) = 0, \alpha = 1, \dots, 2d\} \\ &= \bigcap_{\alpha=1}^{2d} (\text{Ker } \partial\rho^\alpha \cap T^{1,0}M) = \left( \bigcap_{\alpha=1}^{2d} (\text{Ker } \partial\rho^\alpha) \right) \cap T^{1,0}M \end{aligned}$$

has a fiber of complex dimension  $n$  at each point  $x \in N$ , it follows that  $(\partial\rho^1, \dots, \partial\rho^{2d})$  has rank  $m - n = d$  at  $x$ .

(iii)  $\Rightarrow$  (i): Since  $T_x^{1,0}M$  is of complex dimension  $m$  and  $(\partial\rho^1, \dots, \partial\rho^{2d})$  has rank  $d$ , the intersection of the null spaces of  $\partial\rho^\alpha : T_x^{1,0}M \rightarrow \mathbb{C}$ ,  $\alpha = 1, \dots, 2d$ , is of complex dimension  $m - d = n$ , and therefore contains linearly independent vectors  $X_1^{1,0}, \dots, X_n^{1,0}$ , where  $X_k^{1,0} = \frac{1}{2}(X_k - \sqrt{-1}JX_k)$  for some real vector  $X_k$ . Then for each  $\alpha = 1, \dots, 2d$  and each  $k = 1, \dots, n$  we have, by Proposition 2.2,

$$0 = \partial\rho^\alpha(X_k^{1,0}) = \partial\rho^\alpha(X_k) = \frac{1}{2}(d\rho^\alpha(X_k) - \sqrt{-1}d\rho^\alpha(JX_k)),$$

which implies that  $d\rho^\alpha(X_k) = 0$  and  $d\rho^\alpha(JX_k) = 0$  since the  $\rho^\alpha$  are real-valued functions. Therefore,  $\{X_k, JX_k : k = 1, \dots, n\}$  are tangent to  $N$ . Since  $\{X_k^{1,0}, k = 1, \dots, n\}$  are independent, the set of vectors  $X_j, JX_j$  ( $j = 1, \dots, n$ ) forms a  $J$ -invariant basis for  $T_x N$ . Therefore,  $N$  is  $J$ -invariant.  $\square$

The  $J$ -invariance of submanifolds has been studied in [Han and Lee 2010]. As for the special cases of real codimension 2 ( $d = 1$ ) we have the following:

**Corollary 2.4.** *Let  $(s, t)$  be a nondegenerate set of real-valued functions of  $(M^{2m}, J)$  and let  $N^{2(m-1)}$  be the common zero set of  $s$  and  $t$ . Then  $N$  is  $J$ -invariant if and*

only if

$$(2-2) \quad \partial s \wedge \partial t \equiv 0 \pmod{(s, t)}.$$

*Proof.* In the case of  $d = 1$  in Theorem 2.3 the rank condition (iii),

$$\text{rank}(\partial s, \partial t)(x) = 1 \quad \text{for each } x \in N,$$

can be written as (2-2).  $\square$

**Theorem 2.5.** *Let  $(M^{2m}, J)$  be an almost complex manifold. A submanifold  $N^{2n}$  of real codimension  $2d$ , where  $d = m - n$ , is  $J$ -invariant if and only if  $N$  is the common zero set of a set of complex-valued functions  $\zeta = (\zeta^1, \dots, \zeta^d)$  that are  $J$ -holomorphic on the zero set.*

*Proof.* Suppose that  $N$  is a  $J$ -invariant submanifold of real codimension  $2d$ . Let  $(\rho^1, \dots, \rho^{2d})$  be a nondegenerate set of real-valued functions whose common zero set is  $N$ . Since  $(\partial\rho^1, \dots, \partial\rho^{2d})$  has rank  $d$  by Theorem 2.3, we may assume that  $\partial\rho^1 \wedge \dots \wedge \partial\rho^d \neq 0$ . Then, for each  $\alpha = 1, \dots, d$ ,  $\partial\rho^{d+\alpha}$  is a linear combination of  $(\partial\rho^1, \dots, \partial\rho^d)$ , or, equivalently,

$$(2-3) \quad \begin{bmatrix} \bar{\partial}\rho^{d+1} \\ \vdots \\ \bar{\partial}\rho^{2d} \end{bmatrix} = A \begin{bmatrix} \bar{\partial}\rho^1 \\ \vdots \\ \bar{\partial}\rho^d \end{bmatrix}$$

for some invertible matrix  $A = (a_{\beta}^{\alpha})$  of smooth functions. Define  $\zeta = (\zeta^1, \dots, \zeta^d)$  by

$$\begin{bmatrix} \zeta^1 \\ \vdots \\ \zeta^d \end{bmatrix} = \begin{bmatrix} \rho^{d+1} \\ \vdots \\ \rho^{2d} \end{bmatrix} - A \begin{bmatrix} \rho^1 \\ \vdots \\ \rho^d \end{bmatrix}.$$

By (2-3) we have

$$\bar{\partial}\zeta^{\alpha} \equiv 0 \pmod{(\rho)} \quad \text{for each } \alpha = 1, \dots, d.$$

Since  $\mathcal{F}(\rho) = \mathcal{F}(\zeta, \bar{\zeta})$  it follows that the set of complex-valued functions  $\zeta = (\zeta^1, \dots, \zeta^d)$  is  $J$ -holomorphic on the zero set. Conversely, suppose that  $\zeta = (\zeta^1, \dots, \zeta^d)$  with  $d\zeta^1 \wedge \dots \wedge d\zeta^d \neq 0$  is  $J$ -holomorphic on the zero set. Let  $\zeta^{\alpha} = s^{\alpha} + it^{\alpha}$ . Then  $\bar{\partial}\zeta^{\alpha} = \bar{\partial}s^{\alpha} + i\bar{\partial}t^{\alpha} \equiv 0 \pmod{(\zeta, \bar{\zeta})}$ , which implies

$$(2-4) \quad \partial t^{\alpha} \equiv -i\partial s^{\alpha} \pmod{(\zeta, \bar{\zeta})}.$$

Hence the rank of  $(\partial s^1, \partial t^1, \dots, \partial s^d, \partial t^d)$  is at most  $d$ . On the other hand, since

$$\begin{aligned} d\zeta^{\alpha} &\equiv \partial\zeta^{\alpha} \pmod{(\zeta, \bar{\zeta})} \\ &\equiv 2\partial s^{\alpha}, \quad \text{by (2-4),} \end{aligned}$$

we have

$$\begin{aligned} 2^d \partial s^1 \wedge \cdots \wedge \partial s^d &\equiv \partial \zeta^1 \wedge \cdots \wedge \partial \zeta^d \pmod{(\zeta, \bar{\zeta})} \\ &\equiv d\zeta^1 \wedge \cdots \wedge d\zeta^d \pmod{(\zeta, \bar{\zeta})} \\ &\neq 0. \end{aligned}$$

Hence  $(\partial s^1, \partial t^1, \dots, \partial s^d, \partial t^d)$  has rank  $d$ . Then it follows from Theorem 2.3 that  $N$  is  $J$ -invariant.  $\square$

### 3. Nonlinearly perturbed Cauchy–Riemann equations

Let  $z = (z^1, \dots, z^n) \in \mathbb{C}^n$ ,  $n \geq 1$ ,  $z^j = x^j + \sqrt{-1}y^j$ . In this section we discuss the local solvability of the quasilinear Cauchy–Riemann equations for one unknown function  $w$ :

$$(3-1) \quad \frac{\partial w}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k(z, \bar{z}, w, \bar{w}) \frac{\partial w}{\partial z^k} = B_j(z, \bar{z}, w, \bar{w}), \quad j = 1, \dots, n,$$

where  $A_j^k$  and  $B_j$  are complex-valued  $C^\infty$  functions defined on a neighborhood of the origin of  $\mathbb{C}^{n+1} = \{(z, w)\}$ .

We consider a system  $(Z_1, \dots, Z_{n+1})$  of complex vector fields whose complex conjugates are given by

$$(3-2) \quad \bar{Z}_j = \frac{\partial}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k \frac{\partial}{\partial z^k} + B_j \frac{\partial}{\partial w}, \quad j = 1, \dots, n, \quad \bar{Z}_{n+1} = \frac{\partial}{\partial \bar{w}}.$$

Then  $\bar{Z}_1, \dots, \bar{Z}_{n+1}$  are annihilated by the following set of independent 1-forms:

$$(3-3) \quad \theta^\alpha = dz^\alpha - \sum_{j=1}^n A_j^\alpha d\bar{z}^j, \quad \alpha = 1, \dots, n, \quad \theta^{n+1} = dw - \sum_{j=1}^n B_j d\bar{z}^j.$$

Let  $\langle \theta \rangle$  be the linear span of  $\theta = (\theta^1, \dots, \theta^{n+1})$ . Then there exists the unique almost complex structure for which  $\theta^1, \dots, \theta^{n+1}$  are  $(1, 0)$  forms, provided that

$$(3-4) \quad \langle \theta \rangle \cap \langle \bar{\theta} \rangle = \{0\},$$

where  $\bar{\theta}$  is the complex conjugate of  $\theta$ . Equation (3-4) holds if and only if  $\theta^1, \dots, \theta^{n+1}, \bar{\theta}^1, \dots, \bar{\theta}^{n+1}$  are linearly independent. Note that

$$\begin{aligned} \theta^1 \wedge \cdots \wedge \theta^{n+1} \wedge \bar{\theta}^1 \wedge \cdots \wedge \bar{\theta}^{n+1} \\ = \{1 + P(A_j^\alpha, \bar{A}_j^\alpha)\} dz^1 \wedge d\bar{z}^1 \wedge \cdots \wedge dz^n \wedge d\bar{z}^n \wedge dw \wedge d\bar{w}, \end{aligned}$$

where  $P$  is a polynomial in the arguments  $(A_j^\alpha, \bar{A}_j^\alpha)$  without a constant term. Thus (3-4) holds if

$$(3-5) \quad |P(A_j^\alpha, \bar{A}_j^\alpha)| < 1.$$

In particular, if each  $A_j^\alpha$  vanishes at the origin, then (3-5) holds. Henceforth we shall mean (3-5) by the  $A_j^\alpha$  being *sufficiently small*.

**Proposition 3.1.** *For each  $j, k = 1, \dots, n$ , let  $A_j^k$  and  $B_j$  be  $C^\infty$  complex-valued functions defined on a neighborhood of the origin of  $\mathbb{C}^{n+1} = \{(z, w)\}$ . Let  $J$  be the almost complex structure whose  $(0, 1)$ -vectors and  $(1, 0)$ -forms are given by (3-2)–(3-3) assuming  $A_j^k$  are sufficiently small. Suppose that  $J$  has type  $\nu$  and that  $I^{(\nu)}$  has rank  $q$ . Then there exist  $q$  independent  $J$ -holomorphic functions  $\zeta^1, \dots, \zeta^q$ . A function  $\zeta$  is a  $J$ -holomorphic function if and only if  $\zeta$  is holomorphic in the variables  $\zeta^1, \dots, \zeta^q$ .*

*Proof.* The first part of the conclusion follows from Theorems 1.3 and 1.4. To prove the second assertion, suppose that  $\zeta$  is  $J$ -holomorphic. Since  $d\zeta \in \underline{I}^{(\nu)}$  we have

$$(3-6) \quad d\zeta = \sum_{\alpha=1}^q a_\alpha d\zeta^\alpha$$

for some  $C^\infty$  functions  $a_\alpha$ . Without loss of generality assume

$$(z^1, \dots, z^p, \zeta^1, \dots, \zeta^q), \quad p + q = n + 1,$$

are independent functions, so that they serve as  $C^\infty$  local coordinates of  $\mathbb{C}^{n+1}$ . Then (3-6) implies

$$\frac{\partial \zeta}{\partial z^j} = \frac{\partial \zeta}{\partial \bar{z}^j} = \frac{\partial \zeta}{\partial \bar{\zeta}^\alpha} = 0$$

for all  $j = 1, \dots, p$  and  $\alpha = 1, \dots, q$ , which means that  $\zeta$  is holomorphic in  $(\zeta^1, \dots, \zeta^q)$ . Conversely, if  $\zeta$  is a function in the variables  $\zeta^1, \dots, \zeta^q$ , then we have

$$d\zeta \in \mathcal{F}(d\zeta^1, \dots, d\zeta^q) = I^{(\nu)}.$$

Therefore,  $\zeta$  is  $J$ -holomorphic. □

**Theorem 3.2.** *Under the same hypotheses as in Proposition 3.1, let  $\zeta$  be a  $J$ -holomorphic function with  $\partial \zeta / \partial w \neq 0$ . Then*

$$(3-7) \quad \zeta = \text{constant}$$

*is an implicit solution of (3-1).*

*Proof.* Since  $\partial \zeta / \partial w \neq 0$  and  $\partial \zeta / \partial \bar{w} = 0$ , by implicit function theorem we can solve (3-7) for  $w$  to have  $w = f(z, \bar{z})$ ; that is,

$$(3-8) \quad \zeta(z, \bar{z}, f(z, \bar{z}), \overline{f(z, \bar{z})}) = 0.$$

Differentiating (3-8) in  $\bar{z}^j$  and in  $z^k$ , respectively, we obtain

$$(3-9) \quad \frac{\partial \zeta}{\partial \bar{z}^j} + \frac{\partial \zeta}{\partial w} \frac{\partial f}{\partial \bar{z}^j} = 0, \quad \frac{\partial \zeta}{\partial z^k} + \frac{\partial \zeta}{\partial w} \frac{\partial f}{\partial z^k} = 0, \quad j, k = 1, \dots, n.$$



Since  $\zeta$  is  $J$ -holomorphic we have  $L_j \zeta = 0$ ,  $j = 1, \dots, n$ ; namely,

$$(3-10) \quad \frac{\partial \zeta}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k \frac{\partial \zeta}{\partial z^k} + B_j \frac{\partial \zeta}{\partial w} = 0, \quad j = 1, \dots, n.$$

From (3-9) and (3-10) it follows that

$$(3-11) \quad -\frac{\partial \zeta}{\partial w} \left( \frac{\partial f}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k \frac{\partial f}{\partial z^k} - B_j \right) = 0,$$

which implies the conclusion.  $\square$

**Theorem 3.3.** *Let  $(\bar{Z}_1, \dots, \bar{Z}_{n+1})$  and  $(\theta^1, \dots, \theta^{n+1})$  be the same as in (3-2)–(3-3) and let  $J$  be the almost complex structure with  $(0, 1)$ -vectors  $\bar{Z}_j$  (or, equivalently,  $(1, 0)$ -forms  $\theta^j$ ). Then there exists a solution  $w = f(z, \bar{z})$  of (3-1) if and only if there exists a function  $\zeta(z, \bar{z}, w, \bar{w})$  with  $\partial \zeta / \partial w \neq 0$  which is  $J$ -holomorphic on the zero set.*

*Proof.* Suppose that  $w = f(z, \bar{z})$  is a solution of (3-1). Then

$$\zeta(z, \bar{z}, w, \bar{w}) := f(z, \bar{z}) - w$$

satisfies  $\bar{Z}_j \zeta \equiv 0 \pmod{(\zeta, \bar{\zeta})}$ , for all  $j = 1, \dots, n+1$ . Conversely, suppose that  $\zeta(z, \bar{z}, w, \bar{w})$  with  $\partial \zeta / \partial w \neq 0$  is  $J$ -holomorphic on the zero set. Since  $\partial \zeta / \partial \bar{w} = 0$  on the zero set, by the implicit function theorem we can solve  $\zeta = 0$  for  $w$ , to obtain  $w = f(z, \bar{z})$ ; that is,

$$(3-12) \quad \zeta(z, \bar{z}, f(z, \bar{z}), \overline{f(z, \bar{z})}) = 0.$$

Then by differentiating (3-12) with respect to  $\bar{z}^j$  and  $z^k$  and restricting to the zero set of  $\zeta$  we have (3-9)–(3-11) and the proof is same as that of Theorem 3.2.  $\square$

For the existence of solutions of (3-1) the coefficients  $A_j^k$  and  $B_j$  must satisfy certain conditions. To discuss this we first define smooth functions  $T_{ij}^\alpha$  by setting

$$(3-13) \quad d\theta^\alpha \equiv \sum_{1 \leq i < j \leq n+1} T_{ij}^\alpha d\bar{z}^i \wedge d\bar{z}^j \pmod{(\theta)}, \quad \alpha = 1, \dots, n+1,$$

where  $z^{n+1} = w$ ,  $\bar{z}^{n+1} = \bar{w}$ . Arranging the pairs  $(ij)$  with  $i < j$  in lexicographical order, we write (3-13) in matrices as

$$\begin{bmatrix} d\theta^1 \\ \vdots \\ d\theta^{n+1} \end{bmatrix} \equiv \underbrace{\begin{bmatrix} T_{12}^1 & T_{13}^1 & \cdots & T_{n,n+1}^1 \\ \vdots & \vdots & & \vdots \\ T_{12}^{n+1} & T_{13}^{n+1} & \cdots & T_{n,n+1}^{n+1} \end{bmatrix}}_{\mathcal{T}} \begin{bmatrix} d\bar{z}^1 \wedge d\bar{z}^2 \\ d\bar{z}^1 \wedge d\bar{z}^3 \\ \vdots \\ d\bar{z}^n \wedge d\bar{z}^{n+1} \end{bmatrix} \pmod{(\theta)}.$$

The matrix  $\mathcal{T}$  of size  $(n+1) \times \binom{n+1}{2}$  shall be called the torsion of the Pfaffian system (3-3).

If  $\mathcal{T}$  has rank zero, that is, if all  $T_{ij}^\alpha$  are zero, this is the case  $I = I^{(1)}$ , and  $I$  is closed. Then by Theorem 1.2 there exist  $n+1$  independent  $J$ -holomorphic functions.

**Theorem 3.4.** *Suppose there exist  $J$ -holomorphic functions  $\zeta^1, \dots, \zeta^q$ ,  $q \leq n+1$ , with  $d\zeta^1 \wedge \dots \wedge d\zeta^q \neq 0$ . Then  $\mathcal{T}$  has rank at most  $(n+1) - q$ .*

*Proof.* For each  $\lambda = 1, \dots, q$ , let

$$(3-14) \quad d\zeta^\lambda = \sum_{\alpha=1}^{n+1} a_\alpha^\lambda \theta^\alpha$$

for some functions  $a_\alpha^\lambda$ . Applying  $d$  to (3-14) we have

$$(3-15) \quad 0 \equiv \sum_{\alpha} a_\alpha^\lambda d\theta^\alpha \pmod{(\theta)}$$

for each  $\lambda = 1, \dots, q$ . Substituting (3-13) in (3-15) we have

$$0 \equiv \sum_{\alpha=1}^{n+1} a_\alpha^\lambda \sum_{1 \leq i < j \leq n+1} T_{ij}^\alpha d\bar{z}^i \wedge d\bar{z}^j \pmod{(\theta)},$$

which is written in matrices as

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \equiv \underbrace{\begin{bmatrix} a_1^1 & \cdots & a_{n+1}^1 \\ \vdots & & \vdots \\ a_1^q & \cdots & a_{n+1}^q \end{bmatrix}}_{\mathcal{A}} \underbrace{\begin{bmatrix} T_{12}^1 & T_{13}^1 & \cdots & T_{n,n+1}^1 \\ \vdots & \vdots & & \vdots \\ T_{12}^{n+1} & T_{13}^{n+1} & \cdots & T_{n,n+1}^{n+1} \end{bmatrix}}_{\mathcal{T}} \begin{bmatrix} d\bar{z}^1 \wedge d\bar{z}^2 \\ d\bar{z}^1 \wedge d\bar{z}^3 \\ \vdots \\ d\bar{z}^n \wedge d\bar{z}^{n+1} \end{bmatrix} \pmod{(\theta)}.$$

Since  $d\zeta^1 \wedge \dots \wedge d\zeta^q \neq 0$ , (3-14) implies that  $\mathcal{A}$  has maximal rank  $q$ . Each row of  $\mathcal{A}$  gives a linear relation among the rows of  $\mathcal{T}$ . Therefore,  $\mathcal{T}$  has rank at most  $(n+1) - q$ .  $\square$

We construct the sequence (1-4) of subbundles as follows: an element  $\phi = \sum_{\alpha=1}^{n+1} a_\alpha \theta^\alpha$  of  $I$  belongs to  $I^{(1)}$  if and only if  $(a_1, \dots, a_{n+1})$  is a null vector of the matrix  $\mathcal{T}$ , because

$$\begin{aligned} d\phi &\equiv \sum_{\alpha=1}^{n+1} a_\alpha d\theta^\alpha \pmod{(\theta)} \\ &\equiv \sum_{1 \leq i < j \leq n+1} \sum_{\alpha=1}^{n+1} a_\alpha T_{ij}^\alpha d\bar{z}^i \wedge d\bar{z}^j \pmod{(\theta)} \end{aligned}$$

is zero if and only if  $(a_1, \dots, a_{n+1})$  is a null vector of  $\mathcal{T}$ ; that is,

$$\sum_{\alpha=1}^{n+1} a_\alpha T_{ij}^\alpha = 0 \quad \text{for all pairs } (ij).$$

Inductively, let  $\phi = (\phi^1, \dots, \phi^r)$  be a set of generators of  $I^{(k)}$ . Then a 1-form  $\psi = \sum_{\beta=1}^r b_\beta \phi^\beta$  is an element of  $I^{(k+1)}$  if and only if  $(b_1, \dots, b_r)$  is a null vector of the torsion matrix of the Pfaffian system  $\phi$ . Now suppose that (3-3) has derived length  $\nu$ . In the construction of  $I^{(\nu)}$  the coefficients  $A_j^k, B_j$  in (3-3) are differentiated up to  $\nu$  times and then the condition

$$(3-16) \quad dI^{(\nu)} \subset (I^{(\nu)})$$

raises the order of the derivatives by one. Thus we have:

**Proposition 3.5.** *Let  $J$  be the almost complex structure on  $\mathbb{C}^{n+1}$  whose  $(1, 0)$ -forms are given in (3-3). Then its type condition is a system of partial differential equations on  $(A_j^k, B_j)$ : condition (3-16) being of type  $\nu$  is a PDE system of order  $\nu + 1$ . If  $J$  has type  $\nu$  and  $I^{(\nu)}$  has rank  $q$ , then there exists a complex  $q$ -parameter family of solutions of (3-1).*

Summarizing our previous discussions in Theorems 3.3 and 2.5 and Corollary 2.4 we have:

**Theorem 3.6.** *Given a system of quasilinear Cauchy–Riemann equations (3-1) with coefficients  $A_j^\alpha$  sufficiently small, let  $J$  be the almost complex structure on  $\mathbb{C}^{n+1} = \{(z, w)\}$ ,  $z = (z^1, \dots, z^n)$ ,  $n \geq 2$ , with  $(1, 0)$ -forms (3-3). Then (3-1) has a solution if and only if there exists a nondegenerate system of real-valued functions  $(s, t)$  having the following properties:*

- (i) *The determinant of any square submatrices of maximal size of  $\mathcal{T}$  is zero modulo  $(s, t)$ .*
- (ii)  $\partial s \wedge \partial t \equiv 0 \pmod{(s, t)}$ .

Condition (ii) means that the common zero set of  $s$  and  $t$  is  $J$ -invariant. Condition (i) means that we construct  $s$  and  $t$  by finding a nondegenerate set of real-valued functions that generates an ideal to which the determinants of  $n \times n$  submatrices of the torsion belong.

#### 4. Cases of several unknown functions

Our arguments of the previous section can easily be generalized to the cases of several unknown functions. Let  $z = (z^1, \dots, z^n) \in \mathbb{C}^n$ ,  $n \geq 1$ ,  $z^j = x^j + \sqrt{-1}y^j$ . We consider the system of quasilinear Cauchy–Riemann equations for  $w = (w^1, \dots, w^d)$ ,

$d \geq 2$ :

$$(4-1) \quad \frac{\partial w^\alpha}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k(z, \bar{z}, w, \bar{w}) \frac{\partial w^\alpha}{\partial z^k} = B_j^\alpha(z, \bar{z}, w, \bar{w})$$

for each  $j = 1, \dots, n$  and  $\alpha = 1, \dots, d$ , where  $A_j^k$  and  $B_j^\alpha$  are complex-valued  $C^\infty$  functions defined on a neighborhood of the origin of  $\mathbb{C}^{n+d} = \{(z, w)\}$ . We consider a system  $(Z_1, \dots, Z_{n+d})$  of complex vector fields on an open neighborhood of the origin of  $\mathbb{C}^{n+d} = \{(z, w)\}$  whose complex conjugates are given by

$$(4-2) \quad \bar{Z}_j = \frac{\partial}{\partial \bar{z}^j} + \sum_{k=1}^n A_j^k \frac{\partial}{\partial z^k} + \sum_{\alpha=1}^d B_j^\alpha \frac{\partial}{\partial w^\alpha}, \quad j = 1, \dots, n,$$

$$\bar{Z}_{n+\beta} = \frac{\partial}{\partial \bar{w}^\beta}, \quad \beta = 1, \dots, d.$$

Then  $\bar{Z}_1, \dots, \bar{Z}_{n+d}$  are annihilated by the following set of independent 1-forms:

$$(4-3) \quad \theta^k = dz^k - \sum_{j=1}^n A_j^k d\bar{z}^j, \quad k = 1, \dots, n,$$

$$\theta^{n+\alpha} = dw^\alpha - \sum_{j=1}^n B_j^\alpha d\bar{z}^j, \quad \alpha = 1, \dots, d.$$

If  $A_j^k$  are sufficiently small, then the functions  $A_j^k$  and  $B_j^\alpha$  define an almost complex structure  $J$  on  $\mathbb{C}^{n+d}$ , for which  $\bar{Z}_j, \bar{Z}_{n+\beta}$  are  $(0, 1)$ -vector fields, or, equivalently,  $\theta^k, \theta^{n+\alpha}$  are  $(1, 0)$ -forms. The following is a generalization of Theorem 3.3.

**Theorem 4.1.** *Let  $(\bar{Z}_1, \dots, \bar{Z}_{n+d})$  and  $(\theta^1, \dots, \theta^{n+d})$  be the same as in (4-2)–(4-3) and let  $J$  be the almost complex structure with  $(0, 1)$ -vectors  $\bar{Z}_j$  (or, equivalently,  $(1, 0)$ -forms  $\theta^j$ ). Then there exists a set of solutions  $w^\alpha = f^\alpha(z, \bar{z})$ ,  $\alpha = 1, \dots, d$ , of (4-1) if and only if there exists a set of functions  $\zeta^\alpha(z, \bar{z}, w, \bar{w})$ ,  $\alpha = 1, \dots, d$  with  $\det(\partial \zeta^\alpha / \partial w^\beta) \neq 0$  that is  $J$ -holomorphic on the zero set.*

*Proof.* Suppose  $w^\alpha = f^\alpha(z, \bar{z})$ ,  $\alpha = 1, \dots, d$ , is a solution of (4-1). Let  $\zeta^\alpha = f^\alpha(z, \bar{z}) - w^\alpha$ . Then, for each  $j = 1, \dots, n$  and each  $\alpha = 1, \dots, d$ , we have

$$\bar{Z}_j \zeta^\alpha = \frac{\partial f^\alpha}{\partial \bar{z}^j} + A_j^k \frac{\partial f^\alpha}{\partial z^k} - B_j^\alpha \equiv 0 \pmod{(\zeta, \bar{\zeta})}, \quad \bar{Z}_{n+\beta} \zeta^\alpha = 0.$$

Therefore,  $\zeta = (\zeta^1, \dots, \zeta^d)$  is  $J$ -holomorphic on the zero set that satisfies the nondegeneracy condition as in the statement of the theorem. Conversely, suppose that  $\zeta = (\zeta^1, \dots, \zeta^d)$  is  $J$ -holomorphic on the zero set as in the statement of the theorem. Since  $\det(\partial \zeta^\alpha / \partial w^\beta) \neq 0$  and  $\partial \zeta^\alpha / \partial \bar{w}^\beta = 0 \pmod{(\zeta, \bar{\zeta})}$ , we can solve

$\zeta = 0$  for  $w = (w^1, \dots, w^d)$  by implicit function theorem, to obtain  $w^\alpha = f^\alpha(z, \bar{z})$ ; that is,

$$(4-4) \quad \zeta^\alpha(z, \bar{z}, f(z, \bar{z}), \overline{f(z, \bar{z})}) = 0, \quad \alpha = 1, \dots, d,$$

where  $f = (f^1, \dots, f^d)$ . By applying  $\partial/\partial\bar{z}^j$  and  $\partial/\partial z^k$ , respectively, to (4-4) we have

$$(4-5) \quad \frac{\partial\zeta^\alpha}{\partial\bar{z}^j} + \sum_{\beta=1}^d \frac{\partial\zeta^\alpha}{\partial w^\beta} \frac{\partial f^\beta}{\partial\bar{z}^j} = 0, \quad \frac{\partial\zeta^\alpha}{\partial z^k} + \sum_{\beta=1}^d \frac{\partial\zeta^\alpha}{\partial w^\beta} \frac{\partial f^\beta}{\partial z^k} = 0.$$

Then, for each  $j = 1, \dots, n$  and each  $\alpha = 1, \dots, d$ , we have

$$(4-6) \quad \bar{z}_j \zeta^\alpha = \frac{\partial\zeta^\alpha}{\partial\bar{z}^j} + \sum_{k=1}^n A_j^k(z, \bar{z}, w, \bar{w}) \frac{\partial\zeta^\alpha}{\partial z^k} + \sum_{\beta=1}^d B_j^\beta(z, \bar{z}, w, \bar{w}) \frac{\partial\zeta^\alpha}{\partial w^\beta} \\ \equiv 0 \pmod{(\zeta, \bar{\zeta})}.$$

Combining (4-5) and (4-6) we have

$$-\sum_{\beta=1}^d \frac{\partial\zeta^\alpha}{\partial w^\beta} \frac{\partial f^\beta}{\partial\bar{z}^j} - \sum_{k=1}^n \sum_{\beta=1}^d A_j^k \frac{\partial\zeta^\alpha}{\partial w^\beta} \frac{\partial f^\beta}{\partial z^k} + \sum_{\beta=1}^d B_j^\beta \frac{\partial\zeta^\alpha}{\partial w^\beta} \equiv 0 \pmod{(\zeta, \bar{\zeta})},$$

which can be written in matrices as

$$\frac{\partial\zeta}{\partial w} E \equiv \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{d \times n} \pmod{(\zeta, \bar{\zeta})},$$

where

$$\frac{\partial\zeta}{\partial w} = \begin{bmatrix} \frac{\partial\zeta^1}{\partial w^1} & \cdots & \frac{\partial\zeta^1}{\partial w^d} \\ \vdots & \ddots & \vdots \\ \frac{\partial\zeta^d}{\partial w^1} & \cdots & \frac{\partial\zeta^d}{\partial w^d} \end{bmatrix},$$

$$E = \begin{bmatrix} -\frac{\partial f^1}{\partial\bar{z}^1} - \sum_{k=1}^n A_1^k \frac{\partial f^1}{\partial z^k} + B_1^1 & \cdots & -\frac{\partial f^1}{\partial\bar{z}^n} - \sum_{k=1}^n A_n^k \frac{\partial f^1}{\partial z^k} + B_n^1 \\ \vdots & \ddots & \vdots \\ -\frac{\partial f^d}{\partial\bar{z}^1} - \sum_{k=1}^n A_1^k \frac{\partial f^d}{\partial z^k} + B_1^d & \cdots & -\frac{\partial f^d}{\partial\bar{z}^n} - \sum_{k=1}^n A_n^k \frac{\partial f^d}{\partial z^k} + B_n^d \end{bmatrix}_{d \times n}.$$

Since  $\partial\zeta/\partial w$  is invertible  $E$  is identically zero on the zero set of  $\zeta$ , which implies that  $w = f(z, \bar{z})$  is a solution of (4-1).  $\square$

Then the almost complex structure on  $\mathbb{C}^{n+d}$  whose  $(0, 1)$ -forms are given by (4-2) has torsion  $\mathcal{T}$  of dimension  $(n+d) \times \binom{n+d}{2}$ . By the same argument as in the previous section we have:

**Theorem 4.2.** *Given a system of quasilinear Cauchy–Riemann equations (4-1), let  $J$  be the almost complex structure on  $\mathbb{C}^{n+d} = \{(z, w)\}$ ,  $z = (z^1, \dots, z^n)$ ,  $w = (w^1, \dots, w^d)$ , with  $(1, 0)$ -forms (4-3). Then (4-1) has a solution if and only if there exists a nondegenerate system of real-valued functions  $(s^1, \dots, s^{2d})$  having the following properties:*

- (i) *The determinant of any  $n \times n$  submatrix of the  $(n+d) \times \binom{n+d}{2}$  matrix of the torsion  $\mathcal{T}$  is zero modulo  $(s^1, \dots, s^{2d})$ .*
- (ii)  *$(\partial s^1, \dots, \partial s^{2d})$  has rank  $d$ .*

Condition (ii) means that the common zero set of  $s^1, \dots, s^{2d}$  is  $J$ -invariant. Condition (i) means that we construct  $s^1, \dots, s^{2d}$  by finding a nondegenerate set of real-valued functions that generates an ideal to which the determinants of  $n \times n$  submatrices of the torsion belong.

## 5. Quasilinear Cauchy–Riemann equations in one complex variable

Consider the following equation for a complex-valued function  $w = w(z, \bar{z})$ :

$$(5-1) \quad \frac{\partial w}{\partial \bar{z}} + A(z, \bar{z}, w, \bar{w}) \frac{\partial w}{\partial z} = B(z, \bar{z}, w, \bar{w}), \quad |A(z, \bar{z}, w, \bar{w})| < 1.$$

This is a determined system of two real equations for two real unknown functions  $\Re w$  and  $\Im w$ . Equation (5-1) is always solvable for the following reason: in  $\mathbb{C}^2 = \{(z, w)\}$  we consider complex vector fields

$$\bar{Z}_1 = \frac{\partial}{\partial \bar{z}} + A \frac{\partial}{\partial z} + B \frac{\partial}{\partial w}, \quad \bar{Z}_2 = \frac{\partial}{\partial \bar{w}}$$

and 1-forms that annihilate  $\bar{Z}_j$ ,  $j = 1, 2$ :

$$\theta^1 = dz - A d\bar{z}, \quad \theta^2 = dw - B d\bar{z}.$$

An almost complex structure  $J$  on  $\mathbb{C}^2$  is uniquely determined by the functions  $A$  and  $B$  so that  $\bar{Z}_j$ ,  $j = 1, 2$ , are  $(0, 1)$ -vectors and  $\theta^j$ ,  $j = 1, 2$ , are  $(1, 0)$ -forms. A fundamental theorem of [Nijenhuis and Woolf 1963] states that for any real tangent vector  $V$  of  $\mathbb{C}^2$  at the origin there exists a  $J$ -holomorphic curve  $\gamma(z) = (z, f(z)) : D \rightarrow \mathbb{C}^2$ , where  $D$  is a small open disk centered at the origin in  $\mathbb{C}$ , satisfying initial conditions  $\gamma(0) = 0$  and  $d\gamma(0)(\partial/\partial x) = V$ . The graph  $\gamma$  is the zero set of

$$\zeta := f(z, \bar{z}) - w,$$

which is  $J$ -holomorphic on its zero set. Thus (5-1) is solvable by Theorem 3.3.

Now we check the type of  $J$ . Since

$$\begin{bmatrix} d\theta^1 \\ d\theta^2 \end{bmatrix} \equiv \mathcal{T}d\bar{z} \wedge d\bar{w} \pmod{(\theta)}, \quad \text{where } \mathcal{T} = \begin{bmatrix} A_{\bar{w}} \\ B_{\bar{w}} \end{bmatrix},$$

if  $A_{\bar{w}} = B_{\bar{w}} = 0$  the almost complex structure is integrable, and hence by Theorem 1.2 there exist two independent  $J$ -holomorphic functions  $(\zeta^1, \zeta^2)$ . For any function  $\zeta$  that is analytic in  $(\zeta^1, \zeta^2)$  such that  $\zeta_w \neq 0$ ,

$$\zeta = \text{constant}$$

is an implicit solution of (5-1). This is the case of type 0. Next, we assume  $A_{\bar{w}} \neq 0$ . Then  $(-B_{\bar{w}}, A_{\bar{w}})$  is a null vector of the torsion  $\mathcal{T}$ , so that

$$(5-2) \quad \phi := -B_{\bar{w}}\theta^1 + A_{\bar{w}}\theta^2$$

generates  $I^{(1)}$ . If

$$(5-3) \quad d\phi \equiv 0 \pmod{\phi}$$

then  $I^{(1)}$  is closed. Equation (5-3) is a PDE system of second order for  $A$  and  $B$ . In summary we have the following table:

rank $\mathcal{T}$	type $\nu$	number of $J$ -holomorphic functions	order of PDE for $A, B$	integrability
0	0	2	1	integrable
1	1	1	2	$I^{(1)}$ is closed

Let us consider the following special case of type 1:

$$(5-4) \quad \frac{\partial w}{\partial \bar{z}} + A(z, \bar{z}, w, \bar{w}) \frac{\partial w}{\partial z} = B(\bar{z}, w), \quad A_{\bar{w}} \neq 0.$$

Since  $B_{\bar{w}} = 0$ , from (5-2),  $\phi = A_{\bar{w}}\theta^2$  generates  $I^{(1)}$ . Then computation shows

$$d\phi \equiv 0 \pmod{(\phi)}.$$

Thus  $I^{(1)}$  has rank 1 and there is a nondegenerate  $J$ -holomorphic function  $\zeta$ . Since  $d\zeta \in \mathcal{F}(\theta^2)$  we see that  $\zeta_w \neq 0$ . Therefore,

$$\zeta = \text{constant}$$

is a complex 1-parameter family of solutions of (5-4).

## 6. Examples

**Example 6.1.** Consider the following system for  $w(z^1, \bar{z}^1, z^2, \bar{z}^2)$ :

$$(6-1) \quad \begin{aligned} \frac{\partial w}{\partial \bar{z}^1} + w \frac{\partial w}{\partial z^2} &= \frac{-2w}{1 + \bar{z}^1}, \\ \frac{\partial w}{\partial \bar{z}^2} + \bar{w} \frac{\partial w}{\partial z^1} + \bar{z}^1 \frac{\partial w}{\partial z^2} &= 0. \end{aligned}$$

Then the associated almost complex structure on  $\mathbb{C}^3 = \{(z^1, z^2, w)\}$  has  $(1, 0)$ -forms

$$\begin{aligned} \theta^1 &= dz^1 - \bar{w} d\bar{z}^2, \\ \theta^2 &= dz^2 - w d\bar{z}^1 - \bar{z}^1 d\bar{z}^2, \\ \theta^3 &= dw + \frac{2w}{1 + \bar{z}^1} d\bar{z}^1. \end{aligned}$$

Then we have

$$\begin{bmatrix} d\theta^1 \\ d\theta^2 \\ d\theta^3 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\mathcal{F}} \begin{bmatrix} d\bar{z}^1 \wedge d\bar{z}^2 \\ d\bar{z}^1 \wedge d\bar{w} \\ d\bar{z}^2 \wedge d\bar{w} \end{bmatrix} \pmod{(\theta)}.$$

Hence,  $I^{(1)}$  is spanned by  $\theta^3$ . Since

$$d\theta^3 \equiv 0 \pmod{(\theta^3)},$$

this is the case of type 1. There exists a  $J$ -holomorphic function  $\zeta$ . Since  $d\zeta$  is a nonzero multiple of  $\theta^3$  we see that  $\zeta_w \neq 0$ . Each level set  $\zeta = \text{constant}$  is an implicit solution of (6-1).

**Example 6.2.** Consider the following system for  $w(z^1, \bar{z}^1, z^2, \bar{z}^2)$ :

$$(6-2) \quad \begin{aligned} \frac{\partial w}{\partial \bar{z}^1} + w \frac{\partial w}{\partial z^2} &= \frac{-2w}{1 + \bar{z}^1}, \\ \frac{\partial w}{\partial \bar{z}^2} + \bar{w} \frac{\partial w}{\partial z^1} + [z^2 + w(1 + \bar{z}^1)] \frac{\partial w}{\partial z^2} &= 0. \end{aligned}$$

Then the associated almost complex structure on  $\mathbb{C}^3 = \{(z^1, z^2, w)\}$  has  $(1, 0)$ -forms

$$\begin{aligned} \theta^1 &= dz^1 - \bar{w} d\bar{z}^2, \\ \theta^2 &= dz^2 - w d\bar{z}^1 - [z^2 + w(1 + \bar{z}^1)] d\bar{z}^2, \\ \theta^3 &= dw + \frac{2w}{1 + \bar{z}^1} d\bar{z}^1. \end{aligned}$$



Let

$$\zeta := z^2 + w(1 + \bar{z}^1).$$

Then we see that

$$d\zeta = \theta^2 + (1 + \bar{z}^1)\theta^3 + \zeta d\bar{z}^2.$$

Therefore,  $\zeta$  is  $J$ -holomorphic on its zero set. Thus  $\zeta = 0$  is an implicit solution of (6-2).

A pseudoanalytic function in several complex variables satisfies

$$\frac{\partial w}{\partial \bar{z}^j} = \alpha_j(z)w(z) + \beta_j(z)\overline{w(z)} \quad \text{for } j = 1, \dots, n,$$

for some functions  $\alpha_j(z)$  and  $\beta_j(z)$ . See the details in Section 7.

**Example 6.3.** Consider the system (7-4) of pseudoanalytic functions in  $\mathbb{C}^2$ :

$$\frac{\partial w}{\partial \bar{z}_1} = \alpha_1(z, \bar{z})w + \beta_1(z, \bar{z})\bar{w}, \quad \frac{\partial w}{\partial \bar{z}_2} = \alpha_2(z, \bar{z})w + \beta_2(z, \bar{z})\bar{w}.$$

Let  $B_j(z, \bar{z}, w, \bar{w}) = \alpha_j(z, \bar{z})w + \beta_j(z, \bar{z})\bar{w}$  for  $j = 1, 2$ . Then the associated almost complex structure on  $\mathbb{C}^3 = \{(z^1, z^2, w)\}$  has  $(1, 0)$ -forms

$$\theta^1 = dz^1, \quad \theta^2 = dz^2, \quad \theta^3 = dw - B_1 d\bar{z}^1 - B_2 d\bar{z}^2.$$

By applying  $d$  to  $\theta^1, \theta^2, \theta^3$ , we obtain the components of the torsion  $\mathcal{T}$  as follows:

$$\begin{aligned} T_{ij}^\alpha &= 0 \quad \text{for } \alpha = 1, 2, \\ T_{12}^3 &= \frac{\partial B_1}{\partial \bar{z}^2} + \frac{\partial B_1}{\partial w} B_2 - \frac{\partial B_2}{\partial \bar{z}^1} - \frac{\partial B_2}{\partial w} B_1 \\ &= \left( \frac{\partial \alpha_1}{\partial \bar{z}^2} - \frac{\partial \alpha_2}{\partial \bar{z}^1} \right) w + \left( \frac{\partial \beta_1}{\partial \bar{z}^2} - \frac{\partial \beta_2}{\partial \bar{z}^1} + \alpha_1 \beta_2 - \alpha_2 \beta_1 \right) \bar{w}, \\ T_{13}^3 &= \frac{\partial B_1}{\partial \bar{w}} = \beta_1, \quad T_{23}^3 = \frac{\partial B_2}{\partial \bar{w}} = \beta_2. \end{aligned}$$

Then  $\mathcal{T}$  has rank 0 if and only if  $\beta_1 = \beta_2 = 0$  and

$$\frac{\partial \alpha_1}{\partial \bar{z}^2} = \frac{\partial \alpha_2}{\partial \bar{z}^1}.$$

This is the involutive case and there exist three independent pseudoholomorphic functions for the associated almost complex structure. One of them satisfies  $\partial \zeta / \partial w \neq 0$ , which gives implicit solutions  $\zeta = \text{constant}$ .

If  $\text{rank } \mathcal{T} = 1$ , then  $\phi^1 = \theta^1$  and  $\phi^2 = \theta^2$  generates  $I^{(1)}$ . Since  $d\phi^k = 0$  for  $k = 1, 2$ , this is the case of type 2. However, it is easy to check that there cannot be a function  $\zeta$  with  $\partial \zeta / \partial w \neq 0$  that is pseudoholomorphic on the zero set. Therefore, there are no solutions in this case.

## 7. Perturbed Cauchy–Riemann equations and overdetermined PDE systems: a brief survey of history

The main object of complex analysis is the family of holomorphic functions  $w = w(z)$  which are characterized by the Cauchy–Riemann equations

$$\frac{\partial w}{\partial \bar{z}^j} = 0, \quad j = 1, \dots, n.$$

It is not surprising that mathematical literature abounds in natural generalizations of the Cauchy–Riemann equations. For the cases  $n = 1$  the theories of quasiconformal functions and pseudoanalytic functions were developed in the mid-20th century [Ahlfors 1954; 1966; Bers 1956; 1977]. A quasiconformal mapping  $w = w(z)$  satisfies the Beltrami equation

$$(7-1) \quad \frac{\partial w}{\partial \bar{z}} + \mu(z) \frac{\partial w}{\partial z} = 0$$

for some complex-valued Lebesgue measurable function  $\mu(z)$  with  $|\mu(z)| < 1$ . Of central importance in the theory of quasiconformal mappings in the complex plane is the measurable Riemann mapping theorem [Morrey 1938], which generalizes the Riemann mapping theorem from conformal to quasiconformal homeomorphisms.

Pseudoanalytic functions [Bers 1953; Vekua 1962] are solutions of

$$(7-2) \quad \frac{\partial w}{\partial \bar{z}} = \alpha(z)w(z) + \beta(z)\overline{w(z)}$$

for some functions  $\alpha(z)$  and  $\beta(z)$ . Recall that every harmonic function  $\phi(x, y)$  is locally the real part of an analytic function  $h(z)$  and the complex gradient  $w(z) = \partial\phi/\partial x - i\partial\phi/\partial y$  is analytic and  $w(z) = h'(z)$ . On the other hand, a linear elliptic equation for a real-valued function  $\phi(x, y)$  of second order with Hölder continuously differentiable coefficients can be transformed into the canonical form

$$\phi_{xx} + \phi_{yy} + A\phi_x + B\phi_y = 0.$$

Then  $w := \partial\phi/\partial x - i\partial\phi/\partial y$  is a pseudoanalytic function which satisfies (7-2) with  $\alpha = -\frac{1}{4}(A + iB)$  and  $\beta = -\frac{1}{4}(A - iB)$ .

In  $\mathbb{C}^n$  with  $n \geq 2$ , fundamental questions including the Levi problem were solved by means of the inhomogeneous Cauchy–Riemann equations of  $(p, q)$ -type (see [Oka 1953; Kohn 1963; Hörmander 1965]), which are generalizations of the  $(0, 1)$ -type equations

$$\frac{\partial w}{\partial \bar{z}^j} = b_j \quad \text{for } j = 1, \dots, n.$$

A multidimensional version of (7-1) is

$$(7-3) \quad L_j w := \frac{\partial w}{\partial \bar{z}^j}(z) + \sum_{k=1}^n a_j^k(z) \frac{\partial w}{\partial z^k}(z) = 0 \quad \text{for } j = 1, \dots, n,$$

where the  $a_j^k$  vanish at the origin. If the coefficients  $a_j^k(z)$  are sufficiently smooth, it was shown in [Newlander and Nirenberg 1957] that there exist  $n$  linearly independent solutions to (7-3) if and only if  $L_j$  commutes. We state this fundamental result as Theorem 1.2.

A pseudoanalytic function in several complex variables satisfies

$$(7-4) \quad \frac{\partial w}{\partial \bar{z}^j} = \alpha_j(z)w(z) + \beta_j(z)\overline{w(z)} \quad \text{for } j = 1, \dots, n,$$

for some functions  $\alpha_j(z)$  and  $\beta_j(z)$ . The properties of solutions to (7-4) were investigated in [So'n 1990; Hayashi 1996]. In particular, it was proved in [So'n 1990] that (7-4) with  $\beta_j(z) = 0$  has the extension property: if  $D \subset \mathbb{C}^n$  is a domain and  $K$  is a compact subset of  $D$  such that  $D \setminus K$  is connected, then any solution to (7-4) with  $\beta_j(z) = 0$  on  $D \setminus K$  extends to  $D$ . The extension phenomenon of holomorphic functions in several complex variables is the special case of this extension property for (7-4) with  $\alpha_j(z) = \beta_j(z) = 0$ , which was discovered in [Hartogs 1906].

In the 1960s, the theory of overdetermined systems of linear partial differential equations was intensively studied from the algebraic viewpoint based on Spencer complexes [1962; 1965; 1969]. Quillen [1964], Goldschmidt [1967], MacKichan [1968] and Sweeney [1968] investigated the condition for the exactness of the Spencer complexes. The question was whether the Spencer complex is exact if it is elliptic. When the coefficients are real analytic the Spencer complex is exact in the elliptic case. In the  $C^\infty$  category, however, one needs to prove an estimate that implies the solvability of the associated Neumann boundary value problem. It turned out that, if the linear elliptic overdetermined system satisfies the so-called  $\delta$ -estimate, then the Neumann problem for the elliptic system is solvable so that the Spencer sequence is exact and hence such a system is locally solvable.

### References

- [Ahlfors 1954] L. V. Ahlfors, "On quasiconformal mappings", *J. Analyse Math.* **3** (1954), 1–58. MR 16,348d Zbl 0057.06506
- [Ahlfors 1966] L. V. Ahlfors, *Lectures on quasiconformal mappings*, Van Nostrand Mathematical Studies **10**, D. Van Nostrand, Princeton, 1966. MR 34 #336 Zbl 0138.06002
- [Ahn and Han 2012] H. Ahn and C.-K. Han, "Local geometry of Levi-forms associated with the existence of complex submanifolds and the minimality of generic CR manifolds", *J. Geom. Anal.* **22:2** (2012), 561–582. MR 2891737 Zbl 06112517

- [Berhanu et al. 2008] S. Berhanu, P. D. Cordaro, and J. Hounie, *An introduction to involutive structures*, New Mathematical Monographs **6**, Cambridge University Press, 2008. MR 2009b:32048 Zbl 1151.35011
- [Bers 1953] L. Bers, *Theory of pseudo-analytic functions*, New York University, 1953. MR 15,211c Zbl 0051.31603
- [Bers 1956] L. Bers, “An outline of the theory of pseudoanalytic functions”, *Bull. Amer. Math. Soc.* **62** (1956), 291–331. MR 18,470d Zbl 0072.07703
- [Bers 1977] L. Bers, “Quasiconformal mappings, with applications to differential equations, function theory and topology”, *Bull. Amer. Math. Soc.* **83**:6 (1977), 1083–1100. MR 57 #3384 Zbl 0419.30016
- [Bryant et al. 1991] R. L. Bryant, S. S. Chern, R. B. Gardner, H. L. Goldschmidt, and P. A. Griffiths, *Exterior differential systems*, Mathematical Sciences Research Institute Publications **18**, Springer, New York, 1991. MR 92h:58007 Zbl 0726.58002
- [Douglis and Nirenberg 1955] A. Douglis and L. Nirenberg, “Interior estimates for elliptic systems of partial differential equations”, *Comm. Pure Appl. Math.* **8** (1955), 503–538. MR 17,743b Zbl 0066.08002
- [Gardner 1967] R. B. Gardner, “Invariants of Pfaffian systems”, *Trans. Amer. Math. Soc.* **126** (1967), 514–533. MR 35 #2233 Zbl 0161.41301
- [Gilbarg and Trudinger 1998] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, revised 2nd ed., Grundlehren der Math. Wissenschaften **224**, Springer, Berlin, 1998. MR 2001k:35004 Zbl 1042.35002
- [Goldschmidt 1967] H. Goldschmidt, “Existence theorems for analytic linear partial differential equations”, *Ann. of Math. (2)* **86** (1967), 246–270. MR 36 #2933 Zbl 0154.35103
- [Han and Kim 2012] C.-K. Han and H. Kim, “Holomorphic functions on almost complex manifolds”, *J. Korean Math. Soc.* **49**:2 (2012), 379–394. MR 2933604 Zbl 1238.32017
- [Han and Lee 2010] C.-K. Han and K.-H. Lee, “Integrable submanifolds in almost complex manifolds”, *J. Geom. Anal.* **20**:1 (2010), 177–192. MR 2011b:32044 Zbl 1185.32015
- [Han and Park 2013] C.-K. Han and J.-D. Park, “Method of characteristics and first integrals for systems of quasi-linear partial differential equations of first order”, 2013. Submitted for publication.
- [Hartogs 1906] F. Hartogs, “Einige Folgerungen aus der Cauchyschen Integralformel bei Funktionen mehrerer Veränderlichen”, pp. 223–242 in *Sitzungsberichte der Königlich Bayerischen Akademie der Wissenschaften* (München), Mathematisch-Physikalische Klasse **36**, München, Berlin, 1906. JFM 37.0443.01
- [Hayashi 1996] Y. Hayashi, “A global theory of some nonlinear Cauchy–Riemann system in several complex variables”, *Math. Nachr.* **178** (1996), 157–198. MR 97d:32023 Zbl 0855.35091
- [Hörmander 1965] L. Hörmander, “ $L^2$  estimates and existence theorems for the  $\bar{\partial}$  operator”, *Acta Math.* **113** (1965), 89–152. MR 31 #3691 Zbl 0158.11002
- [Kohn 1963] J. J. Kohn, “Harmonic integrals on strongly pseudo-convex manifolds, I”, *Ann. of Math. (2)* **78** (1963), 112–148. MR 27 #2999 Zbl 0161.09302
- [Kruglikov 1998] B. S. Kruglikov, “Nijenhuis tensors and obstructions to the construction of pseudo-holomorphic mappings”, *Math. Notes* **63**:4 (1998), 476–493. MR 2000f:32036 Zbl 0933.32038
- [MacKichan 1968] B. B. MacKichan, *A generalization to overdetermined systems of the notion of diagonal operators*, Ph.D. thesis, Stanford University, 1968. MR 2617842
- [Monge 1803] G. Monge, “Mémoire sur la théorie d’une équation aux dérivées partielles du premier ordre”, *Journal de l’École Polytechnique* **9** (1803), 56–99.

- [Morrey 1938] C. B. Morrey, Jr., “On the solutions of quasi-linear elliptic partial differential equations”, *Trans. Amer. Math. Soc.* **43**:1 (1938), 126–166. MR 1501936 Zbl 0018.40501
- [Mushkarov 1981] O. K. Mushkarov, “Almost complex manifolds without almost holomorphic functions”, *C. R. Acad. Bulgare Sci.* **34**:9 (1981), 1225–1228. MR 83i:53058 Zbl 0487.53030
- [Mushkarov 1986] O. K. Mushkarov, “Existence of holomorphic functions on almost complex manifolds”, *Math. Z.* **192**:2 (1986), 283–295. MR 87i:53055 Zbl 0581.53025
- [Newlander and Nirenberg 1957] A. Newlander and L. Nirenberg, “Complex analytic coordinates in almost complex manifolds”, *Ann. of Math. (2)* **65** (1957), 391–404. MR 19,577a Zbl 0079.16102
- [Nijenhuis and Woolf 1963] A. Nijenhuis and W. B. Woolf, “Some integration problems in almost-complex and complex manifolds”, *Ann. of Math. (2)* **77** (1963), 424–489. MR 26 #6992 Zbl 0115:16103
- [Oka 1953] K. Oka, “Sur les fonctions analytiques de plusieurs variables, IX: Domaines finis sans point critique intérieur”, *Jap. J. Math.* **23** (1953), 97–155. MR 17,82b Zbl 0053.24302
- [Quillen 1964] D. G. Quillen, *Formal properties of over-determined systems of linear partial differential equations*, Ph.D. thesis, Harvard University, 1964. MR 2939514
- [So’n 1990] L. H. So’n, “Extension problem in generalized complex analysis”, pp. 215–229 in *Functional analytic methods in complex analysis and applications to partial differential equations* (Trieste, 1988), edited by A. S. A. Mshimba and W. Tutschke, World Sci., River Edge, NJ, 1990. MR 93c:32019 Zbl 0947.30502
- [Spencer 1962] D. C. Spencer, “Deformation of structures on manifolds defined by transitive, continuous pseudogroups, I: Infinitesimal deformations of structure”, *Ann. of Math. (2)* **76** (1962), 306–398. MR 27 #6287a Zbl 0124.38601
- [Spencer 1965] D. C. Spencer, “Deformation of structures on manifolds defined by transitive, continuous pseudogroups, III: Structures defined by elliptic pseudogroups”, *Ann. of Math. (2)* **81** (1965), 389–450. MR 31 #4052 Zbl 0192.29603
- [Spencer 1969] D. C. Spencer, “Overdetermined systems of linear partial differential equations”, *Bull. Amer. Math. Soc.* **75** (1969), 179–239. MR 39 #3533 Zbl 0185.33801
- [Sweeney 1968] W. J. Sweeney, “The  $D$ -Neumann problem”, *Acta Math.* **120** (1968), 223–277. MR 37 #2250 Zbl 0159.38402
- [Taylor 1997] M. E. Taylor, *Partial differential equations, III: Nonlinear equations*, Applied Mathematical Sciences **117**, Springer, New York, 1997. MR 98k:35001 Zbl 0869.35004
- [Trèves 1981] F. Trèves, *Approximation and representation of functions and distributions annihilated by a system of complex vector fields*, École Polytechnique Centre de Mathématiques, Palaiseau, 1981. MR 84k:58008 Zbl 0515.58030
- [Vekua 1962] I. N. Vekua, *Generalized analytic functions*, Pergamon, London, 1962. MR 27 #321 Zbl 0100.07603

Received June 28, 2012. Revised September 26, 2012.

CHONG-KYU HAN  
DEPARTMENT OF MATHEMATICS  
SEOUL NATIONAL UNIVERSITY  
SAN 56-1, SHILLYM-DONG, GWANAK-GU  
SEOUL 151-742  
SOUTH KOREA  
ckhan@snu.ac.kr

JONG-DO PARK

DEPARTMENT OF MATHEMATICS AND RESEARCH INSTITUTE FOR BASIC SCIENCES

KYUNG HEE UNIVERSITY

SEOUL 130-701

SOUTH KOREA

[mathjdpark@khu.ac.kr](mailto:mathjdpark@khu.ac.kr)

## AN OVERDETERMINED PROBLEM IN POTENTIAL THEORY

DMITRY KHAVINSON, ERIK LUNDBERG AND RAZVAN TEODORESCU

**We investigate a problem posed by L. Hauswirth, F. Hélein, and F. Pacard (*Pacific J. Math.* 250:2 (2011), 319–334): characterize all the domains in the plane admitting a positive harmonic function that solves simultaneously the Dirichlet problem with null boundary data and the Neumann problem with constant boundary data. Hauswirth et al. suggested that essentially only three possibilities exist: the exterior of a disk, a half-plane, and a nontrivial example they found — the image of the strip  $|\Im \zeta| < \pi/2$  under  $\zeta \mapsto \zeta + \sinh \zeta$ . We partially prove their conjecture, showing that these are indeed the only possibilities if the domain is Smirnov and it is either simply connected or its complement is bounded and connected. We also show the nonexistence in  $\mathbb{R}^4$  of an analogous nontrivial example among axially symmetric domains containing their axis of symmetry.**

### 1. Introduction

In [Hauswirth et al. 2011], the authors posed the following problem: find a smooth bounded domain  $\Omega$  in a Riemannian manifold  $\mathcal{M}_g$  with metric  $g$ , such that the first eigenvalue  $\lambda_1$  of the Laplace–Beltrami operator on  $\Omega$  has a corresponding real, positive eigenfunction  $u_1$  satisfying  $u_1 = 0$ ,  $\partial u_1 / \partial n = 1$  on the boundary of  $\Omega$ . Any such domain is called *extremal* because it provides a critical point for the first eigenvalue  $\lambda_1$  of the Laplace–Beltrami operator, under the constraint of fixed total volume of  $\Omega$  (see [Hauswirth et al. 2011] and references therein).

In special cases, one can find a sequence of extremal domains  $\{\Omega_t\}$  with increasing volumes, such that the limit domain  $\Omega = \Omega_{t \rightarrow \infty}$  is unbounded, and its first eigenvalue vanishes as  $t \rightarrow \infty$ . This limit extremal domain is then called *exceptional*, and the corresponding limit function  $(u_{1,t})_{t \rightarrow \infty} \rightarrow u$  is a positive, harmonic function on  $\Omega$  which solves simultaneously the overdetermined boundary value problem with null Dirichlet data and constant Neumann data.

The problem of finding exceptional domains in  $\mathbb{R}^n$  and their corresponding functions  $u$  (called roof functions by the authors of [Hauswirth et al. 2011]) is a

---

*MSC2010:* primary 31A25, 35R35, 35N25; secondary 30C20, 30E25.

*Keywords:* exceptional domain, roof function, vortex dynamics, quadrature domain, null-quadrature domains, non-Smirnov domain, Schwarz function, free boundary.

nontrivial problem of potential theory. There is no obvious variational principle to use, on the one hand because  $\Omega$  is unbounded (so the Dirichlet energy of  $u$  [Astala et al. 2009, Chapter 1] will diverge), and, on the other hand, because the constant Neumann data constraint is not conformally invariant.

In the absence of a suitable variational formulation, we may interpret the scaling  $t \rightarrow \infty$  described above as a dynamical process, in which the pair  $(\Omega_t, u_t)$  evolves so that the limit  $t \rightarrow \infty$  solves the overdetermined problem. In other words, we can turn this observation into a constructive method for finding (building) exceptional domains. In order to do this, it is helpful to note that, upon compactification of the boundary  $\partial\Omega$  (with metric  $d\sigma^2$ ), the pair  $(\Omega, u)$  with flat metric becomes conformal to the half-cylinder  $\mathcal{N} := \mathbb{R}_+ \times \overline{\partial\Omega}$ , with metric  $ds^2 = e^{-2u}(du^2 + d\sigma^2)$ . Under this reformulation, scaling of  $(\Omega_t, u_t)_{t \rightarrow \infty}$  becomes equivalent to scaling of the metric structure given above, defined over the fixed space  $\mathcal{N}$ . This is reminiscent of the Ricci flow, in which the metric structure  $g$  evolves with respect to a deformation parameter  $t \in \mathbb{R}$  according to the equation

$$\frac{dg_{ij}}{dt} = -2R_{ij},$$

with the right side of the equation given by the covariant Ricci tensor. It is known from [Topping 2006] that for the case of a two-dimensional manifold, with metric given by

$$ds^2 = e^{-2u}(dx^2 + dy^2),$$

the Ricci flow equations reduce to a single nonlinear equation

$$\frac{\partial u}{\partial t} = \nabla_g^2 u$$

(since in two dimensions the Riemann tensor has only one independent component). This is a heat equation with the generator given by the Laplace–Beltrami operator corresponding to the metric  $ds^2$ . Therefore, if there is a stationary solution  $\partial u / \partial t \rightarrow 0$  as  $t \rightarrow \infty$ , it will correspond to the scaling of the first eigenvalue  $\lambda_1(t) \rightarrow 0$  and, by conformally mapping back  $\mathcal{N}$  using the solution  $u(t \rightarrow \infty)$ , we will obtain the solution  $(\Omega, u)$ .

In other words, we can summarize this constructive method for finding exceptional domains in  $\mathbb{R}^2$  as follows: starting from a 2-dimensional Riemannian manifold with finite volume and metric encoded through the positive real function  $u$ , and boundary set defined via  $u = 0$ , consider the time evolution given by the Ricci flow, without volume renormalization. Then the manifold will remain Riemannian at all times [ibid.], and in the  $t \rightarrow \infty$  limit, the function  $u$  will become a solution of the nonlinear Laplace–Beltrami equation. Furthermore, if  $u$  remains finite everywhere in the domain, then it is harmonic and satisfies both Dirichlet and Neumann conditions



at all finite boundary components, so it is a solution for the overdetermined potential problem. Considered together with the (boundary) point at infinity, the manifold is equivalent to a pseudosphere — flat everywhere except at the point at infinity, with overall positive curvature. (We wish to emphasize that there is no reason to assume that such constructive methods would be exhaustive.)

Thus, so motivated, it is natural to try to characterize exceptional domains in flat Euclidean spaces. Hauswirth et al. [2011] suggested that in two dimensions there are only three examples: a complement of a disk, a half-plane, and a nontrivial example they found [ibid., Section 2]: the image of the strip  $|\Im \zeta| \leq \pi/2$  under the mapping  $\zeta \mapsto \zeta + \sinh \zeta$ . They posed as an open problem to determine if these are the only examples [ibid., Section 7]. (They gave some evidence by characterizing the half-plane under a global assumption on the gradient of the roof function [ibid., Proposition 6.1].) They also posed the problem of finding nontrivial examples in higher dimensions and suggested the possibility of axially symmetric examples similar to the nontrivial example in the plane [ibid., Remark 2.1].

We address both of these problems. The paper is organized as follows. In Section 2, we review the theory of Hardy spaces in order to address a subtlety that arises in connection with the regularity of the boundary of an exceptional domain. This leads us to assume in our theorems that the domain  $\Omega$  is Smirnov. In Section 3, we characterize exteriors of disks as being the only exceptional domain whose boundary is compact. In Section 4, we establish a connection between the roof function of an exceptional domain and the so-called Schwarz function of its boundary, and we also show that the boundary of a simply connected exceptional domain  $\Omega$  can pass either once or twice through infinity. We then show that in the first case  $\Omega$  is a half-plane (Section 5) and in the second case  $\Omega$  is the nontrivial example of Hauswirth et al. (Section 6). In each of these theorems we assume that  $\Omega$  is Smirnov, but we allow the roof function to be a weak solution merely satisfying the boundary conditions almost everywhere.

In Section 7, we extend the result of Section 3 to higher dimensions. In Section 8, we show that the nontrivial example from Section 6 does not allow an extension to axially symmetric domains in four dimensions, contrary to what was suggested in [Hauswirth et al. 2011, Remark 2.1] (and we conjecture that this example has no analogues in any number of dimensions greater than two).

Sections 3 through 6 together partially confirm what was suggested in [Hauswirth et al. 2011, Section 7], under some assumptions on the topology of  $\Omega$  and assuming that  $\Omega$  is Smirnov. In Section 9, we give concluding remarks, including a conjecture that, up to similarity, there are only three finitely connected exceptional domains. The additional assumption of finite connectivity is due to a remarkable example of an infinitely connected exceptional domains that appeared in the fluid dynamics literature [Baker et al. 1976]. See Section 9 for discussion.

**Remark.** After this paper was submitted, Martin Traizet [2013] announced a more complete classification of exceptional domains after developing an exciting new connection to minimal surfaces. He characterized the three examples as the only ones having finitely many boundary components. Traizet’s preprint, which appeared while we were revising this paper, finds a new beautiful connection of the problem with the theory of minimal surfaces. From this point of view, he noticed the above-mentioned family of infinitely connected examples [Baker et al. 1976] and characterized them as the only periodic exceptional domains for which the quotient by the period has finitely many boundary components [Traizet 2013, Theorem 13]. For this latter result, he invokes a powerful theorem of W. H. Meeks and M. Wolf. Our methods mostly rely on classical function theory ( $H^p$  spaces) and potential theory and in most parts are different from Traizet’s. Interestingly, as noted by Traizet [2013, Remark 5], if one could prove his Theorem 13 by only invoking pure function theory, this would give a new and independent proof of the Meeks–Wolf result from minimal surfaces. An attractive challenge!

## 2. Classical versus weak solutions, regularity of the boundary, and Hardy spaces

From the rigidity of the Cauchy problem, one might expect to obtain, for free, regularity of the boundary of an exceptional domain (as is often the case for solutions of free boundary problems). Unfortunately, the problem at hand is complicated by a remarkable family of examples with rectifiable but nonsmooth boundaries, also known as non-Smirnov domains; see [Duren 1970, Chapter 10]. This results in adding a Smirnov condition to the assumptions on the domains if we desire to consider weak solutions, i.e., harmonic roof functions satisfying the Dirichlet and Neumann boundary conditions almost everywhere with respect to the Lebesgue measure.

In order to address this subtlety, we first give some background from  $H^p$  theory; see [Duren 1970] for details.

An analytic function  $f : \mathbb{D} \rightarrow \mathbb{C}$  is said to belong to the Hardy class  $H^p$ ,  $0 < p < \infty$ , if the integrals

$$\int_0^{2\pi} |f(re^{i\theta})|^p d\theta$$

remain bounded as  $r \rightarrow 1$ .

Recall that a Blaschke product is a function of the form

$$B(z) = z^m \prod_n \frac{|a_n|}{a_n} \frac{a_n - z}{1 - \overline{a_n}z},$$

where  $m$  is a nonnegative integer and  $\sum(1 - |a_n|) < \infty$ . The latter condition ensures convergence of the product [Duren 1970, Theorem 2.4].

A function analytic in  $\mathbb{D}$  is called an inner function if its modulus is bounded by 1 and its modulus has radial limit 1 almost everywhere on the boundary. If  $S(z)$  is an inner function without zeros, then  $S(z)$  is called a singular inner function.

An outer function for the class  $H^p$  is a function of the form

$$(2-1) \quad F(z) = e^{i\gamma} \exp \left\{ \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{it} + z}{e^{it} - z} \log \psi(t) dt \right\},$$

where  $\gamma$  is a real number,  $\psi(t) \geq 0$ ,  $\log \psi(t) \in L^1$ , and  $\psi(t) \in L^p$ .

The following theorem [Duren 1970, Chapter 2, Chapter 5] (see also [Fisher 1983]) provides the parametrization of functions in Hardy classes by their zero sets, associated singular measures, and moduli of their boundary values.

**Theorem 2.1.** *Every function  $f(z)$  of class  $H^p$  ( $p > 0$ ) has a unique (up to a unimodular constant factor) factorization of the form  $f(z) = B(z)S(z)F(z)$ , where  $B(z)$  is a Blaschke product,  $S(z)$  is a singular inner function, and  $F(z)$  is an outer function for the class  $H^p$ .*

Suppose  $\Omega$  is a Jordan domain with rectifiable boundary and  $f : \mathbb{D} \rightarrow \Omega$  is a conformal map. Then  $f' \in H^1$  by Theorem 3.12 in [Duren 1970]. By Theorem 2.1,  $f'$  has a canonical factorization  $f'(z) = B(z)S(z)F(z)$ , and since  $f$  is a conformal map  $f'$  does not vanish, so  $f'(z) = S(z)F(z)$ . Then  $\Omega$  is called a Smirnov domain if  $S(z) \equiv 1$  so that  $f'(z) = F(z)$  is purely an outer function. This definition is independent of the choice of conformal map.

There are examples of non-Smirnov domains with, as above,  $f'(z) = S(z)F(z)$ , but now  $F(z) \equiv 1$  and the singular inner function  $S(z)$  is not constant. Such examples were first constructed by M. Keldysh and M. Lavrentiev [1937] using complicated geometric arguments. Their existence was somewhat demystified by an analytic proof provided by P. Duren, H. S. Shapiro, and A. L. Shields [Duren et al. 1966; Shapiro 1966]. Like the disk, such a domain has harmonic measure at zero (assuming  $f(0) = 0$ ) proportional to arclength. Thus, its boundary is sometimes called a pseudocircle.

Similarly, there are exterior pseudocircles, arising as the boundary of an unbounded non-Smirnov domain [Jones and Smirnov 1999] for which the harmonic measure at infinity is proportional to arclength, and thus Green's function with singularity at infinity provides a roof function that is a weak solution satisfying the boundary conditions almost everywhere. Thus, this provides a pathological example of an exceptional domain in a weak sense. In order to construct such an unbounded non-Smirnov domain, let us follow the method above-mentioned [Duren et al. 1966], which is presented in Duren's book [1970, Section 10.4]. We recall that the construction is carried out by working backwards, first writing down a singular inner function  $S(z)$  as a candidate for the derivative  $f'(z)$  of the conformal map  $f(z)$ . The difficulty is then to show that  $f(z)$  is not only analytic, but is also

univalent, so that it actually gives a conformal map from  $\mathbb{D}$  to some domain  $\Omega$ . Univalence is established using a univalence criterion of Nehari (and Ahlfors and Weill), which states that the growth condition

$$(2-2) \quad |(Sf)(z)| \leq \frac{k}{(1 - |z|^2)^2}, \quad \text{with } k < 2,$$

on the Schwarzian derivative  $Sf$  implies that  $f$  maps  $\mathbb{D}$  univalently onto a Jordan domain on the Riemann sphere.

We follow this procedure, indicating the step that needs to be modified. Start with a nondecreasing measure  $\mu \leq 0$ , singular with respect to Lebesgue measure on the circle, yet sufficiently smooth, so that it belongs to the Zygmund class  $\Lambda_*$  [Duren 1970, Section 10.4].

We will also require  $\mu$  to have first moment zero:

$$(2-3) \quad \int_0^{2\pi} e^{i\theta} d\mu(\theta) = 0.$$

This can always be achieved by symmetrizing  $\mu$  around the origin, thus replacing  $\mu$  by  $\frac{1}{2}(d\mu(\theta) + d\mu(\theta + \pi))$ . Then the center of mass is at the origin, which is (2-3).

As in [Duren 1970, p. 177], let  $F(z)$  be the Schwarz integral of  $\mu$ :

$$F(z) = \frac{1}{2\pi} \int_0^{2\pi} \frac{e^{i\theta} + z}{e^{i\theta} - z} d\mu(\theta).$$

Let  $g(z)$  be the exponential of a constant (to be chosen later) times  $F$ :

$$g(z) = \exp\{-aF(z)\}.$$

Having chosen  $\mu$  in  $\Lambda_*$  and nondecreasing, for  $a$  small enough the antiderivative of  $g$  maps the disk onto a bounded Jordan domain with rectifiable boundary [Duren 1970, Theorem 10.9]. Here is where we depart slightly from [Duren 1970], to get an *unbounded* domain as the image of  $f(z)$ . Instead of taking  $g(z)$  as a candidate for  $f'(z)$ , we take

$$f'(z) = g(z)/z^2.$$

and we check that the same estimates used in the proof of [Duren 1970, Theorem 10.9] also apply to this case. The residue of  $f'(z)$  is zero (since we made the first moment of  $\mu$  zero), so its antiderivative  $f(z)$  is analytic in  $\mathbb{D}$  except for a simple pole at  $z = 0$ . Also,  $|f'(z)| = 1$  a.e. on  $\partial\mathbb{D}$ .

A calculation shows that the Schwarzian derivative  $Sf$  of  $f$  is given by

$$(Sf)(z) = (SG)(z) + \frac{2}{z} \frac{g'(z)}{g(z)} = -aF''(z) - \frac{a^2}{2} F'(z)^2 - \frac{2a}{z} F'(z),$$

where  $G$  is an antiderivative of  $g$ . As explicitly stated in [Duren 1970, Section 10.4],

$F''(z)$ ,  $F'(z)^2$ , and  $F'(z)$  are each  $O((1 - |z|)^{-2})$ . Moreover,  $F'(0) = 0$  by the vanishing of the first moment of  $\mu$ , so that  $F'(z)/z$  is also  $O((1 - |z|)^{-2})$ . Thus, if  $a$  is chosen small enough,  $Sf$  satisfies the Nehari criterion for univalence, (2-2).

Hence,  $f(z)$  is a conformal map taking  $\{|z| < 1\}$  onto the complement of a Jordan domain with rectifiable boundary. To see why the boundary is rectifiable, note that, as stated in [Duren 1970, Section 10.4]),  $g(z) \in H^1$ , and so  $f'(z) = g(z)/z^2$  is in  $H^1$  in an annulus  $0 < r < |z| < 1$ .

This seemingly excessive construction of an exterior pseudocircle cannot be avoided by simply taking an inversion of an interior pseudocircle; the result will be non-Smirnov, but it will not be an exterior pseudocircle. Nor can one simply take the complement. As proved by P. Jones and S. Smirnov [1999], the complement of a non-Smirnov domain is often Smirnov! (This unexpected resolution of a long standing problem put to rest all hopes of characterizing the Smirnov property in terms of a boundary curve.)

**Remark.** Closely related examples of similar function-theoretic problems can be found in [Ebenfelt et al. 2002; Shahgholian 1992]. These examples of non-Smirnov exceptional domains lead to assuming  $\Omega$  is Smirnov in our main theorems; but we allow  $u$  to be a weak solution.

An alternative approach is to require the roof function  $u$  to be a classical solution that satisfies the boundary condition everywhere (and not just a.e.). Then the domain must be Smirnov. This is because the boundary is locally real-analytic, as stated in the next lemma. Thus, the boundary is smooth except possibly at infinity. The preimage of infinity under a conformal map from the disc (or a circular domain, in the multiply connected case) can consist of at most countably many points on the boundary, i.e., infinity represents at most countably many “prime ends” [Markushevich 1977]. Indeed, the complement of  $\bar{\Omega}$  is a disjoint union of domains, and for each boundary component of  $\Omega$  going to infinity there is a separate component of the complement. None of these are void, since boundary components must be petals as opposed to slits — on a slit the normal vector would jump at the finite end point of the slit (violating the local real analyticity stated in the lemma below). There can be at most countably many such boundary components, because otherwise the complement of  $\Omega$  would consist of uncountably many disjoint nonempty domains (impossible). Thus, the point at infinity can represent at most countably many prime ends (for the definition of prime end, see [Markushevich 1977, Chapter 2]). So, if the Smirnov condition is violated, meaning the derivative  $f'$  of the conformal map has a factor that is a nontrivial singular inner function, then the associated singular measure would consist of atoms. The derivative  $f'$  then decays exponentially at those points. This violates basic estimates on the derivatives of univalent functions, as given by Koebe’s distortion theorem [Pommerenke 1975, Theorem 1.3].

**Lemma 2.2.** *If  $\Omega \subset \mathbb{R}^2$  is exceptional and the roof function  $u$  is a classical solution in  $C^1(\overline{\Omega})$ , then  $\partial\Omega$  is locally real-analytic.*

*Proof.* The analytic completion  $f(z) = u + iv$  (possibly multivalued) maps  $\Omega$  into the right half-plane, since  $u$  is positive. The Neumann condition for  $u$  implies that  $|f'(z)| = 1$  on  $\partial\Omega$ . Also,  $u \in C^1(\overline{\Omega})$  implies that  $f' \in C(\overline{\Omega})$ .

Choose a point  $z_0 \in \partial\Omega$ , and let  $\zeta_0 = f(z_0)$ . Let  $g(\zeta) = f^{-1}(\zeta)$  denote the local inverse of  $f(z)$ . Choose a neighborhood  $U$  of  $\zeta$  and let  $F := \overline{U \cap \{\Re(\zeta) \geq 0\}}$ . Choose  $U$  small enough so that  $g \in C(F)$ .

Since  $|g'(\zeta)| = 1$  on  $\partial\Omega$ , we can also choose  $U$  small enough that  $g'$  does not vanish in  $F$ . This implies that  $h(\zeta) = \log(g'(\zeta))$  is analytic in the interior of  $F$  and continuous in  $F$ . Further,  $\Re\{h(\zeta)\}$  vanishes on the imaginary axis, since  $|g'(\zeta)| = 1$  there. Thus  $h(\zeta)$  extends to a neighborhood of  $\zeta_0$  by the Schwarz reflection principle. This allows us to extend  $g'(z)$ , and therefore  $g(z)$  and  $f(z)$  extend analytically across  $z_0$ , since  $u := \Re f = 0$  on  $\partial\Omega$  and  $|\nabla u| = 1$  on  $\partial\Omega$  near  $z_0$ .  $\square$

**Corollary 2.3.** *If  $\partial\Omega$  is  $C^2$ -smooth and  $\Omega$  is exceptional, then  $\partial\Omega$  is locally real-analytic.*

*Proof.*  $C^2$ -smoothness of  $\partial\Omega$  implies that  $u$  is in  $C^1(\overline{\Omega})$ . Now use Lemma 2.2.  $\square$

Using Kellogg's theorem [1929] on regularity of conformal maps up to the boundary (see also [Pommerenke 1992, Chapter 3]), one easily extends the corollary to  $C^{1,\alpha}$  boundaries ( $\alpha > 0$ ) and even  $C^1$  boundaries. We shall not pursue these details here. It would be interesting to find sharp necessary and sufficient conditions for the a priori regularity of the boundary that would guarantee the conclusion of Corollary 2.3. As we have mentioned in the beginning of this section, it is necessary to assume that the domain is Smirnov, but it is not at all obvious that this is sufficient. See [Castro and Khavinson 2013a; 2013b] for a related discussion regarding nonconstant functions in  $E^p$  classes with real boundary values.

### 3. The case when infinity is an isolated boundary point

**Theorem 3.1.** *Suppose  $\Omega$  is an exceptional domain whose complement  $\mathbb{C} \setminus \Omega$  is bounded and connected, and assume  $\Omega$  is Smirnov. Then  $\Omega$  is the exterior of a disk.*

*Proof.* Let  $u$  be a roof function for  $\Omega$ . Positivity of  $u$  implies, by Bôcher's theorem [Axler et al. 1992, Chapter 3],  $u(z) = u_0(z) + C \log|z|$  for some constant  $C$ , where  $u_0(z)$  is harmonic in  $\Omega \cup \{\infty\}$ , and  $u_0(z)$  approaches a constant at infinity (the Robin constant of  $\partial\Omega$  provided  $C = 1$ ). Thus, in view of the Dirichlet data of  $u$ ,  $u(z)$  is a multiple of the Green's function of  $\Omega$  with a pole at infinity, and taking  $v(z)$  to be the harmonic conjugate of  $u(z)/C$ , we have a conformal map  $g(z) = e^{u(z)/C + iv(z)}$  from  $\Omega$  to the exterior of the unit disk (note that  $g(z)$  is single-valued in  $\Omega$ ).

Using both the Dirichlet and Neumann data, we have  $|g'(z)| = 1/C$  a.e. on  $\partial\Omega$ , and therefore

$$|(g^{-1})'(\zeta)| = \frac{1}{|g'(g^{-1}(\zeta))|} = C$$

a.e. on  $\partial\mathbb{D}$ . Since  $g^{-1}$  has a simple pole at infinity,  $(g^{-1})'$  is analytic. Also,  $(g^{-1})'$  is in  $H^1(\mathbb{C} \setminus \mathbb{D})$  since  $\partial\Omega$  is rectifiable. Since  $\Omega$  is Smirnov, the latter function is outer and also has constant modulus on the unit circle a.e., which together imply that it is constant. (Recall from (2-1) that an outer function is determined from its boundary values.) Hence,  $g^{-1}$  is a linear function and  $\partial\Omega$  is a circle.  $\square$

We defer proving a higher-dimensional version of this result until Section 7, but we mention here that under more smoothness assumptions, the higher-dimensional case can be proved using a theorem of W. Reichel [1997].

Under additional smoothness assumptions, the hypothesis of Theorem 3.1 guarantees that  $\Omega$  is a special type of arclength quadrature domain. The following is then an immediate corollary of a result of B. Gustafsson [1987, Remark 6.1].

**Theorem 3.2.** *Suppose  $\Omega$  is a finitely connected exceptional domain, with piecewise  $C^1$  boundary, and infinity is not a point on the boundary of  $\Omega$ . Then  $\Omega$  is the exterior of a disk.*

This removes the condition that the complement of  $\Omega$  is connected.

*Proof.* We will show that  $\Omega$  is an arclength null-quadrature domain (this term is defined in Section 9.3) for analytic functions vanishing at infinity. At first, take the class of test functions to be integrated over  $\partial\Omega$  to consist of rational functions  $r(z)$  in  $\Omega$  vanishing at infinity.

Let  $f(z) = u(z) + iv(z)$  be the analytic completion of the roof function  $u$ . Note that  $f'(z)$  is single-valued (since it is the conjugate of the gradient of  $u$ ), and, by the theorem of Bôcher already cited,  $f'(z) = O(|z|^{-1})$ . The inward normal of  $\partial\Omega$  coincides with the gradient of  $u$ , which equals  $\overline{f'(z)} = 1/f'(z)$ . The unit tangent vector  $dz/ds$  is a  $90^\circ$  rotation of the normal vector  $1/f'(z)$ . Thus,  $if'(z) dz = ds$ . We then have a quadrature formula for integration of  $r(z)$  with respect to arclength:

$$(3-1) \quad \int_{\partial\Omega} r(z) ds = i \int_{\partial\Omega} r(z) f'(z) dz = 0,$$

where the vanishing of this integral is obtained by deforming the contour to infinity where  $f'(z)r(z) = O(|z|^{-2})$ . Indeed,  $r(z) = O(|z|^{-1})$  by our choice of the test class, and  $f'(z) = O(|z|^{-1})$  as mentioned above.

If the boundary of  $\Omega$  is piecewise  $C^1$ , rational functions are dense in  $E^p$  classes (see [Duren 1970, Theorem 10.7] and, for the multiply connected case, [Tumarkin and Havinson 1958a; 1958b; 1960]). In particular, rational functions  $r(z)$  vanishing

at infinity are dense in the space of functions  $E(\Omega)$  considered in [Gustafsson 1987]. Thus, (3-1) shows that  $\Omega$  is an arclength null-quadrature domain for this space of functions, and the result now follows from Remark 6.1 in [Gustafsson 1987].  $\square$

#### 4. The Schwarz function of an exceptional domain

The Schwarz function of a real-analytic curve  $\Gamma$  is the (unique and guaranteed to exist near  $\Gamma$ ) complex-analytic function that coincides with  $\bar{z}$  on  $\Gamma$ . For the basics on the Schwarz function, we refer to [Davis 1974] and [Shapiro 1992].

We recall two basic facts needed in the proof of the next proposition.

- (i) On  $\Gamma$ ,  $|S'(z)| = 1$ .
- (ii) The complex conjugate of  $\sqrt{-S'(z)}$  is normal to  $\Gamma$ .

Statement (i) follows from the chain rule and the fact that the complex conjugate of the Schwarz function,  $\overline{S(z)}$ , is an involution; see [Davis 1974, Chapter 7]. Statement (ii) follows from Formula (7.5) of the same reference, expressing the tangent vector along  $\Gamma$  (i.e., the derivative of  $z$  with respect to arclength):

$$(4-1) \quad T(z) = dz/ds = 1/\sqrt{S'(z)}.$$

**Proposition 4.1.** *If  $\Omega$  is an exceptional domain such that the roof function is a classical solution, then the  $z$ -derivative of the roof function is given by  $u_z(z) = c\sqrt{-S'(z)}$ , where  $c$  is a real constant and  $S(z)$  is the Schwarz function of  $\partial\Omega$ . In particular,  $S'(z)$  is analytic throughout  $\Omega$ .*

**Remark.** If, for instance, the constant Neumann data for the roof function is 1, then the constant above  $c = \pm\frac{1}{2}$ , where the sign depends on the orientation of the boundary.

*Proof.* Lemma 2.2 implies that  $\Gamma$  is locally real-analytic. So  $\Gamma$  has a Schwarz function  $S(z)$ . The complex conjugate of the analytic function  $u_z$  is normal to  $\Gamma$  (since  $u$  has zero Dirichlet data). In light of the constant Neumann data, we then have  $|u_z(z)| = |(u_z(z))^*| = \frac{1}{2}|u_x + iu_y| = \frac{1}{2}\sqrt{u_x^2 + u_y^2}$  is constant on  $\Gamma$ . This, along with the statements (i) and (ii) above, shows that on  $\Gamma$  the vectors  $u_z(z)$  and  $\sqrt{-S'(z)}$  are parallel and each have constant length. Therefore, for some real constant  $c$ , the equation  $u_z(z) = c\sqrt{-S'(z)}$  holds on  $\Gamma$ . But since  $u_z$  and  $\sqrt{-S'(z)}$  are both analytic, the equation is true everywhere that either side is defined. In particular, this guarantees analytic continuation of  $S'(z)$  throughout  $\Omega$ .  $\square$

Let us use the Schwarz function to give a heuristic argument that the boundary of an exceptional domain can pass through infinity at most twice. In fact, the angle between consecutive arcs at infinity must be  $\pi$  (and obviously there cannot be more than two such angles at infinity). Suppose the boundary of a domain has a corner where two arcs meet at an angle different from 0,  $\pi$ , or  $2\pi$ . Then the derivatives



of the Schwarz functions of the two arcs have a branch cut along a third arc that propagates into the domain from the corner. To see why this is the case, note that the Schwarz function of an arc can be approximated near a point by the Schwarz function of the tangent line. Thus, to first order, the jump along the branch cut is linear, so to zeroth order, the jump of  $S'$  is determined by the slopes of the tangent lines. If the angle is 0 or  $2\pi$ , then the tangent line is the same for each arc, but the orientation changes, so there is still a jump due to the sign change. In the case of an angle of  $\pi$ , both the tangent line and the orientation are unchanged. Thus, for any angle other than  $\pi$ ,  $S'(z)$  has a jump across a branch cut between the two boundary components. For an exceptional domain,  $u$  is a global solution throughout  $\Omega$ , and so Proposition 4.1 indicates that the Schwarz function cannot have such branch cuts. Thus, the angle between consecutive boundary arcs at infinity can only be  $\pi$ , and there can be at most two such angles.

In the informal argument above, we have assumed that each arc is real-analytic at infinity, so that the Schwarz function has a Taylor expansion there. A. Eremenko (private communication, 2012) related to us the following proof using ideas from [Barrett and Eremenko 2012] that extend techniques due to C. Pommerenke. Its indisputable advantage is that no regularity assumptions on  $\partial\Omega$  are required. Also, an important part of the theorem readily extends to higher dimensions.

We recall that a Martin function is a positive harmonic function  $M$  in a region  $\Omega$  with the property that for any positive harmonic function  $v$  in  $\Omega$ , the condition  $v \leq M$  implies that  $v = cM$ , where  $c > 0$  is a constant. (Martin functions are also called minimal harmonic functions, as in [Heins 1950].) Martin functions on finitely connected domains are simply Poisson kernels evaluated at points of the Martin boundary, the boundary under Carathéodory compactification (prime ends) of the domain (see [Brelot 1971]).

**Theorem 4.2** (A. Eremenko). *The roof function  $u$  of any exceptional domain  $\Omega$  is a convex combination of at most two Martin functions of  $\Omega$  at infinity. Moreover,  $u(z) = O(|z|)$ , and in two dimensions we also have  $\nabla u(z) = O(1)$  in  $\Omega$ .*

**Remark.** M. Traizet [2013] obtained the estimate  $|\nabla u| \leq 1$  in  $\Omega$  for domains with finitely many boundary components using the Phragmén–Lindelöf principle. For Smirnov domains  $\Omega$ , it suffices to show that  $u_z$  belongs to the class  $N^+$  (see [Castro and Khavinson 2013a]) in order to conclude that the analytic function  $u_z$  is bounded by 1 in  $\Omega$ . However, even this assumption is not needed here, and it is possible to establish the estimate on  $\nabla u$  in full generality. Eremenko has kindly permitted us to include his argument here.

*Proof.* First we note that, as observed in [Barrett and Eremenko 2012, Lemma 1], if  $u$  is a positive harmonic function in a disk (or a ball in higher dimensions),  $D(a, R)$

of radius  $R$  centered at  $a$ , and  $u(z_1) = 0$  for some boundary point  $z_1$ , then

$$(4-2) \quad u(a) \leq 2R|\nabla u(z_1)|.$$

This immediately follows from Harnack's inequality for  $D(a, R)$  as for  $z \in D(a, R)$

$$\frac{u(a)}{R + |z - a|} \leq \frac{u(z)}{R - |z - a|} = \frac{u(z) - u(z_1)}{R - |z - a|},$$

and letting  $z \rightarrow z_1$  establishes (4-2).

Applying (4-2) when  $a \in \Omega$  and  $R$  is the distance from  $a$  to  $\partial\Omega$ , gives  $u(a) \leq 2RC \leq 2(|a| + \text{const})C$ , where  $C$  is the constant value of the Neumann data. So  $u(z) = O(|z|)$ , as  $z \rightarrow \infty$ , showing that  $u(z)$  has order at most one.

That  $u$  is a combination of at most two Martin functions now follows directly from [Kjellberg 1950, Theorem II], which states that if  $\Omega$  has  $n \geq 2$  different (nonproportional) Martin functions of respective orders  $\rho_1, \rho_2, \dots, \rho_n$ , then

$$\sum_{j=1}^n 1/\rho_j \leq 2.$$

(This is proved using an application of Carleman's inequality [1926].) In higher dimensions, one must use [Friedland and Hayman 1976] instead of [Kjellberg 1950].

Next we show, in the two-dimensional case, the additional claim that  $\nabla u(z) = O(1)$ . Let  $R > 0$  and consider an auxiliary function

$$w_R = \frac{|\nabla u|}{u + R},$$

where  $R > 0$  is a parameter. A direct computation shows that

$$(4-3) \quad \Delta \log w_R = w_R^2,$$

and  $w_R(z) = C/R$  for  $z \in \partial D$ . We claim that

$$(4-4) \quad w_R(z) \leq 2C/R \quad \text{for } z \in \Omega,$$

from which the result follows by letting  $R \rightarrow \infty$ , which gives  $|\nabla u| \leq 2C$  in  $\Omega$ .

Suppose, contrary to (4-4), that  $w_R(z_0) > 2/R$  for some  $z_0 \in \Omega$ . Here and in the next lines, we assume for simplicity that  $C = 1$ . Let

$$v(z) = \frac{2R}{R^2 - |z - z_0|^2} \quad \text{for } z \in D(z_0, R) = \{z : |z - z_0| < R\}.$$

Obviously,  $v(z) \geq 2/R$ . A computation reveals that  $\Delta \log v = v^2$ . Let

$$K = \{z \in \Omega \cap D(z_0, R) : w_R(z) > v(z)\}.$$

We have  $z_0 \in K$ , since  $v(z_0) = 2/R$ . Let  $K_0$  be the component of  $K$  containing  $z_0$ . Then we have  $w_R(z) = v(z)$  on  $\partial K_0$ , since  $w_R(z) < v(z)$  on  $\partial\Omega \cap D(z_0, R)$  while  $v(z) = +\infty$  on  $\partial D(z_0, R)$ . On the other hand,

$$\Delta(\log w_R - \log v) \geq w_R^2 - v^2 > 0 \quad \text{in } K_0.$$

So the subharmonic function  $\log w_R - \log v$  is positive in  $K_0$  and vanishes on the boundary, a contradiction.  $\square$

**Remark.** This a priori estimate implies the following corollary showing that the boundaries of exceptional domains are extremely regular. Namely, they are locally real-analytic and can even be parametrized from the unit circle via an antiderivative of a rational function. In particular, it validates the preceding argument using the Schwarz function, and establishes that the boundary passes at most twice through infinity each time with an angle of  $\pi$ . The only additional assumptions needed here are that the domain is Smirnov (compare Section 2) and simply connected.

**Corollary 4.3.** *Let  $\Omega$  be a simply connected Smirnov domain, and let  $h(\zeta)$  be the conformal map from  $\mathbb{D}$  to  $\Omega$ . If  $\Omega$  is exceptional, then  $h'(\zeta)$  is a rational function, and we are in one of two cases:*

- (i)  $h'$  has one pole on  $\partial\mathbb{D}$ .
- (ii)  $h'$  has two poles on  $\partial\mathbb{D}$ .

*Proof.* Let  $u$  be a roof function for  $\Omega$ , and  $f(z) = u + iv$  its analytic completion. Since  $u > 0$ ,  $f(z)$  takes  $\Omega$  into the right half-plane, and  $f(h(\zeta))$  takes the unit disk  $\mathbb{D}$  into the right half-plane. Adding an imaginary constant if necessary, we may assume that  $f(h(0)) > 0$  is real. Then, by Herglotz's theorem [Hoffman 1962, Chapter 3; Duren 1970, Chapter 1], we can represent  $f(h(\zeta))$  as

$$(4-5) \quad f(h(\zeta)) = \int_{\mathbb{T}} \frac{e^{i\theta} + z}{e^{i\theta} - z} d\mu(\theta),$$

with  $\mu$  positive.

Now, since  $\Re f(h(\zeta))$  is the pullback to  $\mathbb{D}$  of the roof function  $u$ , which by Theorem 4.2 is a convex combination of at most two Martin functions,  $\mu$  consists of at most two atoms.

Thus, differentiating (4-5),

$$(4-6) \quad f'(h(\zeta)) \cdot h'(\zeta) = R(\zeta),$$

where  $R(\zeta)$  is a rational function with either one or two double poles on  $\partial\mathbb{D}$  (at the atoms of  $\mu$ ). Since by Theorem 4.2,  $f'(h(\zeta)) = 2u_z(h(\zeta))$  is a bounded analytic function in  $\mathbb{D}$  with  $|f'(h(\zeta))| = 1$  a.e. on  $\partial\Omega$ ,  $f'(h(\zeta))$  is an inner function. Moreover,  $h'(\zeta)$  is an outer function, since  $\Omega$  is Smirnov.

For a rational function such as  $R(\zeta)$ , the canonical factorization given by Theorem 2.1 reduces to

$$R(\zeta) = B(\zeta) \cdot F(\zeta),$$

with  $B$  a Blaschke product and  $F$  a (rational) outer function. (The singular factor  $S(\zeta)$  is trivial, since  $R(\zeta)$  has no essential singularities.) By the uniqueness of the canonical factorization,  $h'(\zeta)$  and  $f'(h(\zeta))$  equal  $F(\zeta)$  and  $B(\zeta)$  respectively (up to multiplication by a unimodular constant). Hence,  $h'(\zeta) = F(\zeta)$  is rational, and  $f'(h(\zeta)) = B(z)$  is a Blaschke product.  $\square$

**Remarks.** (i) It is also possible to see that  $f'(h(\zeta)) = B(z)$  has either no zeros or a single simple zero in  $\mathbb{D}$ . The reason is that the increment of the argument of  $R'(\zeta) d\zeta = dR$  over  $\partial\Omega$  is zero, since  $f'(z) dz$  is positive. Thus, the winding number of  $R'(\zeta)$  over  $\partial\Omega$  is negative one. If  $R'(\zeta)$  has one double pole on the boundary (recall that a pole on the boundary counts half [Bell 1992, p. 48]), then  $R'$  has no zeros inside  $\mathbb{D}$ . If  $R'(\zeta)$  has two double poles (each counts half), then  $R'(\zeta)$  has one simple zero.

(ii) It seems of interest to discuss to what extent the Corollary extends to the multiply connected case (Theorem 4.2 does not assume the domain is simply connected.) Let us make a few comments in this direction. Suppose  $\Omega$  is a finitely connected Smirnov exceptional domain with the boundary passing through infinity and  $n - 1$  additional boundary components that are rectifiable Jordan curves. Let  $z = h(w) : K \rightarrow \Omega$  be a conformal map from a bounded circular domain into  $\Omega$ , with the outer circle  $C_n$  mapped onto the unbounded component of  $\partial\Omega$ . Then it is possible to show that  $u(h(\zeta))$  is a Poisson integral of a positive measure  $\mu$  supported at one or two points on the circle  $C_n$ . Now  $du(h(\zeta))/d\zeta = u_z(h(\zeta))h'(\zeta)$  is a single-valued analytic function in  $K$  with at most two double poles at  $\zeta_1, \zeta_2$ , the support of  $\mu$ . Since the Poisson kernel  $\partial_n g$  of  $K$  is analytic in a neighborhood of  $K$  except for  $\zeta_1, \zeta_2$  on  $C_n$ , the function  $g(\zeta) := du(h(\zeta))/d\zeta$  extends to a meromorphic function in a neighborhood of  $C_n \setminus \{\zeta_1, \zeta_2\}$  and has poles at those points. (This gives us local real-analyticity of the contour going through  $\infty$  on  $\partial\Omega$ .) Applying the argument principle as in the first part of this remark, we see that the total increment of the argument of  $u_z(h(\zeta))h'(\zeta)$  is  $n - 2$ , and  $u_z(h(\zeta))$  has either  $n$  or  $n - 1$  zeros in  $K$ . From an extension of the factorization theorem [Khavinson 1983] to the multiply connected case, it follows that  $B(\zeta) = u_z(h(\zeta))$  is either a constant or a covering of the disk  $\mathbb{D}$  with  $n$  sheets. It is at most  $n$  sheets by the above, and at least  $n$  since each boundary component of  $K$  is mapped to the circle and must have winding number at least one. Otherwise  $B'(\zeta)$  vanishes somewhere on the boundary of  $K$ , and a local expansion of  $B'$  at that point indicates that  $B$  maps part of  $K$  outside of  $\mathbb{D}$ , a contradiction. Putting all this together, either  $h'(\zeta)$  has two double poles

or  $B(\zeta)$  is constant. Thus, if  $C_n$  passes through infinity only once, then  $\Omega$  is a half-plane.

### 5. Infinity as a single point on the boundary

We now characterize the half-plane as the only simply connected exceptional domain having infinity as a single point on the boundary. This extends Proposition 6.1 of [Hauswirth et al. 2011] by removing a hypothesis ( $\partial_x u > 0$  in  $\Omega$ ) on the roof function.

**Theorem 5.1.** *A domain  $\Omega$  as in case (i) of Corollary 4.3 is a half-plane.*

**Remarks.** (i) As mentioned in the introduction and Section 9, M. Traizet [2013] recently used minimal surfaces to establish this result under the assumption of finitely many boundary components. We note that, in the simply connected case, this assumption is stronger than ours, since infinitely many boundary components were allowed in Eremenko's result (Theorem 4.2). The final remark in the previous section explains how to use pure function theory in order to argue that  $\Omega$  is a half-plane without the assumption that  $\Omega$  is simply connected.

(ii) We have not been able to prove a higher-dimensional version of Theorem 5.1 (see also Section 7).

*Proof.* Using the same notation  $f$  and  $h$  from the proof of Corollary 4.3,

$$f(h(\zeta)) = \int_{\partial\mathbb{D}} \frac{e^{i\theta} + \zeta}{e^{i\theta} - \zeta} d\mu(\theta)$$

for some finite positive measure  $\mu$  on  $\partial\mathbb{D}$ . By assumption, we are in the case when  $h'$  has one pole, and according to the proof of Corollary 4.3,  $\mu$  is an atomic measure with a single point mass. Without loss of generality, we can place it at the point  $e^{i\theta} = 1$ .

Thus,  $f(h(\zeta)) = C \frac{1+\zeta}{1-\zeta}$ , which upon differentiation gives

$$(5-1) \quad f'(h(\zeta))h'(\zeta) = 2C \frac{1}{(1-\zeta)^2}.$$

As asserted in the proof of Corollary 4.3,  $f'(h(\zeta))$  is the Blaschke factor of the right side, which has no zeros, so  $f'(h(\zeta))$  is a unimodular constant. Therefore,  $f = u + iv$  is a linear function and  $\Omega$  is a half-plane.  $\square$

### 6. Infinity as a double point of the boundary

In this section we characterize the nontrivial example found in [Hauswirth et al. 2011]. Suppose  $\Omega$  is a simply connected domain and  $\Omega$  is exceptional. By Corollary 4.3, recall that the derivative  $h'(\zeta)$  of the conformal map from the disk onto  $\Omega$  is a rational function with either one or two double poles on  $\partial\mathbb{D}$ .

**Theorem 6.1.** *A domain  $\Omega$  as in case (ii) of Corollary 4.3 is, up to similarity, the image of the strip  $|\Im w| \leq \pi/2$  under the conformal map  $g(w) = w + \sinh w$ , while the analytic completion of the function  $u(g(w))$  is the function  $f(g(w)) = \cosh w$ .*

**Remark.** Together with Theorems 3.1, 4.2 and 5.1, this shows that, under an assumption on the topology, the image of the strip under  $\zeta \mapsto \zeta + \sinh \zeta$  is essentially the only nontrivial example of an exceptional domain in  $\mathbb{R}^2$ . The topological assumption is necessary, since there is a whole one-parameter family of nonsimilar infinitely connected exceptional domains (see Figure 3 on page 106). However, under the assumption of finitely many boundary components, the example described in Theorem 6.1 is the only nontrivial example, as recently proved in [Traizet 2013].

*Proof.* Using the same notation as in the proofs of Corollary 4.3 and Theorem 5.1, we have that  $h'(\zeta)$  is a rational function, and according to (4-6),  $f'(h(\zeta))$  is as well. This justifies applying the argument principle to study  $f(h(\zeta))$  and  $f'(h(\zeta))$ . Namely, we will prove the following.

**Claim.** *The function  $f$  solves the differential equation*

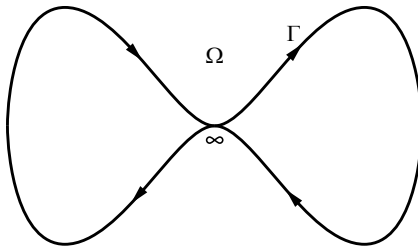
$$(6-1) \quad f' = \sqrt{(f-1)/(f+1)}, \quad z \in \Omega,$$

*after simple normalizations described below.*

Before proving this, we solve the differential equation to see that it gives the desired result. Separating variables,

$$\int \sqrt{(f+1)/(f-1)} df = z + C.$$

Making the substitution  $f = \cosh w$ , we obtain  $z = w + \sinh w$  (fixing the constant of integration  $C = 0$ ). Now using the conditions  $\Re f(z(w)) = 0$  for  $z \in \partial\Omega$  and  $\Re f(z(w)) > 0$  for  $\zeta \in \Omega$ , and the identity  $\Re \cosh(x + iy) = \cosh x \cos y$ , we find that the preimage of the domain in the  $w$ -plane is the strip  $|\Im w| \leq \pi/2$ . Therefore,  $\Omega$  can be described as the image of the strip under the map  $z(w) = w + \sinh w$ , concluding the proof of the theorem.



**Figure 1.** Local geometry of the boundary  $\Gamma = \partial\Omega$  near infinity.

To prove the claim, we will use the argument principle to show that both sides of (6-1) provide a conformal map from  $\Omega$  to  $\mathbb{D}$ . From the formula

$$T(z) = \frac{dz}{ds} = \frac{-i}{f'(z)} = \frac{1}{\sqrt{S'(z)}},$$

which relates the tangent vector  $T(z)$  on  $\partial\Omega$  to the derivative of the analytic completion  $f$  of  $u(z)$ , we obtain using the continuity of  $T(z)$  through the double point at infinity (see Figure 1) that

$$\oint_{\partial\Omega} d \log f'(z) = 2\pi i.$$

We conclude that  $f'$  is a single-sheeted covering of the unit disk by the domain  $\Omega$ , and that it has only one zero, at some point  $z_0 \in \Omega$ .

We may assume that  $f(z_0) = 1$ . If not, say  $f(z_0) = a + ib$ ,  $a > 0$ , then one may subtract the constant  $ib$  from  $f$  (this just amounts to choosing a different harmonic conjugate for the same roof function), so we have  $f(z_0) = a$ . Then one may simply replace 1 with  $a$  in the claim, and integrating the differential equation is done similarly resulting in a dilation of the original solution.

Now consider the function  $g(z) := \sqrt{(f(z) - 1)/(f(z) + 1)}$ , defined on  $\Omega$  and taking values in the unit disk  $\mathbb{D}$ . This too is a univalent map from  $\Omega$  into  $\mathbb{D}$ . Indeed, by the argument principle,  $(f(z) - 1)/(f(z) + 1)$  is a branched, two-sheeted covering of the disk, since it maps each of the two boundary components shown in Figure 1 onto  $\mathbb{T}$ , Since the single branch point  $z_0$  is mapped to the origin, taking the square root gives a single-valued analytic function.

Also,  $f'(z_0) = g(z_0) = 0$ . This uniquely determines the conformal map up to a unimodular constant, which we may assume is 1 (after a rotation), and we then arrive at the differential equation (6-1). □

### 7. An extension of Theorem 3.1 to higher dimensions

In this section, we notice that some results in Section 3 extend to higher dimensions.

**Theorem 7.1.** *Suppose  $\Omega$  is an exceptional domain in  $\mathbb{R}^n$  whose exterior is bounded and connected. If  $\partial\Omega$  is  $C^{2,\alpha}$ -smooth,  $\alpha > 0$ , then  $\partial\Omega$  is a sphere.*

*Proof.* Let  $u$  be a roof function for  $\Omega$ , and let  $v(s) = 1/|s|^{n-2}$  denote the Newtonian kernel. Fix  $y \in \Omega$  and take a small ball  $B_\varepsilon$  centered at  $y$ . Take also a large ball  $B_R$  of radius  $R$  that contains both  $B_\varepsilon$  and the complement of  $\Omega$ .

Since  $u(x)$  and  $v(x - y)$  are harmonic in  $\Omega \setminus B_\varepsilon$ , Green's second identity gives

$$(7-1) \quad \int_{\partial B_R + \partial\Omega - \partial B_\varepsilon} (v(x - y) \partial_n u(x) - u(x) \partial_n v(x - y)) d\sigma_x = 0.$$

Letting  $R \rightarrow \infty$ , we can drop the integration over  $\partial B_R$ , since again by Bôcher's theorem [Axler et al. 1992, Chapter 3], near infinity  $u(x) \approx |x|^{2-n}$ .

Since,  $u(x) = 0$  on  $\partial\Omega$  and  $\partial_n u(x) = 1$  on  $\partial\Omega$ ,

$$(7-2) \quad \int_{\partial\Omega} v(x-y) d\sigma_x = \int_{\partial B_\varepsilon} (v(x-y) \partial_n u(x) - u(x) \partial_n v(x-y)) d\sigma_x.$$

Let  $U$  be a bounded domain such that  $\mathbb{R}^n \setminus \bar{U} = \Omega$ . The outward normal for  $\partial U$  is opposite to that of  $\partial\Omega$ , and since  $v(x-y) = 1/\varepsilon^{n-2}$  on  $\partial B_\varepsilon$ ,

$$(7-3) \quad \int_{\partial U} v(x-y) d\sigma_x = \int_{\partial B_\varepsilon} (-\varepsilon^{-(n-2)} \partial_n u(x) + u(x) \partial_n v(x-y)) d\sigma_x.$$

For the first term on the right, we have

$$\int_{\partial B_\varepsilon} \varepsilon^{-(n-2)} \partial_n u(x) d\sigma_x = \int_{B_\varepsilon} \Delta u(x) dV = 0.$$

So,

$$\int_{\partial U} v(x-y) d\sigma_x = \int_{\partial B_\varepsilon} u(x) \partial_n v(x-y) d\sigma_x \rightarrow u(y),$$

as  $\varepsilon \rightarrow 0$ . So,  $u(y)$  is the single layer potential with charge density 1 on the surface  $\partial U$ . That  $U$  is a ball now follows from Theorem 1 of [Reichel 1997].  $\square$

**Remark.** Reichel's result holds for more general elliptic operators than the Laplacian. In the setting of the Laplacian, J. L. Lewis and A. Vogel [1992] characterized the sphere in terms of its interior Green's function under weaker regularity assumptions, namely, the boundary is assumed Lipschitz. In that case, the Neumann condition can be assumed to hold almost everywhere on the boundary. Thus, the hypothesis of Theorem 7.1 could be weakened by checking that the proof in [Lewis and Vogel 1992] works for the exterior case we are interested in. Yet, we have chosen an easier and more transparent path to apply Reichel's result directly, even though it requires a stronger regularity on the boundary.

## 8. Nonexistence of a higher-dimensional analog of the cosh $z$ example

The authors in [Hauswirth et al. 2011] expressed a suspicion (see Remark 2.1 in [Hauswirth et al. 2011]) that there exist  $n$ -dimensional, rotationally symmetric examples similar to the two-dimensional example  $\{(x, y) \in \mathbb{R}^2 : |y| < \pi/2 + \cosh x\}$ , which appeared in Section 6. We show that there does not exist an exceptional domain in  $\mathbb{R}^4$  whose boundary is generated by rotation about the  $x$ -axis of the (two-dimensional) graph of an even function.

**Theorem 8.1.** *There does not exist a rotationally symmetric exceptional domain  $\Omega$  in  $\mathbb{R}^4$  that contains its own axis of symmetry and whose boundary is obtained by*



rotating the (two-dimensional) graph of an even real-analytic function about the  $x$ -axis.

Our proof will rely heavily on two tricks, one exploiting the assumption that  $n = 4$ , and the other using the assumption that the generating curve is symmetric. However, we strongly suspect a more general nonexistence of such examples in  $\mathbb{R}^n$  for any  $n > 2$ .

Therefore, we conjecture the following.

**Conjecture.** *For  $n > 2$ , there does not exist an axially symmetric, exceptional domain in  $\mathbb{R}^n$  that contains its own axis of symmetry.*

**Remark.** The assumption that the domain contains its axis of symmetry rules out the exteriors of balls and circular (or spherical) cylinders, respectively (which are clearly exceptional domains as was noted in [Hauswirth et al. 2011]). Also, A. Petrosyan and K. Ramachandran pointed out to us (private communication, 2012) that the nonconvex component of the exterior of a certain cone is also an exceptional domain. In  $\mathbb{R}^4$ , using the  $x$ -axis as the axis of rotation, the cone is the rotation of  $\{(x, y) : y^2 - x^2 = 0\}$ , and the roof function in the meridian coordinates  $x, y$  where  $y$  is the distance to the  $x$ -axis in  $\mathbb{R}^4$ , is  $u(x, y) = (y^2 - x^2)/y$  for  $y > 0$ .

*Proof of Theorem 8.1.* Suppose that  $\Omega$  is such a domain in  $\mathbb{R}^4$ . Namely, the boundary  $\partial\Omega$  is obtained from rotation of  $\gamma := \{(x, y) \in \mathbb{R}^2 : y = g(x)\}$ , with  $g(-x) = g(x)$ . That is, the boundary of  $\Omega$  is given by

$$\{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : \sqrt{x_2^2 + x_3^2 + x_4^2} = g(x_1)\}.$$

Considering the boundary data, the rotational symmetry of the domain will be passed to the roof function (this requires uniqueness of the roof function guaranteed by the Cauchy–Kovalevskaya theorem, which may be applied since the boundary is assumed to be real-analytic), so that, abusing notation, we can write

$$u(x_1, x_2, x_3, x_4) = u(x, y).$$

For clarity, we emphasize that the  $x$ -axis corresponds to the axis of symmetry and the  $y$ -coordinate gives the distance from the axis of symmetry.

For axially symmetric potentials  $v$  in  $\mathbb{R}^n$ , the cylindrical reduction of Laplace's equation is

$$\Delta_{(x,y)}v + (n-2)v_y/y = 0,$$

where  $x = x_1$  and  $y = \sqrt{x_2^2 + \cdots + x_n^2}$ . Moreover, in the case we are considering, when  $n = 4$ ,  $u$  satisfies the equation  $\Delta u + (2u_y)/y = 0$ , if and only if  $yu(x, y)$  is a harmonic function of two variables  $x$  and  $y$ . Indeed,

$$\Delta(yu) = y\Delta u + 2\nabla u \cdot \nabla y + u\Delta y = y\Delta u + 2u_y.$$

(The trick that reduces axially symmetric potentials in  $\mathbb{R}^4$  to harmonic functions in the meridian plane is well known: compare [Khavinson 1991; Karp 1992].)

Since  $yu(x, y)$  is then harmonic in the unbounded two-dimensional domain  $D$  bounded by  $\gamma$  and its reflection (which we denote by  $\bar{\gamma}$ ) with respect to the  $x$ -axis, this implies  $\partial(yu(x, y))/\partial z$  is analytic in the domain  $D$ , where as usual  $z = x + iy$ . The Cauchy data (originally posed in  $\mathbb{R}^4$ ) imply that  $u_z = \frac{1}{2}(u_x - iu_y)$  coincides with  $\sqrt{-S'(z)}$  on  $\gamma$  and  $\bar{\gamma}$ . This implies that the analytic function

$$(8-1) \quad W(z) := (yu)_z = (-i/2)u + yu_z$$

coincides with  $((z - S(z))/2i)\sqrt{-S'(z)}$  on  $\gamma$  and  $\bar{\gamma}$ . The latter function is analytic, so this actually gives a formula for  $W(z)$  whose validity is not limited to  $\gamma$  and  $\bar{\gamma}$ :

$$(8-2) \quad W(z) = \frac{z - S(z)}{2i} \sqrt{-S'(z)}.$$

We note that (8-2) can be used to analytically continue  $S(z)$  to all of  $D$ , but this is not needed in our proof.

Let  $f(\zeta)$  be the conformal map from the strip  $\Sigma := \{|\Im \zeta| < \frac{1}{2}\}$  to  $D$  such that  $f(0) = 0$  and  $\arg\{f'(0)\} = 0$ . The two-fold symmetry of  $D$  implies that  $f(\zeta)$  is an odd function. Indeed, otherwise  $h(\zeta) = -f(-\zeta)$  gives another conformal map from the strip  $\Sigma$  to  $D$ . But,  $h(0) = -f(0) = 0$  and  $h'(0) = f'(0)$  implies  $h = f$ , by the uniqueness of the conformal map (up to choice of  $f(0)$  and argument of  $f'(0)$ ).

The Schwarz functions  $S_t, S_b$  of the top and bottom edges of the strip  $\Sigma$  are  $S_t(\zeta) = \zeta - i$  and  $S_b(\zeta) = \zeta + i$ . In terms of the conformal map  $f(\zeta)$ , the pullbacks to the  $\zeta$ -plane of the Schwarz functions  $S_+$  and  $S_-$  of  $\gamma$  and  $\bar{\gamma}$ , respectively, satisfy (see [Davis 1974, Chapter 8, Equation 8.7])

$$S_{\pm}(f(\zeta)) = f(\zeta \mp i) \quad \text{and} \quad S'_{\pm}(f(\zeta)) = \frac{f'(\zeta \mp i)}{f'(\zeta)}.$$

Substituting these into (8-2), we obtain two expressions for the pullback of  $W(z)$  to the strip  $\Sigma$ :

$$(8-3) \quad \frac{f(\zeta) - f(\zeta \mp i)}{2i} \sqrt{-\frac{f'(\zeta \mp i)}{f'(\zeta)}}$$

Even though  $W(f(\zeta))$  is analytic throughout  $\Sigma$ , we caution that these two expressions (one expression for  $+$  and one for  $-$ ) may only be valid near the bottom and top sides (respectively) of the strip  $\Sigma$ .

**Claim.** *The function  $W(f(\zeta))$  is odd.*

In view of the claim,  $W(0) = W(f(0)) = 0$ . By (8-1), we then have that  $(-i/2)u + yu_z = 0$  at  $z = 0$ , which implies that  $u(0, 0) = 0$ . This contradicts the positivity of  $u$ , proving the theorem.

There remains to prove the claim. Set  $V(\zeta) = W(f(\zeta)) + W(f(-\zeta))$ . Using (8-3), we can write

$$V(\zeta) = \frac{f(\zeta) - f(\zeta - i)}{2i} \sqrt{-\frac{f'(\zeta - i)}{f'(\zeta)}} + \frac{f(-\zeta) - f(-\zeta + i)}{2i} \sqrt{-\frac{f'(-\zeta + i)}{f'(-\zeta)}}$$

We show that this formula vanishes where it is valid, which then implies that  $V(\zeta)$  vanishes identically throughout  $\Sigma$ . For this, we use the fact that  $f$  is odd and consequently  $f'$  is even.

$$V(\zeta) = \frac{f(\zeta) - f(\zeta - i)}{2i} \sqrt{-\frac{f'(\zeta - i)}{f'(\zeta)}} + \frac{-f(\zeta) + f(\zeta - i)}{2i} \sqrt{-\frac{f'(\zeta - i)}{f'(\zeta)}} = 0.$$

This establishes the claim. □

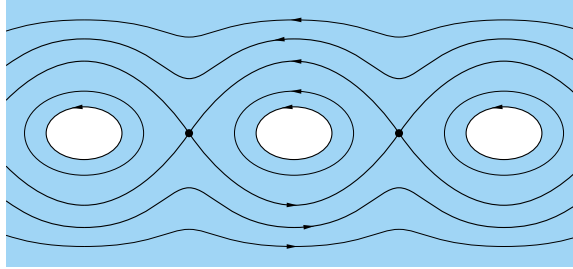
### 9. Concluding remarks and main conjecture

**9.1.** It is tempting to conjecture that the three examples in the plane studied above are the only exceptional domains in the plane, as suggested in [Hauswirth et al. 2011]. However, there is a remarkable family of infinitely connected exceptional domains. They were discovered as solutions to a fluid dynamics problem by Baker, Saffman, and Sheffield [Baker et al. 1976]. (See also [Crowdy and Green 2011] for a more detailed account.) The original problem there was to find hollow vortex equilibria with an infinite periodic array of vortices, known as spinning bubbles, amid a stationary flow of ideal fluid. The domain occupied by fluid turns out to be an exceptional domain with an infinite periodic array of holes, and the roof function is a stream function of the fluid flow; see Figure 2. The constant Dirichlet condition corresponds to the requirement that the boundary of each hollow vortex is a stream line, and the constant Neumann condition corresponds to the requirement that the fluid pressure should be balanced at the interface by the pressure inside each bubble which is assumed constant. The latter correspondence is more subtle; in order to have constant pressure along a stream line, the fluid velocity (which equals the normal derivative of stream function) should be constant, by Bernoulli's law.

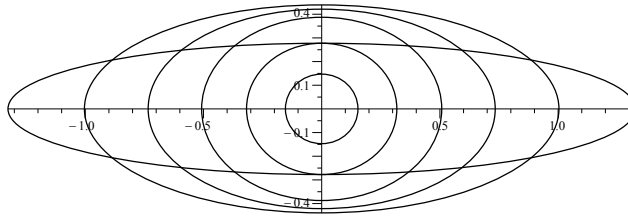
This infinitely connected example leads us to add to the conjecture the assumption that the domain is finitely connected.

**Conjecture.** *The only finitely connected exceptional domains in  $\mathbb{R}^2$  are the exterior of the unit disk, the half-plane, and the domain described in Theorem 6.1.*

**Remark.** As mentioned in the introduction, Martin Traizet [2013] recently announced a classification of exceptional domains. His results confirm our conjecture for domains having finitely many boundary components and also show that the above infinitely connected example is the only periodic exceptional domain for which the quotient by the period has finitely many boundary components. His methods



**Figure 2.** An infinitely connected exceptional domain that also provides a hollow vortex equilibrium. Level curves of the roof function are stream lines. The shape of the bubbles ensures that the pressure dictated by Bernoulli's law is constant at the fluid-bubble interface.



**Figure 3.** One-parameter family of bubble shapes associated with exceptional domains. As stated in [Traizet 2013], each of the three previously known examples can be recovered as scaling limits of this family. In that sense, this family includes all known examples.

use a remarkable nontrivial correspondence to minimal surfaces, perturbing an exceptional domain by harmonically mapping it to another domain in such a way that the graph of the new height function (which pulls back to the roof function in the original domain) satisfies the minimal surface equation. A miraculous (and crucial to his approach) by-product is that, whereas the graph of the roof function meets its boundary at a  $45^\circ$  angle, the minimal graph meets its boundary vertically so that gluing it to its own reflection over the  $xy$ -plane results in a smooth minimal surface (without boundary!) embedded in  $\mathbb{R}^3$ .

**9.2.** Regarding the higher-dimensional case, we conjecture the following extension of Theorem 5.1 to higher dimensions.

**Conjecture.** *Suppose  $\Omega$  is an exceptional domain in  $\mathbb{R}^n$  that is homeomorphic to a half-space. Then  $\Omega$  is a half-space.*

**9.3.** The connection to the Schwarz function in Section 4 reveals that exceptional domains are *arclength null-quadrature domains*. That is, for any function  $f$ , say analytic in  $\Omega$ , continuous in  $\overline{\Omega}$ , integrable over the boundary, and decaying sufficiently at infinity, we have  $\int_{\partial\Omega} f ds = 0$ . Indeed,  $\int_{\partial\Omega} f ds = \int_{\partial\Omega} f(z)(1/T(z))T(z) ds = \int_{\partial\Omega} f(z)\sqrt{S'(z)} dz$ , where  $T(z)$  is the complex unit tangent vector (see Section 4), and now this integral vanishes as long as the integrand decays sufficiently at infinity. Null-quadrature domains were previously studied in the case of area measure. They were characterized in the plane by M. Sakai [1981]. Our current study can be seen as a step toward characterizing null-quadrature domains for arclength.

**9.4.** Other interesting connections involve differentials on Riemann surfaces. The study of Gustafsson [1987] used half-order differentials on the Schottky double of an arclength quadrature domain. From a different point of view, the boundary of an exceptional domain is a trajectory of the positive quadratic differential  $-(df)^2$ , where  $f(z)$  is the analytic completion of the roof function.

**9.5.** The differential equation (6-1) can be solved by a more general substitution using Jacobi elliptic functions [Abramowitz and Stegun 1964, p. 567, §16]:

$$(9-1) \quad f(\zeta, k) = \cos \theta \operatorname{cn}(\zeta, k) + \sin \theta \operatorname{sn}(\zeta, k),$$

$$(9-2) \quad z(\zeta) = \cos \theta \operatorname{sn}(\zeta, k) - \sin \theta \operatorname{cn}(\zeta, k) + \int^\zeta \operatorname{dn}(\xi, k) d\xi,$$

where  $\theta$  is an arbitrary phase,  $\theta \in [0, 2\pi]$ .

For a given value of the elliptic modulus  $k \in [0, 1]$ , we define the corresponding domain  $\mathbb{F}$  through its fundamental periods

$$T_1(k) = 4K(\sqrt{1-k^2}) \quad \text{and} \quad T_2(k) = 4K(k),$$

where

$$K(k) \equiv \int_0^{\pi/2} \frac{1}{\sqrt{1-k^2 \sin^2 \theta}} d\theta$$

is the complete elliptic integral of the first kind [Abramowitz and Stegun 1964, p. 590, §17.3]. It diverges for  $k = 1$  and equals  $\pi/2$  for  $k = 0$ .

Then it is straightforward to check that (6-1) is satisfied by  $f(z)$ , due to the identity [Abramowitz and Stegun 1964, p. 573, §16.9]

$$1 = (-\operatorname{sn} z \cos \theta + \operatorname{cn} z \sin \theta)^2 + (\operatorname{cn} z \cos \theta + \operatorname{sn} z \sin \theta)^2.$$

Let  $\gamma$  be the preimage of  $\partial\Omega$  under  $z(\zeta)$ : it consists of two pieces  $\gamma_\pm$ ,  $\gamma_- = -\gamma_+$ , dividing the fundamental domain  $\mathbb{F}$  into three subdomains. Denote the component which contains the origin by  $D_0$ . Since  $f(0) = 1$ , we conclude that  $\Re f(z) > 0$  for  $z \in D_0 \setminus \gamma_\pm$ , and we have proven the following result.

**Proposition.** *The exceptional domain  $\Omega$  is the image of the domain  $D_0(k)$  under the map  $\zeta \mapsto z(\zeta)$  defined in (9-2).*

**Remark.** The case discussed in the proof of the theorem corresponds to the degenerate elliptic modulus  $k = 0$ . Then the domain  $\mathbb{F}$  becomes the infinite strip

$$T_1(0) = 4K(1) \rightarrow \infty, \quad T_2(0) = 4K(0) = 2\pi,$$

while the functions  $f, g$  become (using the fact that  $\operatorname{dn}(z, 1) \equiv 1$ )

$$z(\zeta) = \zeta + \sinh \zeta, \quad f(z(\zeta)) = \cosh \zeta.$$

As noted before, the conditions  $\Re f(\zeta)|_{\gamma_{\pm}} = 0$  give the preimage  $\gamma_{\pm} := z^{-1}(\partial\Omega) = \{\Im \zeta = \pm\pi/2\}$ , and the preimage of the domain,  $D_0$ , becomes the strip  $|\Im \zeta| \leq \pi/2$ .

**9.6.** The domain  $D_0(k)$  is the preimage of the unit disk under the map  $\zeta : \mathbb{F} \rightarrow \mathbb{D}$  defined by

$$\zeta(w) = \frac{\operatorname{sn}(w, k) - i}{\operatorname{sn}(w, k) + i}, \quad k \in [0, 1],$$

with the support of  $\mu$  at points  $\zeta_{\pm} = \pm(1 - ik)/(1 + ik)$ , where  $\mu$  is the measure discussed in the proof of Corollary 4.3. The case  $k \rightarrow 0$  corresponds to the strip domain and to  $\zeta_{\pm} = \pm 1$ . The reparametrization invariance noted above for the solution  $f(z)$  of (6-1) under rescaling of the elliptic modulus  $k$  is indicative of a deeper invariance of the solution: all the specific solutions in  $\mathbb{C}$  discussed here are associated with fixed points in the moduli space of Riemann surfaces.

Again let  $f(h(z))$  be the analytic completion of a solution, and denote by  $\mathcal{G}$  the group of transformations that leaves  $\operatorname{supp}(\mu)$  invariant up to a global rotation. It follows that  $f$  is an automorphism of the quotient of the group of linear fractional transformations by  $\mathcal{G}$ , which can be in general a Kleinian group. The limit set (accumulation points of the orbits of the group) can be finite (in which case it can consist of only 0, 1, or 2 points), or infinite. It is known (see [Astala et al. 2009, Theorem 10.3.4]) that the set of homeomorphic solutions for a quasilinear elliptic equation of Laplace–Beltrami type forms a group only in the case of finite limit set. The Kleinian groups are called degenerate in this case, and they correspond to either finite groups (with empty limit set), or cyclic groups (generated by one element, with limit set consisting of 1 or 2 points). These correspond to the solutions described in the present paper (isolated point at infinity, respectively simple and double boundary point at infinity).

### Acknowledgements

The authors are indebted to Dimiter Vassilev for bringing the article [Hauswirth et al. 2011] to their attention. We wish to thank Alexandre Eremenko for sharing an

improved proof of Theorem 4.2 and Arshak Petrosyan and Koushik Ramachandran for pointing out the example of a cone as an exceptional domain. We also wish to thank Martin Traizet for helpful discussion regarding his preprint. We are grateful to the anonymous referee whose careful reading of the paper and constructive criticism significantly improved the clarity of the exposition. Khavinson and Lundberg acknowledge partial support from the NSF under the grant DMS-0855597.

## References

- [Abramowitz and Stegun 1964] M. Abramowitz and I. A. Stegun (editors), *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, National Bureau of Standards Applied Mathematics Series **55**, U.S. Government Printing Office, Washington, DC, 1964. Reprinted by Dover, New York, 1974. MR 29 #4914 Zbl 0171.38503
- [Astala et al. 2009] K. Astala, T. Iwaniec, and G. Martin, *Elliptic partial differential equations and quasiconformal mappings in the plane*, Princeton Mathematical Series **48**, Princeton University Press, 2009. MR 2010j:30040 Zbl 1182.30001
- [Axler et al. 1992] S. Axler, P. Bourdon, and W. Ramey, *Harmonic function theory*, Graduate Texts in Mathematics **137**, Springer, New York, 1992. MR 93f:31001 Zbl 0765.31001
- [Baker et al. 1976] G. R. Baker, P. G. Saffman, and J. S. Sheffield, “Structure of a linear array of hollow vortices of finite cross-section”, *J. Fluid Mech.* **74**:3 (1976), 469–476. Zbl 0343.76004
- [Barrett and Eremenko 2012] M. Barrett and A. Eremenko, “Generalization of a theorem of Clunie and Hayman”, *Proc. Amer. Math. Soc.* **140**:4 (2012), 1397–1402. MR 2869124 Zbl 1248.32013
- [Bell 1992] S. R. Bell, *The Cauchy transform, potential theory, and conformal mapping*, CRC Press, Boca Raton, FL, 1992. MR 94k:30013
- [Brelot 1971] M. Brelot, *On topologies and boundaries in potential theory*, Lecture Notes in Mathematics **175**, Springer, Berlin, 1971. MR 43 #7654 Zbl 0222.31014
- [Carleman 1926] T. Carleman, “Extension d’un théorème de Liouville”, *Acta Math.* **48**:3-4 (1926), 363–366. MR 1555232 JFM 52.0316.02
- [Castro and Khavinson 2013a] L. de Castro and D. Khavinson, “Analytic functions in Smirnov classes  $E^P$  with real boundary values”, *Complex Anal. Oper. Theory* **7**:1 (2013), 101–106. MR 3010790
- [Castro and Khavinson 2013b] L. de Castro and D. Khavinson, “Analytic functions in Smirnov classes  $E^P$  with real boundary values II”, *Anal. Math. Phys.* **3**:1 (2013), 21–35. MR 3015628 Zbl 06147910
- [Crowdy and Green 2011] D. G. Crowdy and C. C. Green, “Analytical solutions for von Kármán streets of hollow vortices”, *Phys. Fluids* **23**:12 (2011), Article ID #126602.
- [Davis 1974] P. J. Davis, *The Schwarz function and its applications*, Carus Mathematical Monographs **17**, Mathematical Association of America, Buffalo, NY, 1974. MR 53 #11031 Zbl 0293.30001
- [Duren 1970] P. L. Duren, *Theory of  $H^P$  spaces*, Pure and Applied Mathematics **38**, Academic Press, New York, 1970. Reprinted by Dover, Mineola, NY, 2000. MR 42 #3552 Zbl 0215.20203
- [Duren et al. 1966] P. L. Duren, H. S. Shapiro, and A. L. Shields, “Singular measures and domains not of Smirnov type”, *Duke Math. J.* **33** (1966), 247–254. MR 33 #7506 Zbl 0174.37501
- [Ebenfelt et al. 2002] P. Ebenfelt, D. Khavinson, and H. S. Shapiro, “A free boundary problem related to single-layer potentials”, *Ann. Acad. Sci. Fenn. Math.* **27**:1 (2002), 21–46. MR 2002m:35227 Zbl 1035.31001
- [Fisher 1983] S. D. Fisher, *Function theory on planar domains: a second course in complex analysis*, Wiley, New York, 1983. Reprinted by Dover, Mineola, NY, 2007. MR 85d:30001 Zbl 0511.30022

- [Friedland and Hayman 1976] S. Friedland and W. K. Hayman, “Eigenvalue inequalities for the Dirichlet problem on spheres and the growth of subharmonic functions”, *Comment. Math. Helv.* **51**:2 (1976), 133–161. MR 54 #568 Zbl 0339.31003
- [Gustafsson 1987] B. Gustafsson, “Application of half-order differentials on Riemann surfaces to quadrature identities for arc-length”, *J. Analyse Math.* **49** (1987), 54–89. MR 89b:30032 Zbl 0652.30029
- [Hauswirth et al. 2011] L. Hauswirth, F. Hélein, and F. Pacard, “On an overdetermined elliptic problem”, *Pacific J. Math.* **250**:2 (2011), 319–334. MR 2012g:58046 Zbl 1211.35207
- [Heins 1950] M. Heins, “A lemma on positive harmonic functions”, *Ann. of Math. (2)* **52** (1950), 568–573. MR 12,259b Zbl 0045.18803
- [Hoffman 1962] K. Hoffman, *Banach spaces of analytic functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962. MR 24 #A2844 Zbl 0117.34001
- [Jones and Smirnov 1999] P. W. Jones and S. K. Smirnov, “On V. I. Smirnov domains”, *Ann. Acad. Sci. Fenn. Math.* **24**:1 (1999), 105–108. MR 2000k:30009 Zbl 0921.30004
- [Karp 1992] L. Karp, “Construction of quadrature domains in  $\mathbb{R}^n$  from quadrature domains in  $\mathbb{R}^2$ ”, *Complex Variables Theory Appl.* **17**:3-4 (1992), 179–188. MR 93e:30085 Zbl 0722.30039
- [Keldysh and Lavrentiev 1937] M. Keldysh and M. Lavrentiev, “Sur la représentation conforme des domaines limités par des courbes rectifiables”, *Ann. Sci. École Norm. Sup. (3)* **54** (1937), 1–38. MR 1509363 Zbl 0017.21702
- [Kellogg 1929] O. D. Kellogg, *Foundations of potential theory*, Springer, Berlin, 1929. Reprinted in *Grundlehren der Mathematischen Wissenschaften* **31**, 1967. MR 36 #5369 Zbl 0053.07301
- [Khavinson 1983] D. Khavinson, “Factorization theorems for different classes of analytic functions in multiply connected domains”, *Pacific J. Math.* **108**:2 (1983), 295–318. MR 85j:30009 Zbl 0494.30024
- [Khavinson 1991] D. Khavinson, “On reflection of harmonic functions in surfaces of revolution”, *Complex Variables Theory Appl.* **17**:1-2 (1991), 7–14. MR 92j:31005 Zbl 0702.31004
- [Kjellberg 1950] B. Kjellberg, “On the growth of minimal positive harmonic functions in a plane region”, *Ark. Mat.* **1** (1950), 347–351. MR 12,410f Zbl 0040.05502
- [Lewis and Vogel 1992] J. L. Lewis and A. Vogel, “On some almost everywhere symmetry theorems”, pp. 347–374 in *Nonlinear diffusion equations and their equilibrium states* (Gregynog, 1989), vol. 3, edited by N. G. Lloyd et al., *Progr. Nonlinear Differential Equations Appl.* **7**, Birkhäuser, Boston, 1992. MR 93j:35078 Zbl 0792.35009
- [Markushevich 1977] A. I. Markushevich, *Theory of functions of a complex variable*, vol. III, English ed., Chelsea, New York, 1977. MR 56 #3258
- [Pommerenke 1975] C. Pommerenke, *Univalent functions*, *Mathematische Lehrbücher* **25**, Vandenhoeck & Ruprecht, Göttingen, 1975. MR 58 #22526
- [Pommerenke 1992] C. Pommerenke, *Boundary behaviour of conformal maps*, *Grundlehren der Mathematischen Wissenschaften* **299**, Springer, Berlin, 1992. MR 95b:30008 Zbl 0762.30001
- [Reichel 1997] W. Reichel, “Radial symmetry for elliptic boundary-value problems on exterior domains”, *Arch. Rational Mech. Anal.* **137**:4 (1997), 381–394. MR 98m:35070 Zbl 0891.35006
- [Sakai 1981] M. Sakai, “Null quadrature domains”, *J. Analyse Math.* **40** (1981), 144–154. MR 84e:30069 Zbl 0483.30002
- [Shahgholian 1992] H. Shahgholian, “A characterization of the sphere in terms of single-layer potentials”, *Proc. Amer. Math. Soc.* **115**:4 (1992), 1167–1168. MR 93c:31011 Zbl 0752.31003



- [Shapiro 1966] H. S. Shapiro, “Remarks concerning domains of Smirnov type”, *Michigan Math. J.* **13** (1966), 341–348. MR 34 #4472 Zbl 0179.11501
- [Shapiro 1992] H. S. Shapiro, *The Schwarz function and its generalization to higher dimensions*, University of Arkansas Lecture Notes in the Mathematical Sciences **9**, Wiley, New York, 1992. MR 93g:30059 Zbl 0784.30036
- [Topping 2006] P. Topping, *Lectures on the Ricci flow*, London Mathematical Society Lecture Note Series **325**, Cambridge University Press, 2006. MR 2007h:53105 Zbl 1105.58013
- [Traizet 2013] M. Traizet, “Classification of the solutions to an overdetermined problem in the plane”, preprint, 2013. arXiv 1301.6927
- [Tumarkin and Havinson 1958a] G. C. Tumarkin and S. J. Havinson, “К определению аналитических функций класса  $E_p$  в многосвязных областях”, *Uspekhi Mat. Nauk (N.S.)* **13**:1 (1958), 201–206. MR 20 #114 Zbl 0087.07902
- [Tumarkin and Havinson 1958b] G. C. Tumarkin and S. J. Havinson, “О теореме разложения для аналитических функций класса  $E_p$  в многосвязных областях”, *Uspekhi Mat. Nauk (N.S.)* **13**:2(80) (1958), 223–228. MR 20 #3285 Zbl 0081.29804
- [Tumarkin and Havinson 1960] G. C. Tumarkin and S. J. Havinson, “Classes of analytic functions on multiply connected domains”, pp. 45–77 in Исследования по современным проблемам теории функций комплексного переменного, Gosudarstv. Izdat. Fiz.-Mat. Lit., Moscow, 1960. In Russian; translated as pp. 37–71 in *Fonctions d’une variable complexe: problèmes contemporains*, edited by A. I. Markouchevitch, Gauthier-Villars, Paris, 1962. MR 22 #9616 Zbl 0116.28604

Received May 22, 2012. Revised May 14, 2013.

DMITRY KHAVINSON  
DEPARTMENT OF MATHEMATICS AND STATISTICS  
UNIVERSITY OF SOUTH FLORIDA  
TAMPA, FL 33620  
UNITED STATES  
dkhavins@usf.edu

ERIK LUNDBERG  
DEPARTMENT OF MATHEMATICS  
PURDUE UNIVERSITY  
WEST LAFAYETTE, IN 47907  
UNITED STATES  
elundber@math.purdue.edu

RAZVAN TEODORESCU  
DEPARTMENT OF MATHEMATICS AND STATISTICS  
UNIVERSITY OF SOUTH FLORIDA  
TAMPA, FL 33647  
UNITED STATES  
razvan@usf.edu



## QUASISYMMETRIC HOMEOMORPHISMS ON REDUCIBLE CARNOT GROUPS

XIANGDONG XIE

**We show that quasisymmetric homeomorphisms between (most) reducible Carnot groups are bilipschitz. This implies rigidity for quasi-isometries between certain negatively curved homogeneous manifolds. The proof uses Pansu's differentiability theorem for quasisymmetric homeomorphisms between Carnot groups.**

### 1. Introduction

We study quasisymmetric homeomorphisms between reducible Carnot groups. The main result says that in most cases, the quasisymmetric homeomorphism must be bilipschitz.

A Carnot group is *reducible* if it is isomorphic to the direct product of two Carnot groups. Otherwise, a Carnot group is called *irreducible*. A reducible Carnot group  $G$  can be written as  $G = G_0 \times G_1 \times \cdots \times G_m$ , where  $G_0$  is abelian (i.e., isomorphic to some  $\mathbb{R}^n$ ), and  $G_j$  ( $1 \leq j \leq m$ ) is nonabelian irreducible. Such a decomposition is not unique in general; see Example 2.1.

All Carnot groups in this paper are equipped with the Carnot metric (see Section 3).

**Theorem 1.1.** *Let  $F : G \rightarrow G'$  be a quasisymmetric map between two Carnot groups. Suppose  $G$  is reducible and admits a direct product decomposition of irreducible Carnot groups where at least two of the factors are not isomorphic. Then  $F$  is bilipschitz.*

The same claim remains open in the case when  $G$  is isomorphic to a direct product  $N \times \cdots \times N$ , where  $N$  is nonabelian irreducible.

Quasisymmetric homeomorphisms between general metric spaces are quasiconformal. In the case of Carnot groups (and of Loewner spaces more generally), a map is quasisymmetric if and only if it is quasiconformal (see [Heinonen and Koskela 1998]).

---

Partially supported by NSF grant DMS-1265735.

MSC2010: 53C17, 53C23, 30L10.

Keywords: quasisymmetric homeomorphism, Carnot groups, Pansu differentiability.

Theorem 1.1 has consequences for the rigidity of quasi-isometries between certain negatively curved homogeneous manifolds. Recall that a quasi-isometry between two metric spaces is an almost isometry if it preserves distance up to an additive constant. A quasi-isometry between two negatively curved spaces induces a quasisymmetric homeomorphism between the ideal boundaries (of the negatively curved spaces), where the ideal boundaries are equipped with visual metrics. Conversely, under mild conditions on the negatively curved spaces, each quasisymmetric homeomorphism between the ideal boundaries is the boundary map of a quasi-isometry; see [Bonk and Schramm 2000]. Similarly, almost isometries between negatively curved spaces correspond to bilipschitz maps between the ideal boundaries [ibid.]. On the other hand, Carnot groups arise as the ideal boundary of certain negatively curved homogeneous manifolds (see below for more details). Hence a direct consequence of Theorem 1.1 is that each quasi-isometry between certain negatively curved homogeneous manifolds is an almost isometry.

Heintze [1974] characterized homogeneous manifolds with negative sectional curvature (HMNs): Each HMN is isometric to a simply connected solvable Lie group  $S$  equipped with a left invariant Riemannian metric, and furthermore  $S = N \rtimes \mathbb{R}$  is a semidirect product of a nilpotent Lie group  $N$  with  $\mathbb{R}$ , where  $\mathbb{R}$  acts on  $N$  by expanding (and contracting) automorphisms; conversely, every semidirect product as above admits a left invariant Riemannian metric with negative sectional curvature (hence is an HMN). The ideal boundary of an HMN  $S = N \rtimes \mathbb{R}$  can be naturally identified with (the one-point compactification of)  $N$ . On the other hand, each Carnot group  $N$  is a simply connected nilpotent Lie group having a one-parameter family of dilations (see Section 3 for more details). These dilations induce an action of  $\mathbb{R}$  on the Carnot group by expanding (and contracting) automorphisms, so there is an HMN  $N \rtimes \mathbb{R}$  associated with each Carnot group  $N$ . It follows that each Carnot group can be identified with the ideal boundary of some HMN. Hence Theorem 1.1 implies that each quasi-isometry between these HMNs is an almost isometry.

We next make some comments about the proof of Theorem 1.1. A main step in the proof is to show that the quasisymmetric homeomorphism preserves a certain foliation. Then the arguments in [Shanmugalingam and Xie 2012] show that the quasisymmetric homeomorphism is bilipschitz. To show that the quasisymmetric homeomorphism preserves a foliation, one first proves that infinitesimally it preserves a foliation. The global result then follows by integration. Recall that Pansu's differentiability theorem (see [Pansu 1989] or Section 3) says that a quasisymmetric homeomorphism  $F : G \rightarrow G'$  between Carnot groups is Pansu-differentiable a.e., and the Pansu differential is a.e. a graded isomorphism between the two Carnot groups. Under the assumption of Theorem 1.1, we show that there are (proper) connected and simply connected subgroups  $N \subset G$  and  $N' \subset G'$  such that  $\phi(N) = N'$  for every graded isomorphism  $\phi : G \rightarrow G'$ ; see Section 2.

In Section 2 we show that graded isomorphisms between reducible Carnot algebras preserve certain subalgebras, which implies that graded isomorphisms between reducible Carnot groups preserve certain subgroups (as indicated in the preceding paragraph). And in Section 3 we show that quasisymmetric homeomorphisms are bilipschitz.

## 2. Graded isomorphisms of Carnot algebras

In this section we show that graded isomorphisms between reducible Carnot algebras preserve certain subalgebras. This implies that graded isomorphisms between reducible Carnot groups preserve certain subgroups (see Section 3).

A *Carnot Lie algebra* is a finite-dimensional Lie algebra  $\mathcal{G}$  together with a direct sum decomposition  $\mathcal{G} = V_1 \oplus V_2 \oplus \cdots \oplus V_r$  of vector subspaces such that  $[V_1, V_i] = V_{i+1}$  for all  $1 \leq i \leq r$ , where we set  $V_{r+1} = \{0\}$ . The integer  $r$  is called the degree of nilpotency of  $\mathcal{G}$ . Every Carnot algebra  $\mathcal{G} = V_1 \oplus V_2 \oplus \cdots \oplus V_r$  admits a one-parameter family of automorphisms  $\lambda_t : \mathcal{G} \rightarrow \mathcal{G}$  for  $t \in (0, \infty)$ , where  $\lambda_t(x) = t^i x$  for  $x \in V_i$ . Let  $\mathcal{G} = V_1 \oplus V_2 \oplus \cdots \oplus V_r$  and  $\mathcal{G}' = V'_1 \oplus V'_2 \oplus \cdots \oplus V'_s$  be two Carnot algebras. A Lie algebra homomorphism  $\phi : \mathcal{G} \rightarrow \mathcal{G}'$  is graded if  $\phi$  commutes with  $\lambda_t$  for all  $t > 0$ ; that is, if  $\phi \circ \lambda_t = \lambda_t \circ \phi$ . We observe that  $\phi(V_i) \subset V'_i$  for all  $1 \leq i \leq r$ .

A Carnot algebra  $\mathcal{G}$  is called reducible if there exist two nontrivial Carnot algebras  $\mathcal{G}_1$  and  $\mathcal{G}_2$  and a graded isomorphism between  $\mathcal{G}$  and  $\mathcal{G}_1 \oplus \mathcal{G}_2$ . It is called irreducible otherwise. The finite dimensionality implies that every reducible Carnot algebra  $\mathcal{G}$  can be written as a direct sum of Carnot algebras  $\mathcal{G} = \mathcal{G}_0 \oplus \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$ , where  $\mathcal{G}_0$  is abelian, and each  $\mathcal{G}_i$  with  $i \geq 1$  is nonabelian and irreducible.

Let  $\mathcal{G} = V_1 \oplus V_2 \oplus \cdots \oplus V_r$  be a Carnot algebra. When  $\mathcal{G}$  is reducible, it also has a decomposition  $\mathcal{G} = \mathcal{G}_0 \oplus \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$  as a direct sum of an abelian factor and irreducible nonabelian factors. We are interested in the question whether graded isomorphisms preserve such decompositions (after possibly permuting the factors). This question is equivalent to the uniqueness problem of such a decomposition. In general the decomposition is not unique, as the following example shows. The author thanks Bruce Kleiner for suggesting the example.

**Example 2.1** (Kleiner). Let  $\mathcal{G}$  be a nonabelian irreducible Carnot algebra, and let  $f : \mathcal{G} \rightarrow \mathbb{R}^m$  be a nontrivial Lie algebra homomorphism into an abelian group. Let  $G(f) \subset \mathcal{G} \oplus \mathbb{R}^m$  be the graph of  $f$ . Then  $G(f)$  is a Carnot algebra (being isomorphic to  $\mathcal{G}$ ), and  $\mathcal{G} \oplus \mathbb{R}^m$  has two different decompositions  $\mathcal{G} \oplus \mathbb{R}^m = G(f) \oplus \mathbb{R}^m$ . Alternatively, let  $g : \mathcal{G} \oplus \mathbb{R}^m \rightarrow \mathcal{G} \oplus \mathbb{R}^m$  be the map given by  $g(x, a) = (x, a + f(x))$ . Then  $g$  is a graded isomorphism and it does not preserve the factor  $\mathcal{G}$ .

Despite this example, we show that graded isomorphisms always preserve the abelian factor (Proposition 2.4) and, in the case of a trivial abelian factor, preserve

the decomposition after possibly permuting the factors (Proposition 2.5).

**Definition 2-1.** Let  $\mathcal{G}$  be a Lie algebra and  $x \in \mathcal{G}$ . Define  $d(x) = \dim(\ker(\text{ad } x))$ , where  $\text{ad } x : \mathcal{G} \rightarrow \mathcal{G}$  is the linear map given by  $\text{ad } x(y) = [x, y]$ .

**Lemma 2.2.** *If  $d(x) = \dim \mathcal{G} > 1$  for some  $x \in V_1 \setminus \{0\}$ , then  $\mathcal{G}$  is reducible.*

*Proof.* Note  $[x, y] = 0$  for all  $y \in \mathcal{G}$ . Let  $\mathcal{G}_1$  be the one-dimensional subspace of  $V_1$  spanned by  $x$ , and let  $W$  be a complementary subspace of  $\mathcal{G}_1$  in  $V_1$ . Set  $\mathcal{G}_2 = W \oplus V_2 \oplus \cdots \oplus V_r$ . Then  $\mathcal{G} = \mathcal{G}_1 \oplus \mathcal{G}_2$  is a direct sum of vector subspaces. The assumption on  $x$  now implies that  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are ideals of  $\mathcal{G}$  and that both  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are Carnot algebras. Hence  $\mathcal{G}$  is reducible.  $\square$

The next lemma provides an intrinsic characterization of the abelian factor  $\mathcal{G}_0$ .

**Lemma 2.3.** *Let  $\mathcal{G} = V_1 \oplus V_2 \oplus \cdots \oplus V_r$  be a Carnot algebra, and consider a direct sum decomposition  $\mathcal{G} = \mathcal{G}_0 \oplus \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$  of  $\mathcal{G}$  into an abelian factor and irreducible nonabelian factors. Let  $x \in V_1$ . Then  $x \in \mathcal{G}_0$  if and only if  $d(x) = \dim \mathcal{G}$ .*

*Proof.* It is clear that  $x \in \mathcal{G}_0$  implies  $d(x) = \dim \mathcal{G}$ . We assume  $d(x) = \dim \mathcal{G}$  and shall prove that  $x \in \mathcal{G}_0$ . Note  $[x, y] = 0$  for all  $y \in \mathcal{G}$ . Write  $x = x_0 + x_1 + \cdots + x_m$  with  $x_i \in \mathcal{G}_i \cap V_1$ . Suppose  $x \notin \mathcal{G}_0$ . Then  $x_i \neq 0$  for some  $i \geq 1$ . Since  $[x_i, y] = [x, y] = 0$  for all  $y \in \mathcal{G}_i$ , Lemma 2.2 implies  $\mathcal{G}_i$  is reducible, contradicting the assumption.  $\square$

Recall that the goal of this section is to show that a graded isomorphism of reducible Carnot algebras preserves certain Lie subalgebras. The case when the abelian factor is nontrivial is covered by Proposition 2.4.

**Proposition 2.4.** *Let  $\mathcal{G} = \mathcal{G}_0 \oplus \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$  and  $\mathcal{G}' = \mathcal{G}'_0 \oplus \mathcal{G}'_1 \oplus \cdots \oplus \mathcal{G}'_n$  be two reducible Carnot algebras written as direct sums of an abelian factor and irreducible nonabelian factors. Let  $\phi : \mathcal{G} \rightarrow \mathcal{G}'$  be a graded isomorphism. Then  $\phi(\mathcal{G}_0) = \mathcal{G}'_0$ .*

*Proof.* By Lemma 2.3,  $\phi(\mathcal{G}_0) \subset \mathcal{G}'_0$ . Since  $\phi$  is an isomorphism, the conclusion follows by considering  $\phi^{-1}$ .  $\square$

Proposition 2.5 treats the case when the abelian factor is trivial.

**Proposition 2.5.** *Let  $\mathcal{G} = \mathcal{G}_0 \oplus \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$  and  $\mathcal{G}' = \mathcal{G}'_0 \oplus \mathcal{G}'_1 \oplus \cdots \oplus \mathcal{G}'_n$  be two reducible Carnot algebras written as direct sums of an abelian factor and irreducible nonabelian factors. Let  $\phi : \mathcal{G} \rightarrow \mathcal{G}'$  be a graded isomorphism. Suppose  $\mathcal{G}$  has no abelian factor (that is,  $\mathcal{G}_0 = \{0\}$ ). Then  $\mathcal{G}'_0 = \{0\}$ ,  $m = n$  and after possibly permuting the factors  $\mathcal{G}'_1, \dots, \mathcal{G}'_m$ , there exist graded isomorphisms  $\phi_i : \mathcal{G}_i \rightarrow \mathcal{G}'_i$  such that  $\phi = \phi_1 \oplus \cdots \oplus \phi_m$ .*

Now we start the proof of Proposition 2.5. First observe that Proposition 2.4 implies  $\mathcal{G}'_0 = \{0\}$ . In the following proofs, we shall use both decompositions  $\mathcal{G} = V_1 \oplus \cdots \oplus V_r = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$  of  $\mathcal{G}$ , as well as those for  $\mathcal{G}'$ .

**Lemma 2.6.** *Let  $x \in V_1$ . Write  $x = x_{i_1} + \cdots + x_{i_k}$  ( $1 \leq i_1 < \cdots < i_k \leq m$ ) with  $x_{i_j} \in (\mathcal{G}_{i_j} \cap V_1) \setminus \{0\}$ . If  $k \geq 2$ , then  $d(x_{i_j}) > d(x)$ .*

*Proof.* We first show that  $\ker(\operatorname{ad} x) \subset \ker(\operatorname{ad} x_{i_j})$  for all  $1 \leq j \leq k$ . Let  $y \in \ker(\operatorname{ad} x)$ . Write  $y = y_1 + \cdots + y_m$  with  $y_i \in \mathcal{G}_i$ . Then  $0 = [x, y] = [x_{i_1}, y_{i_1}] + \cdots + [x_{i_k}, y_{i_k}]$ . Since  $[x_{i_j}, y_{i_j}] \in \mathcal{G}_{i_j}$ , we have  $[x_{i_j}, y_{i_j}] = 0$ . Hence  $[x_{i_j}, y] = [x_{i_j}, y_{i_j}] = 0$ ; that is,  $y \in \ker(\operatorname{ad} x_{i_j})$ .

Next we shall find an element  $y \in \ker(\operatorname{ad} x_{i_j}) \setminus \ker(\operatorname{ad} x)$ . Since  $k \geq 2$ , there is some  $1 \leq l \leq k$  with  $l \neq j$ . By Lemma 2.2, since  $\mathcal{G}_{i_l}$  is nonabelian and irreducible, there is some  $y \in \mathcal{G}_{i_l}$  such that  $[x_{i_l}, y] \neq 0$ . Now notice that  $[x_{i_j}, y] = 0$  and  $[x, y] = [x_{i_l}, y] \neq 0$ . □

For each  $1 \leq i \leq m$ , set

$$A_i = \{x \in \mathcal{G}_i \cap V_1 : \phi(x) \in \mathcal{G}'_j \text{ for some } j\}.$$

Let  $N_i \subset \mathcal{G}_i \cap V_1$  be the vector subspace spanned by  $A_i$ . Similarly, for each  $1 \leq j \leq n$  set

$$A'_j = \{y \in \mathcal{G}'_j \cap V'_1 : \phi^{-1}(y) \in \mathcal{G}_i \text{ for some } i\}.$$

Let  $N'_j \subset \mathcal{G}'_j \cap V'_1$  be the vector subspace spanned by  $A'_j$ .

**Lemma 2.7.** *We have  $N_i = \mathcal{G}_i \cap V_1$  for each  $i$  and  $N'_j = \mathcal{G}'_j \cap V'_1$  for each  $j$ .*

*Proof.* We prove by contradiction. Suppose  $N_i \neq \mathcal{G}_i \cap V_1$  for some  $i$  or  $N'_j \neq \mathcal{G}'_j \cap V'_1$  for some  $j$ . Let  $d_1 = 0$  if  $N_i = \mathcal{G}_i \cap V_1$  for all  $i$ ; otherwise, let

$$d_1 = \max\{d(x) : x \in (\mathcal{G}_i \cap V_1) \setminus N_i \text{ for some } i\}.$$

Similarly, let  $d_2 = 0$  if  $N'_j = \mathcal{G}'_j \cap V'_1$  for all  $j$ ; otherwise, let

$$d_2 = \max\{d(y) : y \in (\mathcal{G}'_j \cap V'_1) \setminus N'_j \text{ for some } j\}.$$

Let  $d_0 = \max\{d_1, d_2\}$ . After possibly replacing  $\phi$  with  $\phi^{-1}$ , we may assume  $d_0 = d_1$ . Pick  $x \in (\mathcal{G}_i \cap V_1) \setminus N_i$  (for some  $i$ ) with  $d(x) = d_0$ . By the definition of  $N_i$  we have  $x \notin A_i$ . Hence  $\phi(x)$  can be written as

$$(2-2) \quad \phi(x) = y_1 + \cdots + y_k,$$

where  $k \geq 2$  and  $y_s \in (\mathcal{G}'_{j_s} \cap V'_1) \setminus \{0\}$  for each  $1 \leq s \leq k$ , and  $1 \leq j_1 < \cdots < j_k \leq n$ . By Lemma 2.6,  $d(y_s) > d(\phi(x)) = d(x) = d_0$ . It follows from the definition of  $d_0$  that  $y_s \in N'_{j_s}$ . Hence there is an expression

$$(2-3) \quad y_s = z_{s,1} + \cdots + z_{s,u_s} + w_{s,1} + \cdots + w_{s,v_s}$$

with  $z_{s,p}, w_{s,q} \in A'_{j_s}$  such that  $\phi^{-1}(z_{s,p}) \in \mathcal{G}_i \cap V_1$  and  $\phi^{-1}(w_{s,q}) \in \mathcal{G}_t \cap V_1$  for some  $t \neq i$  (here  $t$  may depend on  $q$ ). Notice that (2-2) and (2-3) imply

$$x = \sum_{s,p} \phi^{-1}(z_{s,p}) + \sum_{s,q} \phi^{-1}(w_{s,q}).$$

Since  $x \in \mathcal{G}_i$  and  $\mathcal{G} = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$  is a direct sum decomposition, we obtain  $x = \sum_{s,p} \phi^{-1}(z_{s,p})$ . Notice that each  $\phi^{-1}(z_{s,p}) \in A_i$ . It follows that  $x \in N_i$ , contradicting the assumption.  $\square$

**Lemma 2.8.** *For each  $i$ , there is some  $j$  such that  $\phi(\mathcal{G}_i \cap V_1) \subset \mathcal{G}'_j \cap V'_1$ .*

*Proof.* Fix  $i$ . By Lemma 2.7,  $\mathcal{G}_i \cap V_1 = N_i$ . Hence there is a vector space basis  $B$  of  $\mathcal{G}_i \cap V_1$  consisting of elements of  $A_i$ . Write  $B$  as a disjoint union  $B = \sqcup B_j$ , where  $B_j$  consists of those elements of  $B$  that are mapped into  $\mathcal{G}'_{j_1}$  under  $\phi$ . Since  $\phi$  is an isomorphism and  $\mathcal{G}'_{j_1}$  and  $\mathcal{G}'_{j_2}$  commute for  $j_1 \neq j_2$ , we see that  $[X, Y] = 0$  for  $X \in B_{j_1}$  and  $Y \in B_{j_2}$ . Let  $E_j \subset \mathcal{G}_i$  be the subalgebra of  $\mathcal{G}_i$  generated by  $B_j$ . Observe that  $E_j$  is an ideal of  $\mathcal{G}_i$  and  $\mathcal{G}_i$  admits the direct sum decomposition  $\mathcal{G}_i = E_1 \oplus \cdots \oplus E_n$ . Since  $\mathcal{G}_i$  is irreducible,  $E_j = \{0\}$  for all  $j$  except exactly one. It follows that for some  $j$ , all the elements in  $B$  are mapped into  $\mathcal{G}'_j$ . Since  $B$  is a basis of  $\mathcal{G}_i \cap V_1$ , we have  $\phi(\mathcal{G}_i \cap V_1) \subset \mathcal{G}'_j$ .  $\square$

Applying Lemma 2.8 to  $\phi^{-1}$ , we see that for each  $j$ , there is some  $i$  such that  $\phi^{-1}(\mathcal{G}'_j \cap V'_1) \subset \mathcal{G}_i \cap V_1$ . From this it is easy to see that  $m = n$ , and after possibly permuting the factors  $\mathcal{G}'_j$  we have  $\phi(\mathcal{G}_i) = \mathcal{G}'_i$ . Proposition 2.5 follows.

### 3. Quasisymmetric homeomorphisms are bilipschitz

In this section we show that in most cases quasisymmetric homeomorphisms between reducible Carnot groups are bilipschitz.

A simply connected nilpotent Lie group is a *Carnot group* if its Lie algebra is a Carnot algebra. Let  $G$  be a Carnot group with Lie algebra  $\mathcal{G} = V_1 \oplus \cdots \oplus V_r$ . The subspace  $V_1$  defines a left invariant distribution  $HG \subset TG$  on  $G$ . We fix a left invariant inner product on  $HG$ . An absolutely continuous curve  $\gamma$  in  $G$  whose velocity vector  $\gamma'(t)$  is contained in  $H_{\gamma(t)}G$  for a.e.  $t$  is called a horizontal curve. By Chow's theorem ([Bellaïche and Risler 1996], Theorem 2.4), any two points of  $G$  can be connected by horizontal curves. Let  $p, q \in G$ ; the *Carnot distance*  $d(p, q)$  between them is defined as the infimum of length of horizontal curves that join  $p$  and  $q$ .

Since the inner product on  $HG$  is left invariant, the Carnot metric on  $G$  is also left invariant. Different choices of inner product on  $HG$  result in Carnot metrics that are bilipschitz equivalent. The Hausdorff dimension of  $G$  with respect to a Carnot metric is given by  $\sum_{i=1}^r i \cdot \dim V_i$ . We use the corresponding Hausdorff measure on  $G$ . When  $G = G_1 \times G_2$  is a direct product of two Carnot groups (with



a suitable choice of inner product on  $HG$ ), the Carnot metric on  $G$  is the product of the Carnot metrics on  $G_1$  and  $G_2$ , and the Hausdorff measure on  $G$  is the product of the Hausdorff measures on  $G_1$  and  $G_2$ .

Recall that, for a simply connected nilpotent Lie group  $G$  with Lie algebra  $\mathcal{G}$ , the exponential map  $\exp : \mathcal{G} \rightarrow G$  is a diffeomorphism. Furthermore, the exponential map induces a one-to-one correspondence between Lie subalgebras of  $\mathcal{G}$  and connected Lie subgroups of  $G$ .

Let  $G$  be a Carnot group with Lie algebra  $\mathcal{G} = V_1 \oplus \dots \oplus V_r$ . Since  $\lambda_t : \mathcal{G} \rightarrow \mathcal{G}$  ( $t > 0$ ) is a Lie algebra automorphism and  $G$  is simply connected, there is a unique Lie group automorphism  $\Lambda_t : G \rightarrow G$  whose differential at the identity is  $\lambda_t$ . For each  $t > 0$ ,  $\Lambda_t$  is a similarity with respect to the Carnot metric:  $d(\Lambda_t(p), \Lambda_t(q)) = t d(p, q)$  for any two points  $p, q \in G$ . A Lie group homomorphism  $f : G \rightarrow G'$  between two Carnot groups is a graded homomorphism if it commutes with  $\Lambda_t$  for all  $t > 0$ ; that is, if  $f \circ \Lambda_t = \Lambda_t \circ f$ . Notice that a Lie group homomorphism  $f : G \rightarrow G'$  between two Carnot groups is graded if and only if the corresponding Lie algebra homomorphism is graded.

A Carnot group is reducible if its Lie algebra is reducible. Equivalently, a Carnot group is reducible if it is isomorphic to the direct product of two Carnot groups. A Carnot group is called irreducible otherwise.

Proposition 2.4 and Proposition 2.5 respectively immediately imply Corollary 3.1 and Corollary 3.2, which say that a graded isomorphism of reducible Carnot groups preserves certain Lie subgroups.

**Corollary 3.1.** *Let  $G = G_0 \times G_1 \times \dots \times G_m$  and  $G' = G'_0 \times G'_1 \times \dots \times G'_n$  be two reducible Carnot groups written as direct products of an abelian factor and irreducible nonabelian factors. Let  $f : G \rightarrow G'$  be a graded isomorphism. Then  $f(G_0) = G'_0$ .*

**Corollary 3.2.** *Let  $G = G_0 \times G_1 \times \dots \times G_m$  and  $G' = G'_0 \times G'_1 \times \dots \times G'_n$  be two reducible Carnot groups written as direct products of an abelian factor and irreducible nonabelian factors. Let  $f : G \rightarrow G'$  be a graded isomorphism. Suppose  $G$  has no abelian factor (that is,  $G_0 = \{e\}$ ). Then  $G'_0 = \{e\}$ ,  $m = n$  and after possibly permuting the factors  $G'_1, \dots, G'_m$ , there exist graded isomorphisms  $f_i : G_i \rightarrow G'_i$  such that  $f = f_1 \times \dots \times f_m$ .*

**Definition 3-1.** Let  $G$  and  $G'$  be two Carnot groups endowed with Carnot metrics. A map  $F : G \rightarrow G'$  is *Pansu-differentiable* at  $x \in G$  if there exists a graded homomorphism  $L : G \rightarrow G'$  such that

$$\lim_{y \rightarrow x} \frac{d(F(x)^{-1}F(y), L(x^{-1}y))}{d(x, y)} = 0.$$

In this case, the graded homomorphism  $L : G \rightarrow G'$  is called the *Pansu differential* of  $F$  at  $x$ , and is denoted by  $dF(x)$ .

**Definition 3-2.** Let  $\eta : [0, \infty) \rightarrow [0, \infty)$  be a homeomorphism. A homeomorphism of metric spaces  $F : X \rightarrow Y$  is an  $\eta$ -*quasisymmetric* homeomorphism if for all distinct triples  $x, y, z \in X$ , we have

$$\frac{d(F(x), F(y))}{d(F(x), F(z))} \leq \eta\left(\frac{d(x, y)}{d(x, z)}\right).$$

A map  $F : X \rightarrow Y$  is a *quasisymmetric homeomorphism* if it is an  $\eta$ -quasisymmetric homeomorphism for some  $\eta$ .

The following result (except the terminology) is due to Pansu [1989].

**Theorem 3.3.** *Let  $F : G \rightarrow G'$  be a quasisymmetric homeomorphism between two Carnot groups. Then  $F$  is a.e. Pansu-differentiable. Furthermore, at a.e.  $x \in G$ , the Pansu differential  $dF(x) : G \rightarrow G'$  is a graded isomorphism.*

In Theorem 3.3 and the proofs below, “a.e.” is with respect to the Hausdorff measure on  $G$ .

For the proof of Theorem 1.1, we need the following:

**Proposition 3.4.** *Let  $G$  and  $G'$  be two Carnot groups,  $W \subset V_1$ ,  $W' \subset V'_1$  be subspaces. Denote by  $\mathcal{G}_W \subset \mathcal{G}$  and  $\mathcal{G}'_{W'} \subset \mathcal{G}'$ , respectively, the Lie subalgebras generated by  $W$  and  $W'$ . Let  $H \subset G$  and  $H' \subset G'$ , respectively, be the connected Lie subgroups of  $G$  and  $G'$  corresponding to  $\mathcal{G}_W$  and  $\mathcal{G}'_{W'}$ . Let  $F : G \rightarrow G'$  be a quasisymmetric homeomorphism. If  $dF(x)(W) \subset W'$  for a.e.  $x \in G$ , then  $F$  sends left cosets of  $H$  into left cosets of  $H'$ .*

*Proof.* For each nonzero vector  $u \in W$ , the set  $\{\exp(tu) : t \in \mathbb{R}\}$  is a subgroup of  $G$ . It is a geodesic with respect to the Carnot metric and shall be called a horizontal line. For each nonzero vector  $u \in W$ , let  $\mathcal{F}_u$  be the set of left cosets of  $\{\exp(tu) : t \in \mathbb{R}\}$  in  $G$ . By the main result in [Balogh et al. 2007],  $F : G \rightarrow G'$  is absolutely continuous on almost every curve. It follows that for almost every  $L \in \mathcal{F}_u$ , the map  $F|_L : L \rightarrow G'$  is an absolutely continuous curve in  $G'$ . On the other hand, by Pansu’s theorem,  $F$  is a.e. Pansu-differentiable and the Pansu differential  $dF(x) : G \rightarrow G'$  is a graded isomorphism for a.e.  $x \in G$ . Also by assumption,  $dF(x)(W) \subset W'$  for a.e.  $x \in G$ . It follows from Fubini’s theorem that, for almost every  $L \in \mathcal{F}_u$ , the Pansu differential  $dF(x) : G \rightarrow G'$  exists, is a graded isomorphism and satisfies  $dF(x)(W) \subset W'$  for a.e.  $x \in L$ . Hence, the tangent vectors of the curve  $F|_L$  lie in  $W'$  almost everywhere. It follows that for almost every  $L \in \mathcal{F}_u$ ,  $F(L)$  lies in a left coset of  $H'$ . Now the continuity of  $F$  and a limiting argument show that the same is true for all  $L \in \mathcal{F}_u$ . Conceivably, it might be possible for distinct  $L_1, L_2 \in \mathcal{F}_u$  to lie in the same left

coset of  $H$ , while their images  $F(L_1)$  and  $F(L_2)$  lie in distinct cosets of  $H'$ . We next show that this cannot happen.

In a Carnot group, every two points can be joined by a piecewise geodesic, where each piece is a left translation of a segment in a horizontal line. The preceding paragraph shows that the image under  $F$  of each piece lies in a left coset of  $H'$ . It follows that the image of the entire piecewise geodesic lies in a left coset of  $H'$ . Hence  $F$  sends left cosets of  $H$  into left cosets of  $H'$ .  $\square$

Now we are ready to prove Theorem 1.1.

*Proof of Theorem 1.1.* Let  $F : G \rightarrow G'$  be a quasisymmetric homeomorphism between two Carnot groups. Suppose  $G$  is reducible and admits a direct product decomposition of irreducible Carnot groups where at least two of the factors are not isomorphic. We first use Proposition 3.4 to show that  $F$  preserves a certain foliation. The arguments in [Shanmugalingam and Xie 2012] then show that  $F$  is bilipschitz.

Write  $G = G_0 \times G_1 \times \cdots \times G_m$  and  $G' = G'_0 \times G'_1 \times \cdots \times G'_n$ , where  $G_0, G'_0$  are abelian and  $G_i, G'_j$  are irreducible nonabelian factors.

First consider the case when  $G_0$  is nontrivial. Let  $\mathcal{F}$  be the foliation of  $G$  consisting of the cosets of  $G_0$ , and similarly let  $\mathcal{F}'$  be the foliation of  $G'$  consisting of the cosets of  $G'_0$ . The leaf space of  $\mathcal{F}$  can be naturally identified with  $N := G_1 \times \cdots \times G_m$ , and that of  $\mathcal{F}'$  with  $N' := G'_1 \times \cdots \times G'_n$ . By Corollary 3.1 and Proposition 3.4, the map  $F$  sends the leafs of  $\mathcal{F}$  to the leafs of  $\mathcal{F}'$ . Hence  $F$  induces a map  $F_1 : N \rightarrow N'$ . Notice that  $G = G_0 \times N$  with the Carnot metric is isometric to the product of  $G_0$  and  $N$  (also equipped with the Carnot metric); similarly for  $G'$ . The arguments in [Shanmugalingam and Xie 2012] go through and imply that  $F_1$  is also quasisymmetric. Since both the leafs and the leaf spaces are geodesic metric spaces, the arguments further show that  $F$  is bilipschitz.

Next we consider the case when  $G_0$  is trivial. Then  $G = G_1 \times \cdots \times G_m$  is a direct product of nonabelian irreducible Carnot groups. We combine all isomorphic factors in the above decomposition to obtain  $G = N_1 \times \cdots \times N_s$ . Each  $N_j$  is a direct product of isomorphic nonabelian irreducible Carnot groups, and the factors in  $N_i$  and  $N_j$  are not isomorphic for  $i \neq j$ . Similarly,  $G'$  can also be written as such a product  $G' = N'_1 \times \cdots \times N'_t$ . Notice that the assumption of Theorem 1.1 implies that  $s \geq 2$ . Corollary 3.2 implies that  $s = t$ , and after possibly permuting the factors  $N'_i$ , the Pansu differential satisfies  $dF(x)(N_i) = N'_i$  for all  $i$  and a.e.  $x \in N$ . Now the arguments in the preceding paragraph show that  $F$  is bilipschitz. The proof of Theorem 1.1 is now complete.  $\square$

### Acknowledgments

I thank Bruce Kleiner for helpful discussions, and the referee for useful remarks.

## References

- [Balogh et al. 2007] Z. M. Balogh, P. Koskela, and S. Rogovin, “Absolute continuity of quasi-conformal mappings on curves”, *Geom. Funct. Anal.* **17**:3 (2007), 645–664. MR 2009g:30023 Zbl 1134.30014
- [Bellaïche and Risler 1996] A. Bellaïche and J.-J. Risler (editors), *Sub-Riemannian geometry* (Paris, 1992), Progress in Mathematics **144**, Birkhäuser, Basel, 1996. MR 97f:53002 Zbl 0848.00020
- [Bonk and Schramm 2000] M. Bonk and O. Schramm, “Embeddings of Gromov hyperbolic spaces”, *Geom. Funct. Anal.* **10**:2 (2000), 266–306. MR 2001g:53077 Zbl 0972.53021
- [Heinonen and Koskela 1998] J. Heinonen and P. Koskela, “Quasiconformal maps in metric spaces with controlled geometry”, *Acta Math.* **181**:1 (1998), 1–61. MR 99j:30025 Zbl 0915.30018
- [Heintze 1974] E. Heintze, “On homogeneous manifolds of negative curvature”, *Math. Ann.* **211** (1974), 23–34. MR 50 #5695 Zbl 0273.53042
- [Pansu 1989] P. Pansu, “Métriques de Carnot–Carathéodory et quasiisométries des espaces symétriques de rang un”, *Ann. of Math. (2)* **129**:1 (1989), 1–60. MR 90e:53058 Zbl 0678.53042
- [Shanmugalingam and Xie 2012] N. Shanmugalingam and X. Xie, “A rigidity property of some negatively curved solvable Lie groups”, *Comment. Math. Helv.* **87**:4 (2012), 805–823. MR 2984572 Zbl 1255.22005

Received May 17, 2012. Revised May 5, 2013.

XIANGDONG XIE  
DEPARTMENT OF MATHEMATICS AND STATISTICS  
BOWLING GREEN STATE UNIVERSITY  
BOWLING GREEN, OH 43403  
UNITED STATES  
xiex@bgsu.edu

## CAPILLARITY AND ARCHIMEDES' PRINCIPLE

JOHN MCCUAN AND RAY TREINEN

**We consider some of the complications that arise in attempting to generalize a version of Archimedes' principle concerning floating bodies to account for capillary effects. The main result provides a means to relate the floating position (depth in the liquid) of a symmetrically floating sphere in terms of other observable geometric quantities.**

**A similar result is obtained for an idealized case corresponding to a symmetrically floating infinite cylinder.**

**These results depend on a definition of equilibrium for capillary systems with floating objects which to our knowledge has not formally appeared in the literature. The definition, in turn, depends on a variational formula for floating bodies which was derived in a special case earlier (*Pacific J. Math.* 231:1 (2007), 167–191) and is here generalized to account for gravitational forces.**

**A formal application of our results is made to the problem of a ball floating in an infinite bath asymptotic to a prescribed level. We obtain existence and nonuniqueness results.**

### 1. Introduction

Archimedes stated the principle that bears his name in a work titled *On floating bodies*. The principle is commonly stated as follows:

*A body immersed in a fluid is buoyed up with a force equal to the weight of the displaced fluid.*

This is actually a reformulation of Archimedes' principle and, as Erlend Graf [2004] points out, it is deficient (and incorrect) in various respects.

Archimedes considered three distinct cases. The first case is that in which the density of the body is equal to the density of the liquid. The assertion is that the body, after it is deposited into the liquid and comes to rest, will not project above the surface of the liquid nor sink lower in the liquid (*On floating bodies*, part I, Proposition 3; see [Archimedes/Heath 1897, p. 255]).

The second case is that in which the density of the body is less than that of the liquid. The assertion is that the body, if left to interact freely with the liquid bath, will project above the surface of the bath and will displace a volume of liquid having the same weight as the object (Propositions 4 and 5; [ibid., pp. 256–257]). Furthermore, if the object is not allowed to float freely, but is manually pushed downward into the liquid from its floating position, then the object will experience an upward force equivalent to the difference of the weight of the object and the weight of the displaced liquid (Proposition 6; [ibid., p. 257]).

Finally, if the body is more dense than the liquid it will sink to the bottom and, if weighed while in the liquid will be found lighter than its true weight by the weight of the displaced liquid (Proposition 7; [ibid., p. 258]).

The reformulation is about the force experienced by a body deposited in a liquid bath (and nothing else). The original principle of Archimedes specifically addresses two additional questions:

- (1) Will the body float<sup>1</sup> or sink?
- (2) At what height will the object come to rest?

The first question is conditional; the second is geometric. The fact that the reformulation ignores these aspects of the problem is a deficiency of the reformulation and no reflection on the acuity of Archimedes.

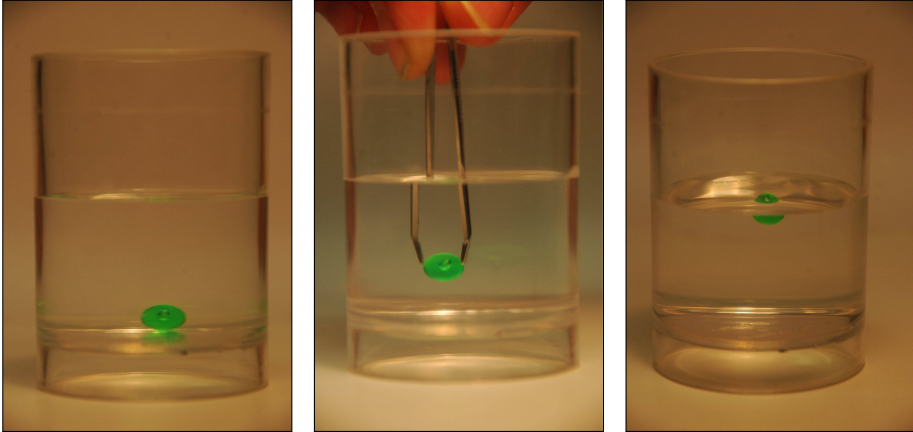
An aspect of the problem that does seem to have escaped the notice of Archimedes involves the effect of surface tension or surface energy associated with wetting. Indeed, simple experiments show that it is possible, under certain circumstances, for even a convex<sup>2</sup> object with density greater than that of a given liquid bath to float (only) partially submerged on the surface of the bath, contradicting Archimedes' Proposition 7; see Figure 1.

Finn [2011] has recently given the first rigorous mathematical proof of this fact, at least in an idealized situation which we describe in Section 4 below. Finn and Vogel [2009] wrote: "One may assume that [Archimedes] was unaware of observations of Aristotles a century earlier" (concerning heavy floating objects). This may be true, or perhaps Archimedes restricted himself to a problem whose solution used the mathematical tools he had at hand. In either case, we find connections with the results of Archimedes, and derive from our new results what can be viewed as a generalization of results which follow from Archimedes' approach. Notice also

---

<sup>1</sup>That is, will the object project above the surface of the liquid?

<sup>2</sup>Convexity is mentioned here in contrast to something like a hollow boat hull often considered in connection with the density considerations of Archimedes. In fact, the possibility that objects with density greater than water might float on the surface of water was already considered by Aristotle a century before Archimedes, and it is surprising Archimedes makes no mention of it. The fact that a thin metal paper clip can float on water makes it clear convexity is not a necessary hypothesis. Nevertheless, we did not know if a sphere could float until we tried it (Figure 1).



**Figure 1.** Photos of a plastic ball in a bath of water: sinking to the bottom (left), being raised to the surface (middle), floating (right)

that the results of [Finn 2011] and [Finn and Vogel 2009] initiate a return to the question addressed by Archimedes: Does the body sink or swim?

Our work below assumes the answer to the question of floating versus sinking is affirmative for floating and seeks to answer a version of Archimedes' second question: What is the geometry? More precisely: What is the height of the floating body and what is the geometry of the interface? We are able to give a partial answer under the assumption of rotational symmetry of the object and the interface. This symmetry appears to hold in the physical system of Figure 1, and similar symmetric interfaces have been shown to exist mathematically in [Treinen 2012] and [Elcrat et al. 2004b]. For further discussion of this point, see Section 6.

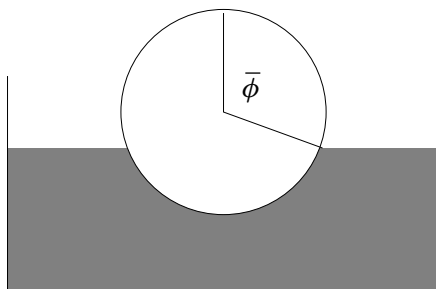
For purposes of comparison, we describe briefly this problem of a floating ball as we imagine Archimedes might have considered it.<sup>3</sup> Given the diagram in Figure 2, with an assumed planar interface meeting a floating sphere  $\Sigma$  along a circular contact line determined by an azimuthal angle  $\bar{\phi}$ , and assuming a density  $\rho$  of the ball less than the density  $\rho_l$  of the liquid, Archimedes' Proposition 5 then becomes

$$(1) \quad \rho_l V_d = \rho |\Sigma|$$

where  $V_d$  is the volume of displaced liquid. Equating this volume of liquid with the volume of the spherical cap below the plane of the interface,

$$V_d = \frac{1}{3}\pi a^3 (\sin^2 \bar{\phi} \cos \bar{\phi} + 2 + 2 \cos \bar{\phi}),$$

<sup>3</sup>The explanation of Vitruvius (in *De architectura*) is of particular interest for this discussion, as it provides some details not contained in Archimedes' work directly. In particular, Vitruvius identified the "displaced fluid" as that which overflows a vessel into which an object is deposited.



**Figure 2.** Azimuthal angle determined by a horizontal contact line.

we obtain this:

**Theorem 1.** *According to Archimedes' principle, a homogeneous sphere of density  $\rho > \rho_l$  will sink to the bottom of a bath of density  $\rho_l$ , and a homogeneous sphere of density  $\rho < \rho_l$  will float at a level determined by*

$$(2) \quad \cos^3 \bar{\phi} - 3 \cos \bar{\phi} = 2 \left( 1 - \frac{2\rho}{\rho_l} \right).$$

It is easily checked that the function  $F(\bar{\phi}) = \cos^3 \bar{\phi} - 3 \cos \bar{\phi}$  is increasing from  $-2$  to  $2$  on  $[0, \pi]$ , with zero derivative at the endpoints and strictly positive derivative interior to the interval. Thus, for each positive value  $0 \leq \rho \leq \rho_l$ , the condition (2) determines a unique azimuthal angle. See Figure 3.

**Definitions of equilibria.** From a more sophisticated point of view, liquid interfaces are rarely planar. Even without the introduction of a floating object, the interface of liquid in a cylinder is usually noticeably curved around the edges. With the introduction of a rigid floating object, one may assume the interface will be further deformed in possibly unexpected ways.

The modern theory of equilibrium capillary configurations developed by Young, Laplace, Gauss, and others (see [Finn 1986]) is now founded on the consideration of energies associated with the area of the outer surface of the liquid where it contacts the surrounding atmosphere and where it contacts the bounding container. This theory has been primarily pursued in the context of solid structures that are rigid and *fixed*. This has led to a commonly adopted definition of a capillary equilibrium [ibid.]:

Up to the determination of a single real parameter ( $\lambda$  below) the problem of finding a capillary surface is a purely geometric one: *to find a surface whose mean curvature is a prescribed function of position and which meets prescribed (rigid) bounding walls in a prescribed angle  $\gamma$ .*



In terms of equations commonly used to model equilibrium capillary surfaces in a gravity field, we have

$$(3) \quad 2H = \kappa z - \lambda \quad \text{and} \quad \cos \gamma = \beta,$$

where  $H$  denotes the mean curvature of the interface,  $z$  denotes the vertical height of a point on the interface,  $\kappa = \rho_l g / \sigma$  is the *capillary constant*, constructed using the gravitational acceleration  $g$  and the *surface tension*  $\sigma$ , and  $\lambda$  is a single real (Lagrange) parameter related to the constraints of the problem; in the second equation one finds the *relative adhesion coefficient*  $\beta$  defined by the assumption that  $\sigma\beta$  is the local energy density<sup>4</sup> associated with contact between the liquid volume and solid structures; one integrates  $\sigma\beta$  over the area of contact, or *wetted area*, to obtain the total *energy of wetting*. The angle  $\gamma$  is assumed to be defined along a curve where the liquid, the container, and the surrounding atmosphere all meet. This curve is called the *contact line* and  $\gamma$  is referred to as the *contact angle*.

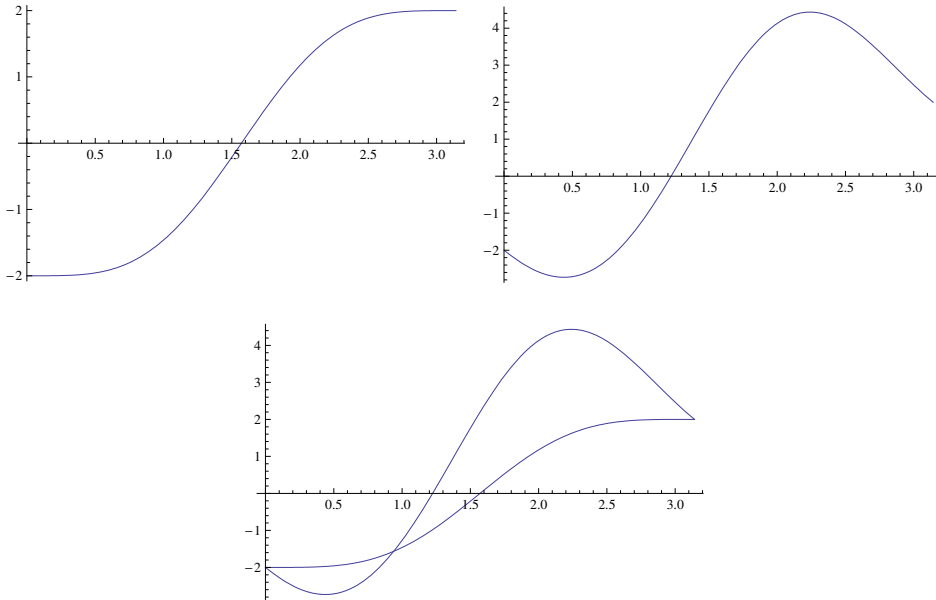
While the problem of a floating object considered here is still purely geometric, the conditions (3) are inadequate to characterize equilibria, even if the object is rigid and the Lagrange parameter  $\lambda$  is known. One still has recourse to the general principle of virtual work, that is, the energy is stationary with respect to variations compatible with the constraints of the problem. Nevertheless, attaining a collection of fundamental necessary conditions analogous to (3) that may be taken as a working definition of equilibrium in particular cases is of evident utility both for applications and the mathematical theory of capillarity. A preliminary discussion of the need for this development was suggested in [McCuan 2007] in the absence of external forces (i.e., zero gravity), and we provide here a general flux condition (13) to augment (3), thus providing a new definition of equilibrium in this context. A discussion of this formula for capillary surfaces is in Section 2.

From the flux formula we obtain the following result which may be compared to Theorem 1 and is proved in Section 3.

**Theorem 2.** *A sphere of radius  $a$  that floats in a centrally symmetric position as described above under the effects of surface tension and adhesion effects of an axially symmetric bath must float at a level determined by the azimuthal angle  $\bar{\phi}$  satisfying*

$$(4) \quad \cos^3 \bar{\phi} - 3 \cos \bar{\phi} + \frac{6}{\kappa a} \left( \bar{H} + \frac{\cos \gamma}{a} \right) \sin^2 \bar{\phi} - \frac{3 \sin \gamma}{\kappa a^2} \sin(2\bar{\phi}) = 2 \left( 1 - \frac{2\rho}{\rho_l} \right),$$

<sup>4</sup>In [Finn 1986], the relative adhesion coefficient is given on page 6 as the difference  $\beta^* - \hat{\beta}^*$  of energy densities associated with contact between one fluid and the container ( $\beta^*$ ) and a complementary fluid and the container ( $\hat{\beta}^*$ ). Using the approximation  $\hat{\beta}^* \approx 0$ , the formulation used here is equivalent. For simplicity, we will also assume  $\sigma$  and  $\beta$  are constants; the reasoning below extends in a straightforward manner to the general case.



**Figure 3.** The azimuthal angles determined by Theorems 1 (top left) and 2 (top right); plotted together on the bottom.

where  $\kappa$  is the capillary constant described above,  $\bar{H}$  is the mean curvature of the liquid interface at the contact line, and  $\gamma \in (0, \pi)$  is the contact angle of the liquid interface with the floating sphere.

The function  $F(\bar{\phi})$  appearing on the left side of (4) takes the values  $-2$  and  $2$  at the endpoints  $\bar{\phi} = 0$  and  $\pi$  respectively. However,  $F$  is decreasing at  $\bar{\phi} = 0$  and decreases to a unique local interior minimum at  $\bar{\phi} = \bar{\phi}_1$ . On the interval from  $\bar{\phi} = \bar{\phi}_1$  to  $\bar{\phi} = \pi$  the function  $F$  has a unique interior local maximum at  $\bar{\phi} = \bar{\phi}_2$ .

If  $\gamma = 0, \pi$ , then the value of the azimuthal angle is uniquely determined by the same function  $F$ , which is increasing and satisfies  $F'(0) = 0 = F'(\pi)$  but is distinct from the function appearing in Theorem 1.

The existence of the unique local interior minimum at  $\bar{\phi} = \bar{\phi}_1$  allows values of  $\rho > \rho_l$  and leads to the determination of a unique maximum density  $\rho_{\max} = \rho_{\max}(a, \gamma, \kappa, \bar{H})$  for which  $\rho > \rho_{\max}$  implies no floating is possible. It will be noted from the properties of  $F$  that a unique azimuthal angle  $\bar{\phi}$  is determined for all values  $0 < \rho < \rho_l$ , and that two values are possible for certain values  $\rho \geq \rho_l$  (as long as  $\rho$  is not too large). We presume by continuity that the physically observed value for heavy floating spheres is the larger one determined by (4). The physical relevance of the other value is discussed in Section 6 of the paper.

We note also that the graph of  $F$  takes values corresponding to negative densities  $\rho$ . This can be imagined to have physical relevance in a situation where a

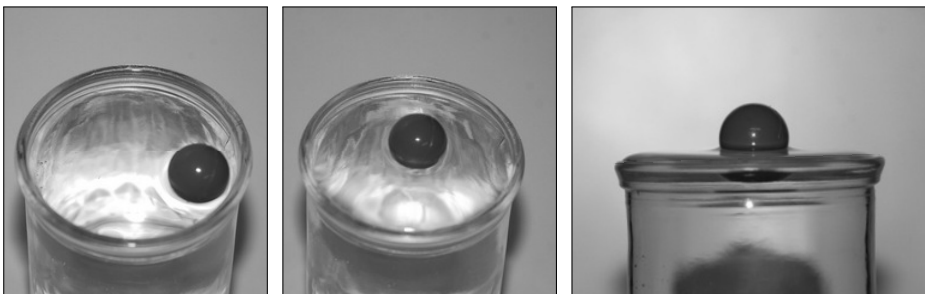
gravitational field acts on the floating object, but one with the opposite direction as that acting on the liquid. It is not readily apparent how such a physical situation would arise, but one can easily imagine a magnetic field producing an upward force on a floating object in a downward gravity field, which would be quite similar.

**Further remarks.** The quantity  $\bar{H}$  appearing in the formula (4) of Theorem 2 is presumed to depend in some manner on other parameters, and perhaps globally imposed geometric constraints in the problem. Perhaps the quantity  $\bar{H}$  and its appearance in (4) is best viewed in contrast to the following specific quantities: the enclosed volume of liquid (n.b., the Lagrange parameter  $\lambda$ ), the outer radius  $R$  of the cylindrical vessel, and the contact angle  $\gamma_{\text{out}}$  between the interface and the outer wall, all of which are conspicuously absent from formula (4). As far as we know, this paper and [McCuan 2007] are the first to consider the global floating configuration for a floating ball including a finite outer bounding wall. Indeed, one might be tempted to dismiss the effects of the interface at the outer bounding wall. Several authors have considered floating objects in an infinite bath asymptotic to a plane (and we do so below in § 6 as well). Under certain assumptions, estimates have been derived [Siegel 1980] to establish the fact that such an interface converges to the planar asymptote exponentially with distance from a floating object.

We offer the following description of an experiment as a caution against assuming the influence of an outer wall is not important.

If a cylinder of water is partially filled, and a ball of density  $\rho < \rho_l$  is deposited in the center of the resulting interface, it will move rapidly to the outer wall. See Figure 4. If the same cylinder is subsequently slightly overfilled so that the (roughly flat) interface curves downward at the edges, then the ball will move rapidly to the center of the interface and remain there in an apparently stable configuration; if the ball is manually moved away from the center it will return.

This experiment brings up a question that is fundamentally different from the one considered in this paper, but it indicates in broad terms that the question of



**Figure 4.** Photos of a plastic ball in a bath of water: tending to the edge (left), stable in the middle (center and right).

how an object floats on a liquid interface can have an answer depending strongly on nonlocal conditions involving the outer bounding wall.

Ideally one would like a formula for the azimuthal angle  $\bar{\phi}$  in terms of the volume of liquid in the bath  $V$ , the radii  $a$  and  $R$  of the ball and the container respectively, and the contact angles  $\gamma$  between the liquid interface and the surface of the floating object and  $\gamma_{\text{out}}$  between the liquid interface and the surface of the container, and from the classical point of view, this is what one would expect. We were unable to attain such a result, and the result we obtain (4) may be viewed simply as a relation between  $\bar{H}$  and  $\bar{\phi}$  for any equilibrium. The interpretation we give in the context of Archimedes' geometric question may then be viewed as the most explicit currently available information arising from (13).

The barrier to getting a more definitive result lies in the complicated nature of the system of ordinary differential equations determining the rotationally symmetric interface. For a survey of recent progress in understanding the family of solutions to these equations, see [Finn 1986; Vogel 1982; Siegel 2006; Siegel 1980; Nickolov 2002; Elcrat et al. 2004a; Turkington 1980; Johnson and Perko 1968; Treinen 2012].

## 2. Variational formulation

The general assumptions of our model are outlined in [McCuan 2007] though the derivation given there was aimed at the zero gravity case in which buoyancy plays no role, and the effects of gravity were not properly considered. For the sake of making this paper somewhat more self-contained we include a short review/summary of the model and amend the deficiencies in the former derivation.

Quite generally, we consider a solid structure

$$\Sigma = \Sigma_s \cup \Sigma_m$$

consisting of a stationary part  $\Sigma_s$  and a movable, or floating, part  $\Sigma_m$ . In addition, we hypothesize an equilibrium liquid interface  $\Lambda$  with corresponding wetted region  $\mathcal{W} = \mathcal{W}_s \cup \mathcal{W}_m$ , so that the liquid volume  $\mathcal{V}$  satisfies  $\partial\mathcal{V} = \Lambda \cup \mathcal{W}$  and the contact line/triple interface is given by  $\partial\Lambda = \partial\mathcal{W}$ . Under these assumptions, we consider the variational problem associated with

$$(5) \quad \mathcal{E} = \sigma|\Lambda| - \sigma\beta|\mathcal{W}| + \mathcal{G}$$

where  $\mathcal{G} = \int_{\mathcal{V} \cup \Sigma_m} G$  and  $G$  is a position dependent function representing field forces such as gravity.<sup>5</sup>

One specific application of the discussion which now follows is that it justifies the following fundamental definition:

<sup>5</sup>We included only  $\int_{\mathcal{V}} G$  in [McCuan 2007].

**Definition 3.** A floating configuration  $\Sigma_s, \Sigma_m, \mathcal{V}$  as described above is said to be in *free-floating equilibrium* for the functional (5) if

1.  $2H = G/\sigma - \lambda$ , where  $H$  is the mean curvature of the free surface interface  $\Lambda$  and  $\lambda$  is some constant,
2.  $\cos \gamma = \beta$  where  $\gamma$  is the angle at which the free surface interface meets the surface of the solid structures measured within  $\mathcal{V}$  and  $\beta$  is the (possibly location dependent) adhesion coefficient, and
3. 
$$\int_{\partial \mathcal{W}_m} \vec{n} + \int_{\mathcal{W}_m} (G/\sigma - \lambda)N - \int_{\partial \Sigma_m} (G/\sigma)N = 0,$$

where  $n$  is the outward pointing unit conormal along  $\partial \Lambda$ , and  $N$  is the unit normal to  $\partial \mathcal{V}$  pointing out of  $\mathcal{V}$ .

Under rather general hypotheses, as described in [McCuan 2007], a family of variations leaving  $\Sigma_m$  fixed leads to the (standard) variational formulas (6)–(8) below:

$$(6) \quad |\dot{\Lambda}| = - \int_{\Lambda} 2H \dot{X} \cdot N + \int_{\partial \Lambda} \dot{X} \cdot \vec{n},$$

where  $H$  is the mean curvature defined on  $\Lambda$ ,  $\dot{X}$  is the variation vector,  $N$  is the unit normal pointing out of the liquid volume  $\mathcal{V}$ , and  $\vec{n}$  is the unit conormal to  $N$  and  $\partial \Lambda$  pointing out of  $\Lambda$ ;

$$(7) \quad |\dot{\mathcal{W}}| = \int_{\partial \Lambda} \dot{X} \cdot \vec{v},$$

where  $\vec{v}$  is the unit conormal to  $N^{\mathcal{W}}$  and  $\partial \mathcal{W}$  pointing out of  $\mathcal{W}$ ; note that  $N^{\mathcal{W}}$  denotes the unit normal to  $\mathcal{W}$  pointing out of  $\mathcal{V}$  and may also be denoted by  $N$  on the interior of  $\mathcal{W}$  where no ambiguity arises;

$$(8) \quad \dot{\mathcal{E}} = \int_{\Lambda} G \dot{X} \cdot N \quad \text{and} \quad |\dot{\mathcal{V}}| = \int_{\Lambda} \dot{X} \cdot N.$$

These last two formulas apparently require an interesting and somewhat delicate application of more general mathematical principles of fluid mechanics, and we outline their derivation under more general assumptions below.

For now, we assemble  $\dot{\mathcal{E}}/\sigma - \lambda|\dot{\mathcal{V}}|$  from the constituent parts above where  $\lambda$  is a Lagrange multiplier associated with the volume constraint:

$$\dot{\mathcal{E}}/\sigma - \lambda|\dot{\mathcal{V}}| = \int_{\Lambda} (-2H + G/\sigma - \lambda)\dot{X} \cdot N + \int_{\partial \Lambda} (\dot{X} \cdot \vec{n} - \beta \dot{X} \cdot \vec{v}).$$

The vanishing of this quantity for all variation vectors  $\dot{X}$  results in the well known

geometric boundary value problem

$$(9) \quad \begin{cases} 2H = G/\sigma - \lambda & \text{on } \Lambda, \\ \cos \gamma = \beta & \text{on } \partial\Lambda, \end{cases}$$

since

$$\vec{n} = (\vec{n} \cdot N^{\mathcal{W}})N^{\mathcal{W}} + \cos \gamma \vec{v}.$$

In the special case under consideration in this paper,  $G$  represents the limiting value  $\rho_l g z$  taken as a limit from inside the liquid, so that

$$2H = \kappa z - \lambda$$

where  $\kappa = \rho_l g/\sigma$  is a capillary constant for the problem. Furthermore, we restrict attention in this paper to cases in which the adhesion coefficient satisfies  $-1 < \beta < 1$  or equivalently, the contact angle  $\gamma$  is strictly between 0 and  $\pi$ .

A more general variation allowing rigid motion of  $\Sigma_m$  takes the form

$$X = X(\mathbf{p}; t, h) : M \times (-\epsilon, \epsilon) \times (-\delta, \delta) \rightarrow \mathbb{R}^3,$$

where  $M = \overline{\Sigma \cup \mathcal{V}}$  is considered as an abstract manifold; see Figure 5.

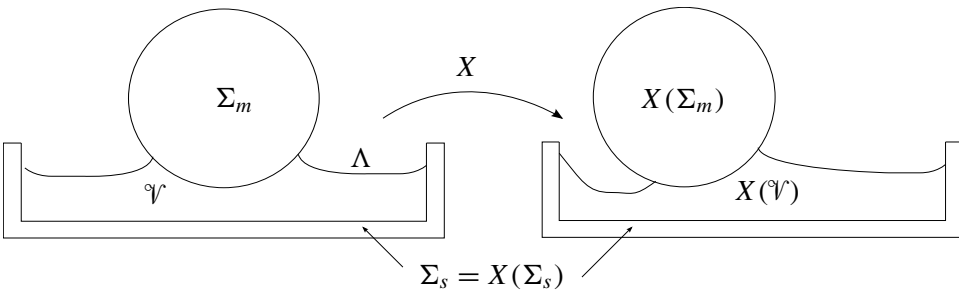
It is assumed here, as indicated in the figure, that  $h$  parametrizes a family of rigid motions  $w = w(x; h)$  to which  $\Sigma_m$  is subject. Denoting derivatives with respect to  $h$  by an acute accent, we find

$$(10) \quad |\acute{\Lambda}| = - \int_{\Lambda} 2H \acute{X} \cdot N + \int_{\partial\Lambda} \acute{X} \cdot \vec{n},$$

$$(11) \quad |\acute{\mathcal{W}}| = - \int_{\mathcal{W}_m} 2H^{\mathcal{W}} \acute{X} \cdot N + \int_{\partial\mathcal{W}_m} \acute{X} \cdot \vec{v},$$

$$(12) \quad \acute{\mathcal{G}} = \int_{\Lambda} G \acute{X} \cdot N + \int_{\mathcal{W}_m} G \acute{X} \cdot N^{\mathcal{W}} + \int_{\partial\Sigma_m} G_m \acute{X} \cdot N^m.$$

This last term requires some explanation. The quantity  $G_m$  denotes the value of the volumetric force field potential taken as a limit from inside the movable solid



**Figure 5.** The variation map and its notation.

structure  $\Sigma_m$ . In the special case of a floating object of density  $\rho$ , we typically take  $G_m = \rho g z$ . Also in this last identity  $N^m$  denotes the unit normal to the boundary  $\partial\Sigma_m$  of the movable/floating solid structure and points out of  $\Sigma_m$ , so that  $N^m = -N^W$  on their common domain of definition  ${}^cW_m$ . Finally, we include a brief derivation.

Up until this point, we have stated all variational formulae in their final form, that is to say with the parameters of the variation set to zero so that  $\dot{X}$  represents

$$\left. \frac{d}{dt} X(\mathbf{p}; t) \right|_{t=0},$$

where  $X = X(\mathbf{p}; t) : M \times (-\epsilon, \epsilon) \rightarrow \mathbb{R}^3$ . For this calculation, we must temporarily assume the parameters  $t$  and  $h$  are not evaluated at zero. Notationally, this is conveniently indicated by a tilde so that  $\tilde{\Sigma}_m = X(\Sigma_m) = X(\Sigma_m; t; h)$ , and we will evaluate at  $t = h = 0$  at the end.

Consideration of the second term should suffice. Setting

$$\mathcal{G}_m = \int_{\tilde{\Sigma}_m} G_m,$$

we have

$$\mathcal{G}_m = \int_{\Sigma_m} G_m \circ X \det DX,$$

where  $X$  represents the restriction of the variation to  $\Sigma_m$  and the derivative is taken in  $M \subset \mathbb{R}^3$  with respect to  $\mathbf{p}$ . Euler's kinematical formula [Serrin 1959] tells us how a material integral changes with the flow of a region of fluid. We can cast our present situation into this framework starting with the preliminary identity

$$\frac{\partial}{\partial h} \det DX = (\operatorname{div}_{\mathbb{R}^3} \mathbf{v}) \circ X \det DX$$

where  $\mathbf{v}(x; h) = \dot{X}(X^{-1}(x; h); h)$  is the spatial velocity associated with the flow  $X = X(\mathbf{p}; h)$  and we have simply suppressed the  $t$  dependence. It might be expected (or hoped) that in our situation the motion/flow associated with the variation should be particularly simple, at least on the solid movable object  $\Sigma_m$ , and that we might have, for example,  $X(\mathbf{p}; h) \equiv w(\mathbf{p}; h)$  there. However, taking into account the motion of the liquid and that of the contact line of the liquid interface  $\Lambda$  in particular, it is clear that this would violate the continuity assumption on the variation  $X : M \times (-\epsilon, \epsilon) \times (-\delta, \delta) \rightarrow \mathbb{R}^3$ . Having made this concession and subjected ourselves to the added complication that other authors seem to have avoided, it is some consolation, as pointed out in [Finn 2005], that the internal motion of the liquid under a variation of the free surface interface could be very complicated, and we are taking account of such possibilities.

In any case, we continue to obtain

$$\dot{\mathcal{G}}_m = \int_{\tilde{\Sigma}_m} DG_m \cdot \mathbf{v} + \operatorname{div}_{\mathbb{R}^3} \mathbf{v} = \int_{\tilde{\Sigma}_m} \operatorname{div}_{\mathbb{R}^3} (G_m \mathbf{v}) = \int_{\partial \tilde{\Sigma}_m} G_m \mathbf{v} \cdot N^m,$$

so that

$$\dot{\mathcal{G}}_m|_{h=0} = \int_{\partial \Sigma_m} G_m \dot{X} \cdot N^m.$$

A similar argument applies to the integral over  $\mathcal{V}$  appearing in  $\mathcal{G}$  and also yields

$$|\dot{\mathcal{V}}| = \int_{\Lambda} \dot{X} \cdot N + \int_{\mathcal{W}_m} \dot{X} \cdot N,$$

where we have returned to the general assumption on evaluation, that  $t = h = 0$ . Combining this with (10)–(12), we have

$$\begin{aligned} \dot{\mathcal{E}}/\sigma - \lambda |\dot{\mathcal{V}}| &= \int_{\Lambda} (-2H + G/\sigma - \lambda) \dot{X} \cdot N + \int_{\partial \Lambda} (\dot{X} \cdot \vec{n} - \beta \dot{X} \cdot \vec{v}) \\ &\quad + \beta \int_{\mathcal{W}_m} 2H^{\mathcal{W}} \dot{X} \cdot N + \int_{\mathcal{W}_m} (G/\sigma - \lambda) \dot{X} \cdot N + \int_{\partial \Sigma_m} (G_m/\sigma) \dot{X} \cdot N^m \\ &= \int_{\partial \mathcal{W}_m} \dot{X} \cdot \vec{n} - \cos \gamma \int_{\partial \mathcal{W}_m} \dot{X} \cdot \vec{v} \\ &\quad + \cos \gamma \int_{\mathcal{W}_m} 2H^{\mathcal{W}} \dot{X} \cdot N + \int_{\mathcal{W}_m} (G/\sigma - \lambda) \dot{X} \cdot N + \int_{\partial \Sigma_m} (G_m/\sigma) \dot{X} \cdot N^m. \end{aligned}$$

Next we refer to a calculation from [McCuan 2007] that uses the fact that

$$w^{-1}(X; h) \in \Sigma_m$$

when  $X = X(\mathbf{p}; h) \in w(\Sigma_m; h)$  to show that

$$\dot{X} - \dot{w} \in T_X \Sigma_m.$$

It follows that  $\dot{X}$  may be replaced with  $\dot{w}$  in the formula above. A second calculation involving an explicit auxiliary variation shows

$$\int_{\mathcal{W}_m} 2H^{\mathcal{W}} \dot{w} \cdot N = \int_{\partial \mathcal{W}_m} \dot{w} \cdot \vec{v}.$$

Making the indicated substitutions, we arrive at our new necessary condition for equilibrium of a floating object:

**Theorem 4.** *If a floating configuration  $\Sigma_m, \mathcal{V}$  subject to forces (having volumetric potentials denoted by  $G$  and  $G_m$  as described above) locally minimizes energy among liquid interface configurations compatible with a smooth family of rigid*



motions  $w = w(x; h)$  with  $w(x; 0) = \text{id}_{\mathbb{R}^3}$  and the wetted region on the floating object is denoted by  ${}^{\mathfrak{W}}\mathcal{W}_m$ , then the configuration must satisfy

$$(13) \quad \int_{\partial {}^{\mathfrak{W}}\mathcal{W}_m} \dot{w} \cdot \vec{n} + \int_{{}^{\mathfrak{W}}\mathcal{W}_m} (G/\sigma - \lambda) \dot{w} \cdot N^{\mathfrak{W}} + \int_{\partial \Sigma_m} (G_m/\sigma) \dot{w} \cdot N^m = 0,$$

where  $\vec{n}$  is the outward pointing unit conormal along the boundary of the liquid interface  $\Lambda$ ,  $N^{\mathfrak{W}}$  is the unit normal to  $\Sigma_m$  pointing out of the liquid,  $N^m = -N^{\mathfrak{W}}$ , and  $\dot{w}$  represents the derivative with respect to  $h$  evaluated at  $h = 0$ .

The condition of the theorem must hold for all  $\dot{w} \in \mathbb{R}^3$  for free floating, or more generally for any collection of directions in which  $\Sigma_m$  is free to move. In the case in which all directions  $\dot{w}$  are possible, the condition (13) simplifies to

$$\int_{\partial {}^{\mathfrak{W}}\mathcal{W}_m} \vec{n} + \int_{{}^{\mathfrak{W}}\mathcal{W}_m} (G/\sigma - \lambda) N^{\mathfrak{W}} + \int_{\partial \Sigma_m} (G_m/\sigma) N^m = 0.$$

One immediately notes the integral over the boundary of the movable wetted surface of the conormal to the free surface interface (the first term) as marking this as a kind of *flux formula* or *force balance formula* as is well known from the work of A. Ros [1996] in minimal surfaces. It is tempting to interpret the other two integrals appearing in the formula as force vectors, and without doubt they are such. We are indebted to a referee for explaining how to do this for a constant vertical gravity field. Similar calculations for that case are also contained in [Bhatnagar and Finn 2006] where a somewhat different problem is considered; see Sections 4 and 6 for further remarks. With this help, we were able to see the following general interpretation.

In order to be dimensionally correct, multiply the equation by the surface tension  $\sigma$ . The first term is then the negative of the force exerted on the object by the interface itself — the surface tension force.

The integrand of the second term  $G - \lambda\sigma$  will be recognized from (9) as the quantity  $2\sigma H$  at the interface and, according to the insight of Thomas Young, the difference in *pressure* across the interface. It is natural to assume that  $G - \lambda\sigma$  gives a pressure field extending throughout the volume of liquid, up to a sign. Since the mean curvature is calculated with respect to the normal  $N$  pointing out of the liquid, we see that the second integral represents the negation of the force this pressure exerts on the floating object, i.e., the buoyancy force.

Let us consider the third term componentwise. If  $e_j$  is the  $j$ -th standard unit vector, then the  $j$ -th component of the third integral is

$$\int_{\partial \Sigma_m} G_m e_j \cdot N^m = \int_{\Sigma_m} \text{div}(G_m e_j) = \int_{\Sigma_m} DG_m \cdot e_j,$$

where the first equality is by the divergence theorem, and we recognize the negation

of the volumetric force density in the gradient of the potential appearing in the last expression. Recombining the components, the third term

$$\int_{\Sigma_m} DG_m$$

evidently lends itself to being interpreted as (minus) the “weight” of the floating object with respect to the potential field  $G_m$ .

In summary, our third equilibrium condition may be read (without the slightest ambiguity in the case of a constant downward gravitational field  $G_m = \rho g z$ ) thus:

*The weight, the pressure/buoyancy force, and the surface tension force on the floating object must sum to zero.*

We next proceed to examine the consequences of (13) for the simple cases of floating suggested in the introduction.

### 3. Floating in three dimensions

Here we assume a vertical circular cylindrical vessel is observed with a sphere  $\Sigma_m$  floating symmetrically along the axis of the vessel and having symmetric circular contact line at azimuthal angle  $\phi = \bar{\phi}$ . Assuming the surface of the liquid is also rotationally symmetric with respect to the same axis, the meridian of the surface with vertical component  $u$  and radial component  $r$  considered as functions of arclength along the meridian must satisfy the boundary value problem

$$(14) \quad \begin{cases} \dot{r} = \cos \psi, \\ \dot{u} = \sin \psi, \\ \dot{\psi} = \kappa u - \lambda - \sin \psi / r, \\ \psi = \gamma - \bar{\phi} \text{ and } u = d + a \cos \bar{\phi} & \text{when } r = r(0) = a \sin \bar{\phi}, \\ \psi = \pi/2 - \gamma_{\text{out}} & \text{when } r = r(l) = R, \end{cases}$$

where we have chosen coordinates so that the center of the floating sphere is  $(0, 0, d)$ , and we have denoted by  $l$  the total length and by  $\psi$  the inclination angle of the meridian.

It would be desirable to preface our discussion of the geometry of the floating ball in Figure 1 with an existence result, but we are unable to obtain such a result for essentially the same reason that our geometric result is somewhat suboptimal: The system of ordinary differential equations appearing in the problem above has been studied extensively, but the structure of the family of all solutions is not well enough understood. Thus, we turn directly to the auxiliary condition (13).

The following formulae, valid in the plane  $y = x_2 = 0$ , are useful in simplifying the integrals in (13):

$$\begin{aligned}
N^m[\phi] &= \sin \bar{\phi} \mathbf{e}_1 + \cos \bar{\phi} \mathbf{e}_3, \\
N^{\mathcal{W}}[\phi] &= -N^m \\
&= -\sin \bar{\phi} \mathbf{e}_1 - \cos \bar{\phi} \mathbf{e}_3, \\
\vec{v}[\phi] &= (N^m)^\perp \\
(15) \quad &= -\cos \bar{\phi} \mathbf{e}_1 + \sin \bar{\phi} \mathbf{e}_3, \\
\vec{n} &= \cos \gamma \vec{v} + \sin \gamma N^{\mathcal{W}} \\
&= -\cos(\bar{\phi} - \gamma) \mathbf{e}_1 + \sin(\bar{\phi} - \gamma) \mathbf{e}_3, \\
N^\Lambda &= (-\vec{n})^\perp \\
&= \sin(\bar{\phi} - \gamma) \mathbf{e}_1 + \cos(\bar{\phi} - \gamma) \mathbf{e}_3.
\end{aligned}$$

In these formulae, the bracketed  $\phi$  indicates validity in the form of the result for an arbitrary azimuthal angle on  $\partial \Sigma_m$  though the main interest is on  $\partial^{\mathcal{W}} \mathcal{W}_m$ ;  $\mathbf{e}_1$  and  $\mathbf{e}_3$  are the standard orthonormal unit vectors in  $\mathbb{R}^3$ .

Taking a vertical translation for the rigid motion of  $\Sigma_m$  so that  $\acute{w} = \mathbf{e}_3$ , the three terms of (13) are as follows:

$$\begin{aligned}
\int_{\partial^{\mathcal{W}} \mathcal{W}_m} \mathbf{e}_3 \cdot \vec{n} &= 2\pi a \sin \bar{\phi} \sin(\bar{\phi} - \gamma), \\
\int_{\mathcal{W}_m} (\kappa z - \lambda) \mathbf{e}_3 \cdot N &= \pi a^2 \left( (\kappa d - \lambda) \sin^2 \bar{\phi} - \frac{2}{3} \kappa a (1 + \cos^3 \bar{\phi}) \right), \\
\int_{\partial \Sigma_m} \kappa \frac{\rho}{\rho_l} z \mathbf{e}_3 \cdot N^m &= \frac{4}{3} \pi \kappa a^3 \frac{\rho}{\rho_l}.
\end{aligned}$$

Combining these terms and rearranging:

$$(16) \quad \frac{6 \sin \bar{\phi} \sin(\bar{\phi} - \gamma)}{\kappa a^2} + \frac{3(\kappa d - \lambda) \sin^2 \bar{\phi}}{\kappa a} - 2 \cos^3 \bar{\phi} = 2 \left( 1 - \frac{2\rho}{\rho_l} \right).$$

Next, we make the substitution

$$2\bar{H} = \kappa(d + a \cos \bar{\phi}) - \lambda,$$

which follows directly from (9). This leads to

$$\frac{6 \sin \bar{\phi} \sin(\bar{\phi} - \gamma)}{\kappa a^2} + \frac{3(2\bar{H} - \kappa a \cos \bar{\phi}) \sin^2 \bar{\phi}}{\kappa a} - 2 \cos^3 \bar{\phi} = 2 \left( 1 - \frac{2\rho}{\rho_l} \right).$$

This last condition simplifies directly into condition (4) of Theorem 2. It remains to verify the description of the function

$$F(\bar{\phi}) = \cos^3 \bar{\phi} - 3 \cos \bar{\phi} + \frac{6}{\kappa a} \left( \bar{H} + \frac{\cos \gamma}{a} \right) \sin^2 \bar{\phi} - \frac{3 \sin \gamma}{\kappa a^2} \sin(2\bar{\phi}),$$

where  $\bar{H}$  is taken to be a given constant. The values at the endpoints are immediate. We find also that

$$\begin{aligned} \frac{F'(\bar{\phi})}{3} &= -\cos^2 \bar{\phi} \sin \bar{\phi} + \sin \bar{\phi} + \frac{4}{\kappa a} \left( \bar{H} + \frac{\cos \gamma}{a} \right) \sin \bar{\phi} \cos \bar{\phi} - \frac{2 \sin \gamma}{\kappa a^2} \cos(2\bar{\phi}) \\ &= \sin^3 \bar{\phi} + \frac{2}{\kappa a} \left( \bar{H} + \frac{\cos \gamma}{a} \right) \sin(2\bar{\phi}) - \frac{2 \sin \gamma}{\kappa a^2} \cos(2\bar{\phi}). \end{aligned}$$

Thus,  $F'(0) = F'(\pi) = -(6/\kappa a^2) \sin \gamma < 0$ . From this it is clear that  $F$  must attain an absolute min at some value less than  $-2$  and an absolute max greater than  $2$ . At these points,  $F'$  must vanish, and it only remains to show these are the only zeros of  $F'$  on  $[0, \pi]$ . In fact, we see that

$$\frac{1}{3} F'(\bar{\phi}) = \sin^3 \bar{\phi} + A \sin(2\bar{\phi} - B)$$

for some quantities  $A > 0$  and  $B$  independent of  $\bar{\phi}$ . The fact that  $F'(0) < 0$  tells us that we may assume  $0 < B < \pi$ . Clearly, since  $0 \leq \bar{\phi} \leq \pi$ , we have  $\sin^3 \bar{\phi} \geq 0$  and there can be no zero of  $F'$  on the interval  $[B/2, \pi/2 + B/2]$ . For the rest, we consider two cases.

**Case I:**  $0 < B \leq \pi/2$ , i.e.,  $F''(0) \geq 0$ . In this case, both terms in the expression for  $F'$  are increasing on the interval  $0 < \bar{\phi} < B/2$ , so  $F'$  can have at most one zero there. (And since  $F'(B/2) > 0$  it does have exactly one.)

$F'$  must also have a zero on  $[\pi/2 + B/2, \pi]$ . We note that

$$\begin{aligned} \frac{1}{3} F''(\bar{\phi}) &= 3 \cos \bar{\phi} \sin^2 \bar{\phi} + 2A \cos(2\bar{\phi} - B) \\ &= \frac{3}{2} \sin(2\bar{\phi}) \sin \bar{\phi} + 2A \cos(2\bar{\phi} - B) \end{aligned}$$

and consider two subcases, depending on the sign of  $A - \sin^3(3\pi/4 + B/2)$ .

- First assume that  $A \geq \sin^3(3\pi/4 + B/2)$ .

Since  $F'(\pi/2 + B/2)/3 = \sin^3(\pi/2 + B/2) + A > 0$ , and  $F'(3\pi/4 + B/2)/3 = \sin^3(3\pi/4 + B/2) - A \leq 0$ , there is some zero of  $F'$  on the interval

$$(\pi/2 + B/2, 3\pi/4 + B/2].$$

Since both  $\sin^3 \bar{\phi}$  and  $A \sin(2\bar{\phi} - B)$  are decreasing on this interval, there is exactly one zero of  $F'$  there.

Let us assume there is another zero  $\bar{\phi}_0$  of  $F'$  with  $3\pi/4 + B/2 < \bar{\phi}_0 < \pi$ . Since  $F''(3\pi/4 + B/2) < 0$  and  $F'(3\pi/4 + B/2) \leq 0$ , we conclude that  $F''$  must have a zero  $\bar{\phi}_1$  on the interval  $(3\pi/4 + B/2, \bar{\phi}_0)$  at a negative local minimum of  $F'$ . Furthermore, since  $F'(\pi) < 0$ , it must be the case that  $F''$  has another zero  $\bar{\phi}_2$  on the interval  $(\bar{\phi}_1, \pi)$  at a nonnegative local maximum of  $F'$ .

We now show this situation leads to a contradiction by establishing that  $F''$  has exactly one zero on the interval  $(3\pi/4 + B/2, \pi]$ . In fact, we will show more:

**Lemma 5.** *If  $0 < B < \pi/2$ , then  $F''$  has exactly one zero in  $[\pi/2 + B/2, \pi]$ , and it occurs on the interval  $(3\pi/4 + B/2, \pi)$  at a local minimum of  $F'$ .*

This is because  $3 \sin^2 \bar{\phi} \cos \bar{\phi}$  is increasing on the interval  $(\pi - \arccos \frac{1}{\sqrt{3}}, \pi)$ . Indeed,

$$\frac{d}{d\bar{\phi}}(\sin^2 \bar{\phi} \cos \bar{\phi}) = 2 \sin \bar{\phi} \cos^2 \bar{\phi} - \sin^3 \bar{\phi} = \sin \bar{\phi}(3 \cos^2 \bar{\phi} - 1).$$

Furthermore, it is easily checked that  $\pi - \arccos(1/\sqrt{3}) < 3\pi/4$ . Thus,  $F''$  is increasing on the interval  $[3\pi/4 + B/2, \pi]$  and has exactly one zero there. Finally,  $F''$  is negative on the interval  $[\pi/2 + B/2, 3\pi/4 + B/2]$ , so we have established the lemma and finished this subcase.

• Still under the assumption  $0 < B \leq \pi/2$  (Case I), we now suppose instead that  $A < \sin^3(3\pi/4 + B/2)$ .

In this case  $F'$  is positive throughout the interval  $[\pi/2 + B/2, 3\pi/4 + B/2]$ . Thus, the first zero  $\bar{\phi}_0$  of  $F'$  on  $[\pi/2 + B/2, \pi]$  must occur inside  $(3\pi/4 + B/2, \pi)$ . Since  $F'(\pi) < 0$ , and  $F''(\pi) > 0$ , the unique zero  $\phi_1$  of  $F''$  given by Lemma 5 must satisfy

$$\max\{\bar{\phi}_0, 3\pi/4 + B/2\} < \bar{\phi}_1.$$

If we assume the existence of a second zero of  $F'$  on the interval  $(\bar{\phi}_0, \pi)$ , we obtain a zero of  $F''$  at a local maximum of  $F'$  (and a contradiction) as before.

**Case II:**  $\pi/2 \leq B < \pi$ , i.e.,  $F''(0) \leq 0$ . The reflection  $\bar{\phi} \rightarrow \pi - \bar{\phi}$  transforms this case into the first one with  $B \rightarrow \pi - B$ .  $\square$

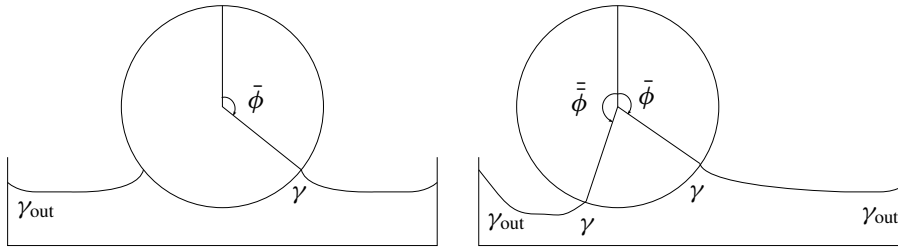
The reader will have no trouble verifying that under Archimedes' assumptions  $\bar{H} = 0$  (a planar interface) and  $\bar{\phi} = \gamma$  (the appropriate azimuthal angle for a horizontal plane to meet the sphere at the correct contact angle) the formula in Theorem 2 reduces to the condition of Archimedes.

#### 4. Floating in two dimensions

The result of [Finn 2011] referred to in the introduction and termed by the author a “criterion for floating” concerns a variational problem considered earlier in [Bhatnagar and Finn 2006] for the energy

$$(17) \quad \mathcal{E} = -\sigma|\hat{\Lambda}| - \sigma\beta|\mathcal{W}| + \mathcal{G},$$

where  $\hat{\Lambda}$  is the linear segment of intersection of a planar/linear interface with a two-dimensional convex body and  $\mathcal{G}$  is the specific gravitational energy we have considered above. The measures appearing in the first two terms in this functional are one-dimensional (length) and the integral is an area integral. There is no volume constraint in Bhatnagar and Finn's problem, nor outer container. With certain other



**Figure 6.** Azimuthal angles determined by a horizontal contact line (left) and differing azimuthal angles in the two-dimensional case (right).

assumptions, they also find that the interface always lies along a fixed line. From this point, Finn goes on to obtain the striking result that for some values of  $\rho > \rho_l$ ,  $\sigma$  and  $\beta$  there will be an equilibrium which is a local minimum for energy in which the convex body contacts the interface, i.e., floats. We now formulate and extend our results to a problem dimensionally similar to the problem of Bhatnagar and Finn.

Physically, we envision a trough consisting of two vertical walls and a horizontal bottom. The trough is assumed to extend infinitely in the  $y = x_2$  direction and to be filled with a sea of liquid. Into this sea is introduced a horizontal floating circular cylinder (an infinitely long log) with axis parallel to  $e_2$ . Let us assume that the free surface interface  $\Lambda$  also is always of cylindrical form with generator parallel to  $e_2$ , so that if the log is centrally located between the walls and the interface shares the same midplane symmetry, then the projection of the system onto the  $x, z$ -plane resembles that of the system considered in the previous section (Figure 6, left), though the equation satisfied by the generating curve (and hence its shape) will be different from that of the meridian previously considered.

The energy of such a system can be taken to have the form of (5):

$$\mathcal{E} = \sigma|\Lambda| - \sigma\beta|\mathcal{W}| + \mathcal{G},$$

where the dimensions of the measures have been lowered by one and  $\mathcal{G} = \int_{\mathcal{V} \cup \Sigma_m} G$  is an area integral. The first-order necessary conditions take the form

$$\begin{aligned} k &= G/\sigma - \lambda && \text{on the curve } \Lambda, \\ \cos \gamma &= \beta && \text{at the endpoints of } \Lambda, \end{aligned}$$

and

$$(18) \quad \dot{w} \cdot \vec{n} \Big|_{\partial \Lambda} + \int_{\mathcal{W}_m} (G/\sigma - \lambda) \dot{w} \cdot N^{\mathcal{W}} + \int_{\partial \Sigma_m} (G_m/\sigma) \dot{w} \cdot N^m = 0,$$

where  $k$  is the curvature of  $\Lambda$  and  $\lambda$  arises from an area constraint on the cross section of liquid in the trough. In analogy to the three-dimensional case, we assume

an area density  $\rho$  for the object, that the object floats in a liquid of area density  $\rho_l$ , a capillary constant  $\kappa = \rho_l g / \sigma$ , and that the radius of the log is  $a$ .

Before we begin an analysis of this variational problem in earnest, let us pause to note what Archimedes' principle would state in this lower-dimensional case (because it will appear in a surprising way later):

**Theorem 6.** *According to Archimedes' principle in one lower dimension, a homogeneous disk/log of density  $\rho > \rho_l$  will sink to the bottom of a bath of density  $\rho_l$ , and a homogeneous disk/log of density  $\rho < \rho_l$  will float at a level determined by*

$$(19) \quad 2\bar{\phi} - \sin(2\bar{\phi}) = 2\pi \left(1 - \frac{\rho}{\rho_l}\right).$$

We assume initially the contact line (i.e., the two points where  $\Lambda$  meets  $\Sigma_m$ ) is determined by two azimuthal angles, one  $\bar{\phi}$  as before and a second  $\bar{\bar{\phi}}$  measured counterclockwise from the vertical  $\mathbf{e}_3$ ; see Figure 6, right. In addition to (15), the following identities have been found useful.

$$(20) \quad \begin{aligned} N^m[\phi] &= -\sin \bar{\bar{\phi}} \mathbf{e}_1 + \cos \bar{\bar{\phi}} \mathbf{e}_3, \\ N^{\mathcal{W}}[\phi] &= -N^m \\ &= \sin \bar{\bar{\phi}} \mathbf{e}_1 - \cos \bar{\bar{\phi}} \mathbf{e}_3, \\ \vec{v}[\phi] &= (N^{\mathcal{W}})^\perp \\ &= \cos \bar{\bar{\phi}} \mathbf{e}_1 + \sin \bar{\bar{\phi}} \mathbf{e}_3, \\ \vec{n} &= \cos \gamma \vec{v} + \sin \gamma N^{\mathcal{W}} \\ &= \cos(\bar{\bar{\phi}} - \gamma) \mathbf{e}_1 + \sin(\bar{\bar{\phi}} - \gamma) \mathbf{e}_3, \\ N^\Lambda &= (\vec{n})^\perp \\ &= -\sin(\bar{\bar{\phi}} - \gamma) \mathbf{e}_1 + \cos(\bar{\bar{\phi}} - \gamma) \mathbf{e}_3. \end{aligned}$$

Taking first a horizontal motion of the floating sphere, so that  $\dot{w} = \mathbf{e}_1$ , we find

$$\mathbf{e}_1 \cdot \vec{n} \Big|_{\partial \mathcal{W}_m} = \cos(\bar{\bar{\phi}} - \gamma) - \cos(\bar{\phi} - \gamma) = -2 \sin B \sin(A - \gamma),$$

where  $A = \frac{1}{2}(\bar{\bar{\phi}} + \bar{\phi})$ ,  $B = \frac{1}{2}(\bar{\bar{\phi}} - \bar{\phi})$ ,

$$\begin{aligned} \int_{\mathcal{W}_m} (\kappa z - \lambda) \mathbf{e}_1 \cdot N^{\mathcal{W}} &= a(\kappa d - \lambda)(\cos \bar{\bar{\phi}} - \cos \bar{\phi}) + \frac{1}{2} \kappa a^2 (\cos^2 \bar{\bar{\phi}} - \cos^2 \bar{\phi}) \\ &= -2a \sin B \sin A (\kappa d - \lambda + \kappa a \cos A \cos B), \end{aligned}$$

and

$$\int_{\partial \Sigma_m} \left( \kappa \frac{\rho}{\rho_l} z - \lambda \right) \mathbf{e}_1 \cdot N^m = 0.$$

Since each of these terms has a factor  $\sin B$ , we see from condition (18), that one possibility is  $\sin B = 0$ . If this holds, it can readily be determined that  $\bar{\bar{\phi}} = \bar{\phi}$ .

Once this occurs, then since the left and right interfaces must start from the same height and with the same inclination angle, we have a proof that the axis of the floating cylinder must lie on the midplane between the vertical walls. This is the conclusion we would like to make. The other alternative is that

$$\sin(A - \gamma) + a \sin A(\kappa d - \lambda + \kappa a \cos A \cos B) = 0,$$

which we rewrite as

$$(21) \quad (\cos \gamma + a(\kappa d - \lambda)) \sin A + \frac{1}{2} \kappa a^2 \sin(2A) \cos B - \sin \gamma \cos A = 0.$$

Leaving this open as a possibility for the moment, we turn to an independent vertical translation of  $\Sigma_m$  with  $\dot{w} = e_3$ . In this case

$$e_3 \cdot \vec{n} \Big|_{\partial \mathcal{W}_m} = \sin(\bar{\phi} - \gamma) + \sin(\bar{\phi} - \gamma) = 2 \cos B \sin(A - \gamma);$$

moreover

$$\begin{aligned} \int_{\mathcal{W}_m} (\kappa z - \lambda) e_3 \cdot N^{\mathcal{W}} \\ &= a(\kappa d - \lambda)(\sin \bar{\phi} + \sin \bar{\phi}) + \frac{1}{4} \kappa a^2 (\sin(2\bar{\phi}) + \sin(2\bar{\phi})) + \frac{1}{2} \kappa a^2 (\bar{\phi} + \bar{\phi}) - \kappa a^2 \pi \\ &= 2a \cos B \sin A(\kappa d - \lambda) + \frac{1}{2} \kappa a^2 \sin(2A) \cos(2B) + \frac{1}{2} \kappa a^2 (\bar{\phi} + \bar{\phi}) - \kappa a^2 \pi \end{aligned}$$

and

$$\int_{\partial \Sigma_m} \kappa \frac{\rho}{\rho_l} z e_3 \cdot N^m = \kappa a^2 \pi \frac{\rho}{\rho_l}.$$

Combining these terms to form the expression in (18), we arrive at a second necessary condition,

$$(22) \quad (\cos \gamma + a(\kappa d - \lambda)) \sin A \cos B - \sin \gamma \cos A \cos B \\ + \frac{1}{2} \kappa a^2 \sin A \cos A (1 - 2 \sin^2 B) + \frac{1}{4} \kappa a^2 (\bar{\phi} + \bar{\phi}) = \frac{\kappa a^2 \pi}{2} \left(1 - \frac{\rho}{\rho_l}\right).$$

Multiplying the equation in (21) by  $\cos B$  and subtracting the result from (22) and simplifying, we obtain the surprising condition

$$(23) \quad \bar{\phi} + \bar{\phi} - \sin(\bar{\phi} + \bar{\phi}) = 2\pi \left(1 - \frac{\rho}{\rho_l}\right).$$

This is surprising because it says that if the log floats anywhere but in the middle between the vertical walls of the trough, then the wetted region must match the wetted region predicted by (19) of Theorem 6, which is based on Archimedes' assumptions, including that of a flat interface. In particular, the portion that is wetted is independent of all parameters except the density fraction! We view this scenario as highly unlikely. The fact that we cannot rule out this possibility leads to the following curious result.



**Theorem 7.** *In the two-dimensional floating log problem, either the axis of the log lies in the vertical midplane determined by the sides of the vessel, or the wetted/nonwetted region is determined by the generalized version of Archimedes' condition given in (23).*

At this point, we proceed as in the three-dimensional case by assuming symmetry of the interface with respect to the midplane. When  $\bar{\phi} = \bar{\phi}$ , condition (22) associated with the vertical translation is still nonvacuous and becomes

$$F(\bar{\phi}) = 2\bar{\phi} + \sin(2\bar{\phi}) + \frac{4}{\kappa a^2} \sin(\bar{\phi} - \gamma) + \frac{4}{\kappa a} (\kappa d - \lambda) \sin \bar{\phi} = 2\pi \left(1 - \frac{\rho}{\rho_l}\right).$$

Again following the three-dimensional case, we let

$$\bar{k} = \kappa(d + a \cos \bar{\phi}) - \lambda$$

denote the curvature of the interface at the contact line on the object. Substitution yields

**Theorem 8.** *A log that floats in a centrally symmetric position under the effects of surface tension and adhesion must float at a level determined by the azimuthal angle  $\bar{\phi}$  satisfying*

$$(24) \quad 2\bar{\phi} - \sin(2\bar{\phi}) + \frac{4}{\kappa a^2} \sin(\bar{\phi} - \gamma) + \frac{4\bar{k}}{\kappa a} \sin \bar{\phi} = 2\pi \left(1 - \frac{\rho}{\rho_l}\right),$$

where  $\bar{k}$  is the curvature of the interface at the contact line, and  $\gamma$  is the contact angle of the interface with the floating log.

We emphasize that  $\bar{k}$  is assumed to be given and constant. The behavior of the function

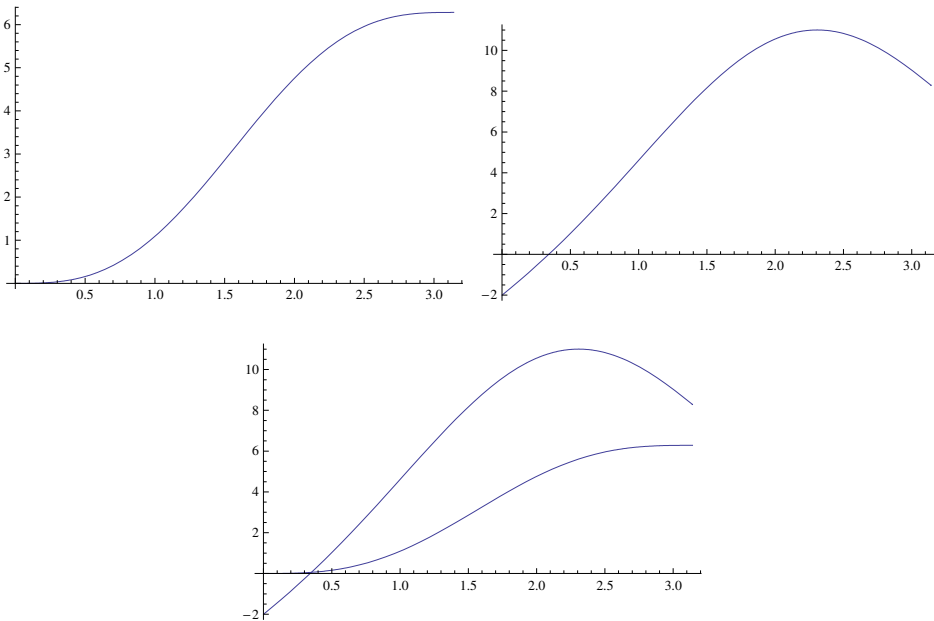
$$F(\bar{\phi}) = 2\bar{\phi} - \sin(2\bar{\phi}) + \frac{4}{\kappa a^2} \sin(\bar{\phi} - \gamma) + \frac{4\bar{k}}{\kappa a} \sin \bar{\phi}$$

is somewhat different than that in the three-dimensional case; see Figure 7. One sees first of all that

$$F(0) = -\frac{4}{\kappa a^2} \sin \gamma < 0 \quad \text{and} \quad F(\pi) = 2\pi + \frac{4}{\kappa a^2} \sin \gamma > 2\pi.$$

Thus, the endpoint values do not coincide with the extremes of the expression on the right in (24) associated with  $\rho = 0$  and  $\rho = \rho_l$ . Nevertheless, the interval between 0 and  $2\pi$  is clearly covered by the values of  $F(\bar{\phi})$  and, in fact, each value is taken exactly once. To see this we compute

$$\frac{F'(\bar{\phi})}{2} = 1 - \cos(2\bar{\phi}) + \frac{2}{\kappa a^2} \cos(\bar{\phi} - \gamma) + \frac{2\bar{k}}{\kappa a} \cos \bar{\phi}$$



**Figure 7.** The azimuthal angles determined by Theorems 6 (top left) and 8 (top right); plotted together on the bottom.

and observe first that

$$\frac{F'(0)}{2} = \frac{2}{\kappa a^2} \cos(\gamma) + \frac{2\bar{k}}{\kappa a} = -\frac{F'(\pi)}{2}.$$

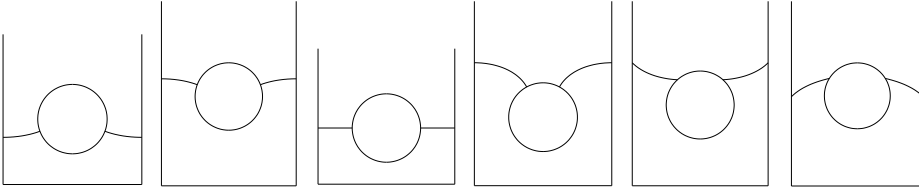
It follows that  $F'$  is nonpositive at one of the endpoints and has the opposite sign at the other. Using this, reasoning similar to that found in Section 3 shows  $F'$  can have at most one zero on  $[0, \pi]$ .

Thus, some salient features of Theorem 2 hold also in this lower-dimensional case. For fixed  $\bar{k}$  and  $\gamma$ , if  $\rho \leq \rho_l$ , there is a unique height at which the disk/log can float; there is an interval  $\rho_l < \rho < \rho_{\max}$  on which there is at least one (and sometimes two) possible heights at which floating can occur. One expects that if two azimuthal angles are determined by (24), the larger is the physically relevant one.

## 5. Global solutions numerically computed

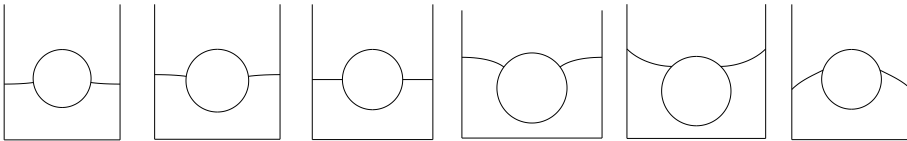
We have obtained global configurations of floating numerically for the problems considered above both in two and three dimensions. The stability and uniqueness of most of these configurations is not presently known.

In Figures 1 and 2 we give representative global configurations which have been obtained and a list of the relevant parameters.



	$\rho/\rho_l$	$\gamma$	$\gamma_{out}$	$d$	$\lambda$	$\bar{\phi}$
(1) Lightest	0.0	$\pi/2$	$\pi/2$	1.8850	1.0860	1.9284
(2) Heavy	1.0	$\pi/2$	$\pi/2$	2.6504	3.4494	1.2132
(3) Flat	0.5	$\pi/2$	$\pi/2$	1.6427	1.6427	1.5708
(4) Denser	1.6	$\pi/2$	$\pi/2$	1.9934	3.9663	0.4973
(5) Unstable (?)	0.5	$\pi/4$	$\pi/4$	2.3777	2.7878	0.7328
(6) Stable (?)	0.5	$\pi/4$	$3\pi/4$	2.7382	3.4704	1.0150

**Table 1.** Two-dimensional case (floating logs). Parameters for each configuration on the top, from left to right. In all cases  $a = 1$ ,  $\kappa = 1$ ,  $R = 2$ , and the cross-sectional area of liquid is 10.



	$\rho/\rho_l$	$\gamma$	$\gamma_{out}$	$d$	$\lambda$	$\bar{\phi}$
(1) Lightest	0.0	$\pi/2$	$\pi/2$	2.1174	1.8486	1.7086
(2) Heavy	1.0	$\pi/2$	$\pi/2$	1.8689	2.1377	1.4330
(3) Flat	0.5	$\pi/2$	$\pi/2$	1.9902	1.9902	1.5708
(4) Denser	2.1	$\pi/2$	$\pi/2$	1.4293	2.5321	0.9293
(5) Unstable (?)	0.5	$\pi/4$	$\pi/4$	1.3646	1.4679	0.7790
(6) Stable (?)	0.5	$\pi/4$	$3\pi/4$	2.0192	2.6403	1.2570

**Table 2.** Three-dimensional case (floating balls). Parameters for each configuration on the top, from left to right. In all cases  $a = 1$ ,  $\kappa = 1$ ,  $R = 2$ , and the volume of liquid is 25.

### 6. Existence and uniqueness

A referee has requested that we provide an existence and uniqueness result for some floating configurations at least superficially like those to which our main result applies. As the referee suggests, we provide in this section an existence result for a ball floating symmetrically in an infinite three-dimensional bath. We also prove

that uniqueness does not hold in that case in general, and provide some remarks suggesting that uniqueness does not hold in the problem we consider either.

This problem has been considered in [Keller 1998; Vella and Mahadevan 2005] though not from a fundamentally variational point of view and with existence (and presumably some statement of uniqueness) assumed. Various aspects of the problem make it fundamentally simpler than the physical problem of floating in a finite container and, as we shall see, we can say much more in this case.

Analogues of the results below are shown numerically in the lower-dimensional case of Bhatnagar and Finn’s problem [2006]. Also, a partial existence result is given in [Finn 2011] in the two-dimensional case and in [Finn and Vogel 2009] in the three-dimensional case. The methods below may be adapted to give versions of our results in this section for the two-dimensional problem.

As is customary for this kind of problem, we assume a prescribed zero level to which our symmetric interface, satisfying the first four requirements of the boundary value problem (14), is asymptotic. The requirement that the interface be asymptotic (to first order) to the zero level plane necessitates the additional conditions

$$\lim_{r \rightarrow \infty} u = \lim_{r \rightarrow \infty} \psi = 0.$$

These conditions along with the third equation in (14) imply that the constant  $\lambda$  is zero. In order to show existence, we must obtain a solution to this system which satisfies the additional requirement of Theorem 2. We stress that our application of Theorem 2 to this situation in which the energies we considered in the proof are infinite is somewhat formal, though under a suitable modification of the energies, it is fairly clear that condition (4) is the correct equilibrium condition for floating in this situation as well. With the aforementioned modifications, our problem becomes one of finding a height  $d$  for the center of the sphere of radius  $a$ , an azimuthal angle  $\bar{\phi}$ , and a meridian  $(r, u)$  with inclination angle  $\psi$  such that

$$(25) \quad \begin{cases} \dot{r} = \cos \psi, \\ \dot{u} = \sin \psi, \\ \dot{\psi} = \kappa u - \sin \psi / r, \\ \psi = \gamma - \bar{\phi} \text{ and } u = d + a \cos \bar{\phi} \text{ when } r = r(0) = a \sin \bar{\phi}, \\ \lim_{r \rightarrow \infty} u = \lim_{r \rightarrow \infty} \psi = 0, \end{cases}$$

and

$$(26) \quad \cos^3 \bar{\phi} - 3 \cos \bar{\phi} + \frac{6}{\kappa a} \left( \bar{H} + \frac{\cos \gamma}{a} \right) \sin^2 \bar{\phi} - \frac{3 \sin \gamma}{\kappa a^2} \sin(2\bar{\phi}) = 2 \left( 1 - \frac{2\rho}{\rho_l} \right),$$

where  $\bar{H} = \kappa(d + a \cos \bar{\phi})/2$ .

It has been shown by Elcrat, Neel, and Siegel [Elcrat et al. 2004b] that given any  $\bar{r} = a \sin \bar{\phi} > 0$  and any inclination angle  $\bar{\psi} = \gamma - \bar{\phi}$ , there is a unique solution  $(r, u)$  of the system (25) *except* for the condition  $u = d + a \cos \bar{\phi}$  on the contact line. Since  $d$  has not been specified, we can obviously take the sphere of center height  $d = u(0) - a \cos \bar{\phi}$  to get this condition as well. In this way, everything becomes a function of  $\bar{\phi}$ , and we have only to find  $\bar{\phi}$  satisfying the following simplified version of (26):

$$(27) \quad \cos^3 \bar{\phi} - 3 \cos \bar{\phi} + \frac{3}{a} \left( u(0) + \frac{2 \cos \gamma}{\kappa a} \right) \sin^2 \bar{\phi} - \frac{3 \sin \gamma}{\kappa a^2} \sin(2\bar{\phi}) = 2 \left( 1 - \frac{2\rho}{\rho_l} \right).$$

Unfortunately, the dependence of  $u(0) = u(0; \bar{\phi})$  on  $\bar{\phi}$  is not explicit and not well understood. This fact prevents us from giving a full analysis of the solutions of (27). Nevertheless, we can set

$$G(\bar{\phi}) = \cos^3 \bar{\phi} - 3 \cos \bar{\phi} + \frac{3}{a} \left( u(0) + \frac{2 \cos \gamma}{\kappa a} \right) \sin^2 \bar{\phi} - \frac{3 \sin \gamma}{\kappa a^2} \sin(2\bar{\phi}),$$

which is a well defined smooth function of  $\bar{\phi}$ .

When  $\bar{\phi}$  tends to zero (a sinking ball), we have that  $r = a \sin \bar{\phi}$  tends to zero and necessarily  $u(0; \bar{\phi})$  tends to zero as well. Thus,

$$\lim_{\bar{\phi} \rightarrow 0} G(\bar{\phi}) = -2.$$

Similarly,

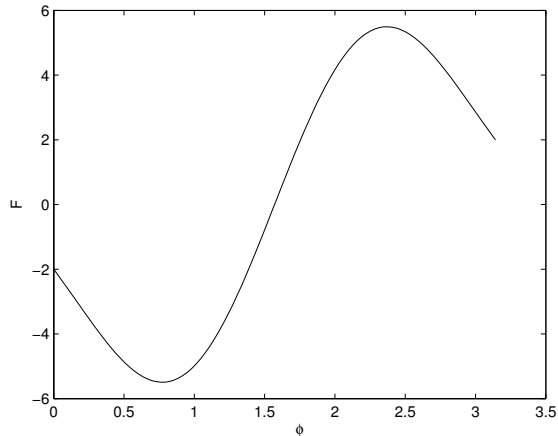
$$\lim_{\bar{\phi} \rightarrow \pi} G(\bar{\phi}) = 2.$$

We draw attention to the fact that these values are shared by the function  $F$  considered in Section 3. In fact, we can numerically graph the function  $G$  for specific choices of  $\kappa$  and  $\gamma$  to see that  $G$  shares the qualitative properties of the function  $F$  analyzed in Section 3, initially decreasing to a unique minimum, then increasing to a unique maximum greater than 2, and decreasing on the remainder of the interval; see Figure 8. We expect that these qualitative features are always shared, but we are unable to prove that.

We are able to compute the following:

$$\lim_{\bar{\phi} \rightarrow 0, \pi} G'(\bar{\phi}) = -\frac{6 \sin \gamma}{\kappa a^2}.$$

This means that  $G$  is always decreasing at  $\bar{\phi} = 0$  and  $\bar{\phi} = \pi$ . By continuity we obviously have enough to obtain existence for any density  $\rho$  between zero and the density of the liquid  $\rho_l$ . The last computation also gives existence for some range of densities greater than  $\rho_l$  and some “negative densities” as described in the discussion of the main result in Section 1. We have shown the following:



**Figure 8.** Numerical plot of the function  $G$  for  $a = 1$ ,  $\kappa = 1$  and  $\gamma = \pi/2$ .

**Theorem 9.** *There are positive numbers  $\epsilon$  and  $\delta$  depending on the capillarity constant  $\kappa$ , the radius of the sphere  $a$ , the contact angle  $\gamma$ , and the density of the liquid  $\rho_l$ , such that the floating ball problem for an infinite bath has a well defined equilibrium configuration (satisfying the flux condition obtained in this paper) for each density  $\rho$  with  $-\epsilon < \rho < \rho_l + \delta$ .*

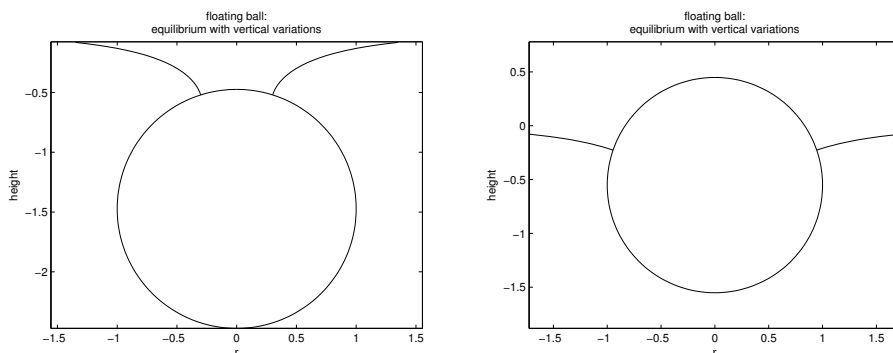
It follows also that there is some  $\bar{\phi} = \bar{\phi}_{\min}$  where  $G$  takes a minimum value  $m < -2$ . If we take a density  $\rho$  with

$$\rho_l < \rho < \rho_l(1 - m/2),$$

then we see there are at least two values  $\bar{\phi}_1$  and  $\bar{\phi}_2$  with  $\bar{\phi}_1 < \bar{\phi}_{\min} < \bar{\phi}_2$  which correspond to distinct equilibrium configurations for different heights  $d$  of the ball.

**Theorem 10.** *The problem of a floating ball in an infinite bath with capillarity taken into account and  $\gamma \in (0, \pi)$  does not have a unique equilibrium solution in general. More precisely, there is an interval  $(\rho_m, \rho_M) \supset \supset (0, \rho_l)$  and for any density  $\rho$  in  $(\rho_m, \rho_M) \setminus (0, \rho_l)$ , there exist at least two equilibria.*

If we calculate a modified energy for the specific choices of parameters considered above in Figure 8 and the two distinct equilibria shown in Figure 9, we find that the one of smaller azimuthal angle and lower height  $d$  has greater energy. This strongly suggests that when a heavy ball is floating, the energy increases to a maximum (at another equilibrium) as the ball is pushed down. After the ball is pushed below the second equilibrium height (maximum energy), it will sink. These qualitative observations are consistent with experiments.



**Figure 9.** Distinct equilibria showing nonuniqueness for  $\rho = 3/2 > \rho_l = 1$ ,  $a = 1$ ,  $\kappa = 1$  and  $\gamma = \pi/2$ .

Comparison to the graphs shown in Figure 7, suggests that the same situation holds in finite containers. It should be noted, however, that Figure 7 does not show this is the case, because  $\bar{H}$  is considered constant there, and the value of  $\bar{H}$  will undoubtedly be different in the two distinct equilibrium configurations.

### Acknowledgment

We thank the referees, the editor, and the copy editor for helpful suggestions on both the content and exposition of our work.

### References

- [Archimedes/Heath 1897] E. by T. L. Heath (editor), *The works of Archimedes*, Cambridge University Press, 1897. Reprinted Dover, Mineola, NY, 2002. MR 2005a:01003
- [Bhatnagar and Finn 2006] R. Bhatnagar and R. Finn, "Equilibrium configurations of an infinite cylinder in an unbounded fluid", *Phys. Fluids* **18**:4 (2006), article ID 047103. MR 2007f:76032
- [Elcrat et al. 2004a] A. Elcrat, T.-E. Kim, and R. Treinen, "Annular capillary surfaces", *Arch. Math. (Basel)* **82**:5 (2004), 449–467. MR 2006e:76055 Zbl 1136.76353
- [Elcrat et al. 2004b] A. Elcrat, R. Neel, and D. Siegel, "Equilibrium configurations for a floating drop", *J. Math. Fluid Mech.* **6**:4 (2004), 405–429. MR 2005j:76017 Zbl 1064.76019
- [Finn 1986] R. Finn, *Equilibrium capillary surfaces*, Grundlehren der Math. Wiss. **284**, Springer, New York, 1986. MR 88f:49001 Zbl 0583.35002
- [Finn 2005] R. Finn, "Floating and partly immersed balls in a weightless environment", *Funct. Differ. Equ.* **12**:1-2 (2005), 167–173. MR 2005m:76072 Zbl 1085.76016
- [Finn 2011] R. Finn, "Criteria for floating, I", *J. Math. Fluid Mech.* **13**:1 (2011), 103–115. MR 2012j:76030
- [Finn and Vogel 2009] R. Finn and T. I. Vogel, "Floating criteria in three dimensions", *Analysis (Munich)* **29**:4 (2009), 387–402. MR 2011a:76020 Zbl 05657214
- [Graf 2004] E. Graf, "Just what *did* Archimedes say about buoyancy?", *Phys. Teach.* **42**:5 (2004), 296–299.

- [Johnson and Perko 1968] W. E. Johnson and L. M. Perko, “Interior and exterior boundary value problems from the theory of the capillary tube”, *Arch. Rational Mech. Anal.* **29** (1968), 125–143. MR 36 #6686 Zbl 0162.57002
- [Keller 1998] J. B. Keller, “Surface tension force on a partly submerged body”, *Phys. Fluids* **10**:11 (1998), 3009–3010. MR 99h:76033 Zbl 1185.76546
- [McCuan 2007] J. McCuan, “A variational formula for floating bodies”, *Pacific J. Math.* **231**:1 (2007), 167–191. MR 2008f:76035 Zbl 1148.76012
- [Nickolov 2002] R. Nickolov, “Uniqueness of the singular solution to the capillary equation”, *Indiana Univ. Math. J.* **51**:1 (2002), 127–169. MR 2003e:35096 Zbl 1042.34015
- [Ros 1996] A. Ros, “Embedded minimal surfaces: forces, topology and symmetries”, *Calc. Var. Partial Differential Equations* **4**:5 (1996), 469–496. MR 98c:53010 Zbl 0861.53008
- [Serrin 1959] J. Serrin, “Mathematical principles of classical fluid mechanics”, pp. 125–263 in *Handbuch der Physik*, vol. 8/1, edited by S. Flügge, Springer-Verlag, Berlin, 1959. MR 21 #6836b
- [Siegel 1980] D. Siegel, “Height estimates for capillary surfaces”, *Pacific J. Math.* **88**:2 (1980), 471–515. MR 82h:35037 Zbl 0411.35043
- [Siegel 2006] D. Siegel, “Approximating symmetric capillary surfaces”, *Pacific J. Math.* **224**:2 (2006), 355–365. MR 2007a:76022 Zbl 1116.76015
- [Treinen 2012] R. Treinen, “Extended annular capillary surfaces”, *J. Math. Fluid Mech.* **14**:4 (2012), 619–632. MR 2992032 Zbl 1254.76040
- [Turkington 1980] B. Turkington, “Height estimates for exterior problems of capillarity type”, *Pacific J. Math.* **88**:2 (1980), 517–540. MR 82i:35070 Zbl 0474.76012
- [Vella and Mahadevan 2005] D. Vella and L. Mahadevan, “The ‘Cheerios effect’”, *Am. J. Phys.* **73**:9 (2005), 817–825.
- [Vogel 1982] T. I. Vogel, “Symmetric unbounded liquid bridges”, *Pacific J. Math.* **103**:1 (1982), 205–241. MR 84f:53007 Zbl 0504.76025

Received January 15, 2009. Revised April 25, 2013.

JOHN MCCUAN  
SCHOOL OF MATHEMATICS  
GEORGIA TECH  
686 CHERRY STREET  
ATLANTA, GA 30332  
UNITED STATES  
mccuan@math.gatech.edu

RAY TREINEN  
MATHEMATICS DEPARTMENT  
TEXAS STATE UNIVERSITY  
601 UNIVERSITY DRIVE  
SAN MARCOS, TX 78666-4616  
UNITED STATES  
rt30@txstate.edu



## GENERALIZED EIGENVALUE PROBLEMS OF NONHOMOGENEOUS ELLIPTIC OPERATORS AND THEIR APPLICATION

DUMITRU MOTREANU AND MIEKO TANAKA

We consider the equation  $-\operatorname{div}(a(x, |\nabla u|) \nabla u) = \lambda |u|^{p-2} u$  (whose special case  $a(x, t) = t^{p-2}$  is the  $p$ -Laplace equation) on a bounded domain  $\Omega \subset \mathbb{R}^N$  with  $C^2$  boundary, with null boundary condition. We prove that there are  $\lambda \in \mathbb{R}$  for which the equation has a nontrivial solution. As an application, by variational methods, we present the existence of a positive solution to  $-\operatorname{div}(a(x, |\nabla u|) \nabla u) = f(x, u)$  in  $\Omega$ , where  $f$  is asymptotically  $(p-1)$ -linear near zero and  $\infty$ , considering the nonresonant, resonant, and doubly resonant cases. We show that, generally, the spectrum of the operator  $-\operatorname{div}(a(x, |\nabla u|) \nabla u)$  on  $W_0^{1,p}(\Omega)$  is not discrete.

### 1. Introduction

Let  $1 < p < \infty$  and let  $\Omega \subset \mathbb{R}^N$  be a bounded domain with  $C^2$  boundary  $\partial\Omega$ . We are interested in values of  $\lambda \in \mathbb{R}$  such that a nontrivial solution exists to the equation

$$(EV; \lambda) \quad \begin{cases} -\operatorname{div} A(x, \nabla u) = \lambda |u|^{p-2} u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega; \end{cases}$$

such a  $\lambda$  is called an *eigenvalue* for  $A$ . Here  $A: \bar{\Omega} \times \mathbb{R}^N \rightarrow \mathbb{R}^N$  is a map that is strictly monotone in the second variable and satisfies the regularity conditions in Assumption A below.

The  $p$ -Laplace equation is the special case of  $(EV; \lambda)$  with  $A(x, y) = |y|^{p-2} y$ , and in this case the eigenvalues for  $A$  are the usual eigenvalues of the  $p$ -Laplacian. However, we do not suppose that  $A$  is  $(p-1)$ -homogeneous in the second variable. Instead, these are the assumptions we make on the map  $A$ :

**Assumption A.**  $A(x, y) = a(x, |y|)y$ , where  $a(x, t) > 0$  for all  $x \in \bar{\Omega}$  and all  $t \in (0, +\infty)$ ; furthermore:

- (i)  $A \in C^0(\bar{\Omega} \times \mathbb{R}^N, \mathbb{R}^N) \cap C^1(\bar{\Omega} \times (\mathbb{R}^N \setminus \{0\}), \mathbb{R}^N)$ .

*MSC2010:* 35P30, 35J62, 49R05.

*Keywords:* quasilinear elliptic equations, nonhomogeneous operators, nonlinear eigenvalue problems, positive solutions, mountain pass theorem.

(ii) There exists  $C_1 > 0$  such that

$$|D_y A(x, y)| \leq C_1 |y|^{p-2} \quad \text{for every } x \in \bar{\Omega} \text{ and } y \in \mathbb{R}^N \setminus \{0\}.$$

(iii) There exists  $C_0 > 0$  such that

$$D_y A(x, y) \xi \cdot \xi \geq C_0 |y|^{p-2} |\xi|^2 \quad \text{for every } x \in \bar{\Omega}, y \in \mathbb{R}^N \setminus \{0\} \text{ and } \xi \in \mathbb{R}^N;$$

(iv) there exists  $C_2 > 0$  such that

$$|D_x A(x, y)| \leq C_2 (1 + |y|^{p-1}) \quad \text{for every } x \in \bar{\Omega} \text{ and } y \in \mathbb{R}^N \setminus \{0\}.$$

(v) There exist  $C_3 > 0$  and a positive  $t_0 \leq 1$  such that

$$|D_x A(x, y)| \leq C_3 |y|^{p-1} (-\log |y|)$$

for every  $x \in \bar{\Omega}$ ,  $y \in \mathbb{R}^N$  with  $0 < |y| < t_0$ .

From now on, we assume that  $C_0 \leq p - 1 \leq C_1$  which leads to no loss of generality, as can be seen from Assumption A(ii)–(iii).

A similar hypothesis to Assumption A is considered in the study of quasi-linear elliptic problems; see [Motreanu and Papageorgiou 2011, Example 2.2; Damascelli 1998; Motreanu et al. 2011; Miyajima et al. 2012; Tanaka 2012a]. We also refer to [García-Huidobro et al. 1995; Kim 2009; Kim and Kim 2010; Fukagai and Narukawa 2007; Prado and Ubilla 1998; Robinson 2004] for generalized  $p$ -Laplace operators. In particular, when  $A(x, y) = |y|^{p-2}y$  — that is, when  $\operatorname{div} A(x, \nabla u)$  is the usual  $p$ -Laplacian  $\Delta_p u$  — we can take  $C_0 = C_1 = p - 1$  in Assumption A. Conversely, if  $C_0 = C_1 = p - 1$  in Assumption A, the inequalities in Remark 1(ii)–(iii) below show that  $a(x, t) = |t|^{p-2}$ , whence  $A(x, y) = |y|^{p-2}y$ . In the  $p$ -Laplace case, the first eigenvalue  $\lambda_1$  is obtained by the Rayleigh quotient:  $\lambda_1 = \inf\{\int_{\Omega} |\nabla u|^p dx / \|u\|_p^p : u \neq 0\}$ . But since our operator is nonhomogeneous,  $\inf\{\lambda \in \mathbb{R} : \lambda \text{ is an eigenvalue of } A\}$  is in general not obtained by such a Rayleigh quotient corresponding to  $A$ . In Section 3, since the Rayleigh quotient plays an important role, we study its behavior as  $\|u\|_p \rightarrow 0$  or  $\|u\|_p \rightarrow \infty$  under an additional condition describing an asymptotic  $(p-1)$ -homogeneity. For example, we can consider

$$\operatorname{div} A(x, \nabla u) = \operatorname{div}((a_0(x)|\nabla u|^{p-2} + a_{\infty}(x)|\nabla u|^{q-2})(1 + |\nabla u|^q)^{(p-q)/q} \nabla u)$$

for  $1 < p \leq q < \infty$ ,  $a_0, a_{\infty} \in C^1(\bar{\Omega})$  with  $\min_{\bar{\Omega}} a_0 > 0$  and  $\min_{\bar{\Omega}} a_{\infty} > 0$ . This satisfies

$$\begin{aligned} A(x, y) - a_0(x)|y|^{p-2}y &= o(|y|^{p-1}) \quad \text{as } |y| \rightarrow 0, \\ A(x, y) - a_{\infty}(x)|y|^{p-2}y &= o(|y|^{p-1}) \quad \text{as } |y| \rightarrow \infty. \end{aligned}$$

Under these these conditions (see (AH0) and (AH) in Section 3), we shall prove

that

$$\min \left\{ \int_{\Omega} \int_0^{|\nabla u(x)|} \frac{a(x, t)t}{r^p} dt dx : \|u\|_p = r \right\}$$

approaches  $\lambda_1(a_0)/p$  as  $r \rightarrow +0$  and  $\lambda_1(a_{\infty})/p$  as  $r \rightarrow +\infty$ ; here

$$\lambda_1(a_0) = \min \left\{ \int_{\Omega} a_0(x) |\nabla u|^p dx : \|u\|_p = 1 \right\},$$

$$\lambda_1(a_{\infty}) = \min \left\{ \int_{\Omega} a_{\infty}(x) |\nabla u|^p dx : \|u\|_p = 1 \right\}.$$

Concerning the eigenvalue problem for a nonhomogeneous operator, we can refer to [Robinson 2004; Tanaka 2012b] under the Neumann boundary condition.

In Section 4, as an application of Section 3, we present the existence of a positive solution for the quasilinear elliptic equation

$$(P) \quad \begin{cases} -\operatorname{div} A(x, \nabla u) = f(x, u) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where  $f$  satisfies the following assumption.

**Assumption (f).**  $f$  is a Carathéodory function on  $\Omega \times \mathbb{R}$  with  $f(x, 0) = 0$  for a.e.  $x \in \Omega$ ,  $f$  is bounded on bounded sets and  $f$  is asymptotically  $(p-1)$ -linear near  $+0$  and  $+\infty$  in the following sense:

- (i)  $\lim_{u \rightarrow +0} \frac{f(x, u)}{u^{p-1}} = \alpha_0$  uniformly in a.e.  $x \in \Omega$ ,
- (ii)  $\lim_{u \rightarrow +\infty} \frac{f(x, u)}{u^{p-1}} = \alpha$  uniformly in a.e.  $x \in \Omega$ ,

for some constants  $\alpha_0$  and  $\alpha$ .

Regarding the existence of a positive solution under the Dirichlet boundary condition, we can refer to [Fukagai and Narukawa 2007; Prado and Ubilla 1998] for nonhomogeneous operators. However, we can not apply these results to our nonlinear term which is only asymptotically  $(p-1)$ -linear near  $+0$  and  $+\infty$ , and furthermore with possibly different weights. In [García-Huidobro et al. 1995], it is proved the existence of a positive radial solution for nonhomogeneous operators.

For the  $p$ -Laplace equation, it is well known that if  $(\alpha - \lambda_1)(\alpha_0 - \lambda_1) < 0$  (where  $\lambda_1$  denotes the first eigenvalue of  $-\Delta_p$  under a Dirichlet boundary condition),

$$-\Delta_p u = f(x, u) \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega,$$

has a positive solution (see [Dancer and Perera 2001]). One of our main purposes is to extend this existence result from the  $p$ -Laplace equation to the corresponding problem involving our nonhomogeneous operator  $A$ . This is done in Theorem 25. We mention that in the special case of  $A(x, y) = A(y)$ , the result in [Kyritsi

et al. 2010] provides the existence of a positive solution if  $\alpha < \lambda_1 C_0 / (p - 1)$  and  $\lambda_1 C_1 / (p - 1) < \alpha_0$  hold (note that we can apply this result only to the case where  $\alpha < \alpha_0$ ). We emphasize that, for our general operator, the case  $\lambda_1(a_0) \neq \lambda_1(a_1)$  can occur. Note that in such a situation, contrary to the  $p$ -Laplacian case, we can still apply our theorem when  $\alpha_0 = \alpha$  provided this number is between  $\lambda_1(a_0)$  and  $\lambda_1(a_1)$ . The known result for the  $p$ -Laplacian case is obtained from our theorem simply by setting  $a_0 \equiv 1$  and  $a_\infty \equiv 1$ .

In particular, our theorem implies that if  $\lambda_1(a_0) \neq \lambda_1(a_\infty)$ , then every  $\lambda$  between  $\lambda_1(a_0)$  and  $\lambda_1(a_\infty)$  is an eigenvalue of  $A$  (see Corollary 26) and has a positive eigenfunction. This shows that, generally, the spectrum of the operator  $-\operatorname{div} A(x, \nabla \cdot)$  on  $W_0^{1,p}(\Omega)$  is not discrete.

In the final part of the paper, we treat the one side resonant and doubly resonant cases under additional conditions on  $f$ . For the  $p$ -Laplace equation, we refer to [Tanaka 2009] for the resonant and doubly resonant cases. Our Theorem 31 provides the existence of a positive solution in all cases of resonance for problem (P) with a nonhomogeneous operator in the principal part.

## 2. The properties of the map $A$

In what follows, the norm on  $W_0^{1,p}(\Omega)$  is given by

$$\|u\|^p := \|\nabla u\|_p^p,$$

where  $\|u\|_q$  denotes the usual norm of  $L^q(\Omega)$  for  $u \in L^q(\Omega)$  ( $1 \leq q \leq \infty$ ). Setting

$$(1) \quad G(x, y) := \int_0^{|y|} a(x, t) t \, dt,$$

we can easily see that

$$\nabla_y G(x, y) = A(x, y) \quad \text{and} \quad G(x, 0) = 0$$

for every  $x \in \bar{\Omega}$ ; see [Motreanu et al. 2011] for details.

**Remark 1.** The following assertions hold under Assumption A:

- (i) For all  $x \in \bar{\Omega}$ ,  $A(x, y)$  is maximal monotone and strictly monotone in  $y$ .
- (ii)  $|A(x, y)| \leq \frac{C_1}{p-1} |y|^{p-1}$  for every  $(x, y) \in \bar{\Omega} \times \mathbb{R}^N$ .
- (iii)  $A(x, y)y \geq \frac{C_0}{p-1} |y|^p$  for every  $(x, y) \in \bar{\Omega} \times \mathbb{R}^N$ .
- (iv)  $G(x, y)$  is strictly convex in  $y$  for all  $x$  and satisfies the inequalities

$$(2) \quad A(x, y)y \geq G(x, y) \geq \frac{C_0}{p(p-1)} |y|^p \quad \text{and} \quad G(x, y) \leq \frac{C_1}{p(p-1)} |y|^p$$

for every  $(x, y) \in \bar{\Omega} \times \mathbb{R}^N$ .

The following result is important for the proof of the Palais–Smale condition for the functionals related to our problem.

**Proposition 2** [Motreanu et al. 2011, Proposition 1]. *Let  $V : W_0^{1,p}(\Omega) \rightarrow W_0^{1,p}(\Omega)^*$  be the map defined by*

$$\langle V(u), v \rangle = \int_{\Omega} A(x, \nabla u) \nabla v \, dx$$

for  $u, v \in W_0^{1,p}(\Omega)$ . Then any sequence  $\{u_m\}$  that converges weakly to  $u$  and satisfies  $\limsup_{m \rightarrow \infty} \langle V(u_m), u_m - u \rangle \leq 0$  also converges strongly to  $u$ .

**Remark 3.** (i) *If  $u \in W_0^{1,p}(\Omega)$  is a solution of (P), then  $u \in C^{1,\alpha}(\bar{\Omega})$  for some  $0 < \alpha < 1$ .*

(ii) *If  $u \in W_0^{1,p}(\Omega)$  is a nontrivial solution of (P) such that  $u \geq 0$ , then  $u > 0$  in  $\Omega$  and  $\partial u / \partial \nu < 0$  on  $\partial \Omega$ , where  $\nu$  denotes the outward unit normal vector on  $\partial \Omega$ .*

*Sketch of proof.* (i) Let  $u \in W_0^{1,p}(\Omega)$  be a solution of (P). Then, because  $u \in L^\infty(\Omega)$  as shown by using the Moser iteration process (cf. [Miyajima et al. 2012, Appendix]), we see that  $u \in C^{1,\alpha}(\bar{\Omega})$  ( $0 < \alpha < 1$ ) by the regularity result in [Lieberman 1988].

(ii) Let  $u \in W_0^{1,p}(\Omega)$  be a solution of (P) satisfying  $u \geq 0$  and  $u \not\equiv 0$ . Then, by Assumption (f), we obtain a constant  $\lambda > 0$  satisfying

$$-\operatorname{div} A(x, \nabla u) + \lambda u^{p-1} \geq 0 \quad \text{in } \Omega.$$

Noting that  $u \in C^{1,\alpha}(\bar{\Omega})$  ( $0 < \alpha < 1$ ) by (i), we have  $u(x) > 0$  for every  $x \in \Omega$  by [Miyajima et al. 2012, Appendix, Theorem B]. In addition, using the strong maximum principle [ibid., Appendix, Theorem A], we easily see that  $\partial u(x) / \partial \nu < 0$  for every  $x \in \partial \Omega$ . □

**Proposition 4.** *Let  $f_n : \Omega \times \mathbb{R} \rightarrow \mathbb{R}$  be a Carathéodory function satisfying*

$$|f_n(x, t)| \leq D(1 + |t|^{r-1}) \quad \text{for every } x \in \Omega, t \in \mathbb{R}$$

with some positive constant  $D$  independent of  $n$  and  $r \in [p, p^*)$ , where  $p^* = \infty$  if  $N \leq p$  and  $p^* = pN / (N - p)$  if  $N > p$ . Assume that  $A_n : \Omega \times \mathbb{R}^N \rightarrow \mathbb{R}^N$  is a map satisfying parts (i)–(iv) of Assumption A with positive constants  $C'_1, C'_0$ , and  $C'_2$  independent of  $n$ . If  $u_n$  is a solution for

$$-\operatorname{div} A_n(x, \nabla u) = f_n(x, u) \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial \Omega$$

and  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ , then there exist a subsequence  $\{u_{n_l}\}$  of  $\{u_n\}$  and  $u_0 \in C_0^1(\bar{\Omega})$  such that  $u_{n_l} \rightarrow u_0$  in  $C_0^1(\bar{\Omega})$  as  $l \rightarrow \infty$ .

*Proof.* Since  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ , we may assume that  $u_n$  converges weakly to some  $u_0$  in  $W_0^{1,p}(\Omega)$  by choosing a subsequence. We can show that there exists a  $C > 0$  depending only on  $|\Omega|, p, N, D, C'_0, C'_1$ , and the embedding constant of

$W_0^{1,p}(\Omega)$  into  $L^{\bar{p}^*}(\Omega)$  such that  $\|u_n\|_\infty \leq C \max\{1, \|u_n\|^{(\bar{p}^*-p)/(\bar{p}^*-r)}\}$  by the Moser iteration process to [Miyajima et al. 2012, Theorem C], where  $\bar{p}^* = p^*$  if  $N > p$  and  $\bar{p}^* > r$  is any constant if  $N \leq p$ . Since  $D$ ,  $C'_1$ , and  $C'_0$  are independent of  $n$ ,  $\|u_n\|_\infty$  is bounded. Therefore, the regularity result in [Lieberman 1988] guarantees that there exist  $\gamma \in (0, 1)$  and  $M > 0$  independent of  $n$  such that  $u_n \in C_0^{1,\gamma}(\bar{\Omega})$  and  $\|u_n\|_{C_0^{1,\gamma}(\bar{\Omega})} \leq M$  (where we use the fact that  $C'_2$  is independent of  $n$ ). Since the inclusion of  $C_0^{1,\gamma}(\bar{\Omega})$  to  $C_0^1(\bar{\Omega})$  is compact,  $u_n$  converges to  $u_0$  in  $C_0^1(\bar{\Omega})$  (note that  $u_n \rightharpoonup u_0$  in  $W_0^{1,p}(\Omega)$ ).  $\square$

### 3. Eigenvalue problems

We introduce a function  $J : W_0^{1,p}(\Omega) \rightarrow \mathbb{R}$  by

$$(3) \quad J(u) = \int_{\Omega} G(x, \nabla u) \, dx \quad \text{for all } u \in W_0^{1,p}(\Omega).$$

It is clear that  $J$  is of class  $C^1$ . We also note that

$$(4) \quad rS := \{u \in W_0^{1,p}(\Omega) : \|u\|_p = r\} \quad \text{for } r > 0$$

is a  $C^1$  Finsler manifold (cf. [Deimling 1985, Sections 27.4 and 27.5]) because  $r$  is a regular value of the function  $u \mapsto \|u\|_p$  on  $W_0^{1,p}(\Omega)$ . Hence the norm of the derivative at  $u \in (rS)$  of the restriction  $\tilde{J}$  of  $J$  to  $rS$  is defined by

$$\begin{aligned} \|\tilde{J}'(u)\|_* &:= \min\{\|J'(u) - t\Phi'(u)\|_{W_0^{1,p}(\Omega)^*} : t \in \mathbb{R}\} \\ &= \sup\{\langle J'(u), v \rangle : v \in T_u(rS), \|v\| = 1\}, \end{aligned}$$

where  $\Phi(u) := (1/p)\|u\|_p^p$  and  $T_u(rS)$  denotes the tangent space of  $rS$  at  $u$ , that is,  $T_u(rS) = \{v \in W_0^{1,p}(\Omega) : \int_{\Omega} |u|^{p-2}uv \, dx = 0\}$ . It follows that the restriction  $\tilde{J} = J|_{(rS)}$  is a  $C^1$ -function on  $rS$  in the sense of manifolds.

**Proposition 5.** *For  $r > 0$ , the infimum*

$$(5) \quad \mu_1(A, r) = \inf_{u \in (rS)} \int_{\Omega} G(x, \nabla u) \, dx$$

*is attained at points  $\pm \hat{u}_r \in (rS)$  with  $\hat{u}_r \in C^{1,\alpha}(\bar{\Omega})$  and  $\hat{u}_r > 0$  in  $\Omega$ . Moreover,  $\pm \hat{u}_r$  are solutions of (EV;  $\lambda$ ) with  $\lambda = \lambda_1(A, \hat{u}_r)/r^p$ , where*

$$(6) \quad \lambda_1(A, \hat{u}_r) = \int_{\Omega} A(x, \nabla \hat{u}_r) \nabla \hat{u}_r \, dx \geq \frac{C_0}{p-1} \lambda_1 r^p.$$

*Proof.* Let  $\{u_n\} \subset (rS)$  be a minimizing sequence for (5). Using (2), it follows that  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ , so along a relabeled subsequence we have  $u_n \rightharpoonup u$  in  $W_0^{1,p}(\Omega)$  and  $u_n \rightarrow u$  in  $L^p(\Omega)$  for some  $u \in W_0^{1,p}(\Omega)$ , thus  $u \in (rS)$ . Since

$G(x, \cdot)$  is convex and continuous for all  $x \in \Omega$ ,  $J$  is weakly lower semicontinuous on  $W_0^{1,p}(\Omega)$  [Mawhin and Willem 1989, Theorem 1.2]. Therefore, we derive that

$$\mu_1(A, r) \leq \int_{\Omega} G(x, \nabla u) dx \leq \liminf_{n \rightarrow \infty} \int_{\Omega} G(x, \nabla u_n) dx,$$

which yields

$$\mu_1(A, r) = \int_{\Omega} G(x, \nabla u) dx.$$

The fact that the functional  $J$  is even implies that  $|u|$  is also a global minimizer of  $\tilde{J}_r$ . Consequently, we may assume that  $u \geq 0$ . On the other hand, the Lagrange multiplier rule leads to the existence of  $t \in \mathbb{R}$  such that

$$(7) \quad \int_{\Omega} A(x, \nabla u) \nabla v dx = t \int_{\Omega} u^{p-1} v dx \quad \text{for all } v \in W_0^{1,p}(\Omega).$$

Inserting  $v = u$  in (7) entails

$$(8) \quad \text{tr}^p = \int_{\Omega} A(x, \nabla u) \nabla u dx \geq \frac{C_0}{p-1} \|\nabla u\|_p^p \geq \frac{C_0 \lambda_1}{p-1} \|u\|_p^p = \frac{C_0 \lambda_1}{p-1} r^p.$$

Therefore, we have

$$t = \frac{\lambda_1(A, u)}{r^p} \geq \frac{C_0 \lambda_1}{p-1}.$$

From (7), it follows that  $u$  is a solution of (EV;  $\lambda$ ) with  $\lambda = t = \lambda_1(A, u)/r^p$ . According to Remark 3 with  $f(x, u) = t|u|^{p-2}u$ , it follows that  $u \in C^{1,\alpha}(\bar{\Omega})$  ( $0 < \alpha < 1$ ) and  $u > 0$  in  $\Omega$ . Since  $J$  is even and  $\lambda_1(A, u) = \lambda_1(A, -u)$ , we have that  $J(-u) = J(u) = \mu_1(A, r)$  and  $-u$  is a negative solution of (EV;  $\lambda$ ) with  $\lambda = t = \lambda_1(A, u)/r^p$ . The result is thus established with  $\hat{u}_r = u$ .  $\square$

We define

$$K_1(A, r) := \{u \in (rS) : J(u) = \mu_1(A, r)\}.$$

Then it follows from Proposition 5 that  $K_1(A, r)$  is not empty for each  $r > 0$ .

Because we do not know whether the minimizers of  $\tilde{J}_r$  are only  $\pm \hat{u}_r$ , we introduce the following:

$$\underline{\lambda}_1(A, r) := \inf \left\{ \int_{\Omega} A(x, \nabla u) \nabla u dx : u \in K_1(A, r) \right\},$$

$$\bar{\lambda}_1(A, r) := \sup \left\{ \int_{\Omega} A(x, \nabla u) \nabla u dx : u \in K_1(A, r) \right\}.$$

**Lemma 6.** *For every  $r > 0$ ,  $\underline{\lambda}_1(A, r)$  and  $\bar{\lambda}_1(A, r)$  are attained.*

*Proof.* We only deal with  $\underline{\lambda}_1(A, r)$  because  $\bar{\lambda}_1(A, r)$  can be treated similarly. Fix any  $r > 0$ . Let  $u_n \in K_1(A, r)$  satisfy  $\lambda_1(A, u_n) \rightarrow \underline{\lambda}_1(A, r)$  as  $n \rightarrow \infty$ . Then we

see that  $\|\nabla u_n\|_p$  is bounded from the inequality

$$\frac{C_0}{p(p-1)} \|\nabla u_n\|_p^p \leq \int_{\Omega} G(x, \nabla u_n) dx = \mu_1(A, r) \leq \int_{\Omega} G(x, \nabla w) dx$$

for  $w \in rS$ , where we use the definition of  $\mu_1(A, r)$  and (2). Recall that each  $u_n$  is a solution of (EV;  $\lambda$ ) with  $\lambda = \lambda_1(A, u_n)/r^p$ . Moreover, we have

$$\frac{C_0}{p-1} \lambda_1 r^p \leq \lambda_1(A, u_n) \leq \frac{C_1}{p-1} \|\nabla u_n\|_p^p$$

by Remark 1(ii) (see (6) for the first inequality), whence  $\lambda_1(A, u_n)/r^p$  is bounded. As a result, due to Proposition 4, we may assume that there exists  $u_0 \in W_0^{1,p}(\Omega)$  such that  $u_n \rightarrow u_0$  in  $C_0^1(\bar{\Omega})$  by choosing a subsequence if necessary. Since  $J$  and  $\lambda_1(A, \cdot)$  are continuous in  $W_0^{1,p}(\Omega)$ , we see that  $J(u_0) = \lim_{n \rightarrow \infty} J(u_n) = \mu_1(A, r)$ ,  $u_0 \in K_1(A, r)$ , and  $\lambda_1(A, u_0) = \lim_{n \rightarrow \infty} \lambda_1(A, u_n) = \underline{\lambda}_1(A, r)$ . Thus, our conclusion holds.  $\square$

Define

$$\lambda_1(A) := \inf_{u \neq 0} \int_{\Omega} \frac{A(x, \nabla u) \nabla u}{\|u\|_p^p} dx \quad \text{and} \quad \mu_1(A) := \inf_{u \neq 0} \int_{\Omega} \frac{G(x, \nabla u)}{\|u\|_p^p} dx.$$

**Lemma 7.**

$$\frac{C_0}{p-1} \lambda_1 \leq \lambda_1(A) \leq \min \left\{ \inf_{r>0} \frac{\lambda_1(A, r)}{r^p}, \frac{C_1}{p-1} \lambda_1 \right\} \quad \text{and} \quad \mu_1(A) = \inf_{r>0} \frac{\mu_1(A, r)}{r^p}.$$

*Proof.* First, we consider  $\lambda_1(A)$ . For every  $0 \neq u \in W_0^{1,p}(\Omega)$ , we have

$$(9) \quad \frac{C_0}{p-1} \frac{\|\nabla u\|_p^p}{\|u\|_p^p} \leq \int_{\Omega} \frac{A(x, \nabla u) \nabla u}{\|u\|_p^p} dx \leq \frac{C_1}{p-1} \frac{\|\nabla u\|_p^p}{\|u\|_p^p}$$

by Remark 1(ii)–(iii). Thus  $(C_0/(p-1))\lambda_1 \leq \lambda_1(A) \leq (C_1/(p-1))\lambda_1$  by taking the infimum with respect to  $u$ .

Here we fix any  $\varepsilon > 0$ . Then there exists an  $r_\varepsilon > 0$  such that  $\underline{\lambda}_1(A, r_\varepsilon)/r_\varepsilon^p \leq \inf_{r>0} (\underline{\lambda}_1(A, r)/r^p) + \varepsilon$ . By Lemma 6, we can choose  $u_\varepsilon \in (r_\varepsilon S)$  such that  $\lambda_1(A, u_\varepsilon) = \underline{\lambda}_1(A, r_\varepsilon)$ , that is,  $\int_{\Omega} A(x, \nabla u_\varepsilon) \nabla u_\varepsilon dx = \underline{\lambda}_1(A, r_\varepsilon)$ . By the definition of  $\lambda_1(A)$ , we obtain

$$\lambda_1(A) \leq \int_{\Omega} \frac{A(x, \nabla u_\varepsilon) \nabla u_\varepsilon}{\|u_\varepsilon\|_p^p} dx = \frac{\lambda_1(A, r_\varepsilon)}{r_\varepsilon^p} \leq \inf_{r>0} \frac{\lambda_1(A, r)}{r^p} + \varepsilon.$$

Because  $\varepsilon > 0$  is arbitrary, we have  $\lambda_1(A) \leq \inf_{r>0} (\lambda_1(A, r)/r^p)$ .

Next we treat  $\mu_1(A)$ . Fix any  $\varepsilon > 0$ . Then there exists an  $r_\varepsilon > 0$  such that  $\mu_1(A, r_\varepsilon)/r_\varepsilon^p \leq \inf_{r>0} (\mu_1(A, r)/r^p) + \varepsilon$ . On the other hand, because  $\mu_1(A, r_\varepsilon)$  is



attained at some  $u_\varepsilon \in (r_\varepsilon S)$ , we have

$$\inf_{u \neq 0} \int_{\Omega} \frac{G(x, \nabla u)}{\|u\|_p^p} dx \leq \int_{\Omega} \frac{G(x, \nabla u_\varepsilon)}{\|u_\varepsilon\|_p^p} dx = \frac{\mu_1(A, r_\varepsilon)}{r_\varepsilon^p} \leq \inf_{r>0} \frac{\mu_1(A, r)}{r^p} + \varepsilon.$$

Because  $\varepsilon > 0$  is arbitrary, this yields that  $\mu_1(A) \leq \inf_{r>0} (\mu_1(A, r)/r^p)$ .

For any  $\varepsilon > 0$ , we take  $v_\varepsilon \neq 0$  such that  $\int_{\Omega} (G(x, \nabla v_\varepsilon)/\|v_\varepsilon\|_p^p) dx \leq \mu_1(A) + \varepsilon$ . Then  $r_\varepsilon := \|v_\varepsilon\|_p > 0$  and so

$$\frac{\mu_1(A, r_\varepsilon)}{r_\varepsilon^p} \leq \int_{\Omega} \frac{G(x, \nabla v_\varepsilon)}{\|v_\varepsilon\|_p^p} dx \leq \mu_1(A) + \varepsilon.$$

This leads to  $\mu_1(A) \geq \inf_{r>0} (\mu_1(A, r)/r^p)$ . □

**Proposition 8.** *If  $\lambda < \lambda_1(A)$ , (EV;  $\lambda$ ) has no nontrivial solutions.*

*Proof.* Let  $u$  be a nontrivial solution of (EV;  $\lambda$ ) with  $\lambda < \lambda_1(A)$ . Then we have

$$\lambda_1(A) \leq \int_{\Omega} \frac{A(x, \nabla u) \nabla u}{\|u\|_p^p} dx = \lambda$$

by the definition of  $\lambda_1(A)$ . This is a contradiction. □

Set

$$(10) \quad A_p := \frac{C_1}{p-1} \left( \frac{C_1}{C_0} \right)^{p-1} \geq 1,$$

which is equal to 1 exactly in the case of  $A(x, y) = |y|^{p-2}y$  (that is, the special case of the  $p$ -Laplacian) because we can choose  $C_0 = C_1 = p - 1$ .

**Lemma 9** [Tanaka 2012a, Lemma 16]. *Let  $\varepsilon > 0$ . For every*

$$u, \varphi \in W^{1,p}(\Omega) \cap C^1(\Omega) \cap L^\infty(\Omega)$$

*with  $u \geq 0$  and  $\varphi \geq 0$  in  $\Omega$ , we have*

$$\int_{\Omega} A(x, \nabla u) \nabla \left( \frac{\varphi^p}{(u + \varepsilon)^{p-1}} \right) dx \leq A_p \|\nabla \varphi\|_p^p.$$

**Proposition 10.** *Any nontrivial solution of (EV;  $\lambda$ ) with  $\lambda > A_p \lambda_1$  changes sign.*

*Proof.* By way of contradiction, assume there is a solution  $u$  that does not change sign. Then we may suppose that  $u \geq 0$  because  $A$  is odd. Due to the strong maximum principle and the regularity theorem (see Remark 3), it follows that  $u \in C_0^1(\bar{\Omega})$  and  $u > 0$  in  $\Omega$ . Let  $\varphi_1$  be the positive eigenfunction of  $-\Delta_p$  corresponding to  $\lambda_1$  such that  $\|\varphi_1\|_p = 1$ . According to Lemma 9, we obtain

$$A_p \lambda_1 = A_p \|\nabla \varphi_1\|_p^p \geq \int_{\Omega} A(x, \nabla u) \nabla \left( \frac{\varphi_1^p}{(u + \varepsilon)^{p-1}} \right) dx = \lambda \int_{\Omega} \left( \frac{u}{u + \varepsilon} \right)^{p-1} \varphi_1^p dx$$

for every  $\varepsilon > 0$ . By taking  $\varepsilon \downarrow 0$ , we have  $\lambda \leq A_p \lambda_1$ . This is a contradiction. □

**Proposition 11.** *Assume  $A_p \lambda_1 < C_0 \lambda_2 / (p - 1)$ , where  $\lambda_2 > \lambda_1$  is the second eigenvalue of  $-\Delta_p$ . If  $A_p \lambda_1 < \lambda < C_0 \lambda_2 / (p - 1)$ ,  $(\text{EV}; \lambda)$  has no nontrivial solutions.*

*Proof.* By way of contradiction, we assume that  $(\text{EV}; \lambda)$  has a nontrivial solution  $u$ . Then it follows from Proposition 10 that  $u$  changes sign. Moreover, by taking  $u_{\pm}$  as a test function in  $(\text{EV}; \lambda)$ , we have

$$\frac{C_0}{p-1} \|\nabla u_{\pm}\|_p^p \leq \int_{\Omega} A(x, \nabla u)(\pm \nabla u_{\pm}) dx = \lambda \|u_{\pm}\|_p^p,$$

whence

$$(11) \quad \|\nabla u_{\pm}\|_p^p < \lambda_2 \|u_{\pm}\|_p^p.$$

This inequality guarantees the existence of a continuous path  $\gamma_0$  on  $S$  such that  $\gamma_0(0) = \varphi_1$ ,  $\gamma_0(1) = -\varphi_1$  and  $\max_{t \in [0,1]} \|\nabla \gamma_0(t)\|_p^p < \lambda_2$ ; refer to [Cuesta et al. 1999, Lemma 5.3]. This contradicts the equality

$$\lambda_2 = \inf_{\gamma \in \Sigma} \max_{t \in [0,1]} \Phi(\gamma(t)),$$

where  $\Phi(u) := \|\nabla u\|_p^p$  and  $\Sigma := \{\gamma \in C([0, 1], S) : \gamma(0) = \varphi_1, \gamma(1) = -\varphi_1\}$ ; see [Anane 1987; Cuesta et al. 1999]. This contradiction proves our result.

For the reader's convenience, we give the sketch of the construction of a path  $\gamma_0$  as required above. Define paths as follows:

$$\begin{aligned} \gamma_1(t) &:= \frac{tu + (1-t)u_+}{\|tu + (1-t)u_+\|_p} = \frac{u_+ - tu_-}{\|u_+ - tu_-\|_p}, & \gamma_2(t) &:= \frac{tu_+ + (1-t)u_-}{\|tu_+ + (1-t)u_-\|_p}, \\ \gamma_3(t) &:= \frac{(1-t)u - tu_-}{\|(1-t)u - tu_-\|_p} = \frac{(1-t)u_+ - u_-}{\|(1-t)u_+ - u_-\|_p} \end{aligned}$$

for  $t \in [0, 1]$ . Then, setting  $\tilde{\Phi} := \Phi|_S$ , we obtain by (11)

$$\max_{t \in [0,1]} \tilde{\Phi}(\gamma_i(t)) < \lambda_2, \quad \text{for } i = 1, 2, 3.$$

We recall that any component of  $\mathbb{O}(r) := \{u \in S : \tilde{\Phi}(u) < r\}$  contains at least one critical point of  $\tilde{\Phi}$ , where  $r > 0$  [Cuesta et al. 1999, Lemma 3.6]. Note that  $\mathbb{O}(\lambda_2)$  contains just two critical points  $\varphi_1$  and  $-\varphi_1$  because a critical value  $c$  of  $\tilde{\Phi}$  corresponds to the eigenvalue  $c$  of the negative  $p$ -Laplacian. Since any component of  $\mathbb{O}(\lambda_2)$  is path connected [ibid., Lemma 3.5], there exists a path  $\gamma_4$  joining from  $u_-/\|u_-\|_p$  to  $\varphi_1$  or  $-\varphi_1$  in  $\mathbb{O}(\lambda_2)$ . Thus, by noting that  $\Phi$  is even, we can construct a path  $\gamma_0 \in \Sigma$  such that  $\max_t \tilde{\Phi}(\gamma_0(t)) < \lambda_2$  by considering  $\gamma_4^{-1} \cdot \gamma_2 \cdot \gamma_1 \cdot \gamma_3 \cdot (-\gamma_4)$  or its inverse, where  $\gamma_i^{-1}(t) := \gamma_i(1-t)$  and  $\gamma_i \cdot \gamma_j$  denotes the path defined by  $\gamma_i(2t)$  if  $0 \leq t \leq \frac{1}{2}$  and  $\gamma_j(2t-1)$  if  $\frac{1}{2} < t \leq 1$ .  $\square$

**3.1. Asymptotically homogeneous case near zero.** We now consider the case where  $A$  is asymptotically  $(p-1)$ -homogeneous near zero in the following sense.

(AH0) *There exist a positive function  $a_0 \in C^1(\bar{\Omega}, \mathbb{R})$  and a continuous function  $\tilde{a}_0(x, t)$  on  $\bar{\Omega} \times [0, +\infty)$  such that*

$$A(x, y) = a_0(x)|y|^{p-2}y + \tilde{a}_0(x, |y|)y \quad \text{for every } x \in \Omega, y \in \mathbb{R}^N,$$

where

$$\lim_{t \rightarrow +0} \frac{\tilde{a}_0(x, t)}{t^{p-2}} = 0 \quad \text{uniformly in } x \in \bar{\Omega}.$$

For this weight function  $a_0$ , we define

$$(12) \quad \lambda_1(a_0) := \inf \left\{ \int_{\Omega} a_0(x)|\nabla u|^p dx : \|u\|_p = 1 \right\}.$$

Because  $0 < \min_{x \in \bar{\Omega}} a_0(x) \leq \max_{x \in \bar{\Omega}} a_0(x) < \infty$ , by the same argument as the one for the first eigenvalue of the negative  $p$ -Laplacian, we can prove that  $\lambda_1(a_0)$  is the first eigenvalue of

$$(13) \quad -\operatorname{div}(a_0(x)|\nabla u|^{p-2}\nabla u) = \lambda|u|^{p-2}u \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega.$$

Moreover,  $\lambda_1(a_0)$  has a positive eigenfunction  $\varphi_{a_0} \in C^1(\bar{\Omega})$  and it is simple. It is proved that (13) has no constant sign solutions other than 0 provided  $\lambda \neq \lambda_1(a_0)$ .

**Theorem 12.** *Assume (AH0). For every  $\varepsilon > 0$  there exists  $r_0 > 0$  such that equation (EV;  $\lambda$ ) has no nontrivial solutions in  $B_p(r_0) := \{v \in W_0^{1,p}(\Omega) : \|v\|_p < r_0\}$  provided  $\lambda < \lambda_1(a_0) - \varepsilon$ .*

*Proof.* We argue by contradiction. Thus we assume that there exist  $\varepsilon_0 > 0$ ,  $\{\lambda_n\}$  and  $\{u_n\}$  such that  $\lambda_n < \lambda_1(a_0) - \varepsilon_0$ ,  $u_n \in B_p(1/n)$  and  $u_n$  is a nontrivial solution of (EV;  $\lambda_n$ ). By taking  $u_n$  as a test function in (EV;  $\lambda_n$ ), we have

$$(14) \quad \frac{C_0}{p-1} \|\nabla u_n\|_p^p \leq \int_{\Omega} A(x, \nabla u_n) \nabla u_n dx = \lambda_n \|u_n\|_p^p \leq (\lambda_1(a_0) - \varepsilon_0)/n^p \rightarrow 0$$

as  $n \rightarrow \infty$ . Therefore,  $u_n \rightarrow 0$  in  $W_0^{1,p}(\Omega)$ . In addition, by noting that  $u_n$  is a nontrivial solution of (EV;  $\lambda_n$ ) and  $0 \leq \lambda_n < \lambda_1(a_0) - \varepsilon_0$ , Proposition 4 yields that  $u_n$  converges to 0 in  $C^1(\bar{\Omega})$ .

Set  $v_n := u_n/\|u_n\|_p$ . Then it follows from (14) and the boundedness of  $\{\lambda_n\}$  that  $\{v_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ . Hence, by choosing a subsequence, we may assume that  $v_n$  converges to some  $v_0$  weakly in  $W_0^{1,p}(\Omega)$  and strongly in  $L^p(\Omega)$ . Again by taking  $u_n/\|u_n\|_p^p$  as a test function in (EV;  $\lambda_n$ ), we obtain

$$\begin{aligned}
\lambda_1(a_0) - \varepsilon_0 > \lambda_n &= \int_{\Omega} \frac{a_0(x)|\nabla u_n|^p}{\|u_n\|_p^p} dx + \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2}{\|u_n\|_p^p} dx \\
&= \int_{\Omega} a_0(x)|\nabla v_n|^p dx + \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2}{\|u_n\|_p^p} \\
&\geq \lambda_1(a_0) + \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2}{\|u_n\|_p^p} =: \lambda_1(a_0) + I
\end{aligned}$$

because of the characterization of  $\lambda_1(a_0)$ . Hypothesis (AH0) guarantees that for every  $\delta > 0$  there exists  $\rho_0 > 0$  such that  $|\tilde{a}_0(x, t)| \leq \delta|t|^{p-2}$  if  $|t| \leq \rho_0$ . Since  $\|u_n\|_{C^1(\bar{\Omega})} \rightarrow 0$  and in view of (14), we can get

$$|I| \leq \delta \int_{\Omega} |\nabla v_n|^p dx \leq \frac{\delta(p-1)}{C_0} \lambda_n \leq \frac{\delta(p-1)}{C_0} (\lambda_1(a_0) - \varepsilon_0)$$

for sufficiently large  $n$ . As a result, by taking a sufficiently small  $\delta > 0$ , we have a contradiction for sufficiently large  $n$ .  $\square$

**Theorem 13.** *Assume (AH0). For every  $\varepsilon > 0$  there exists  $r_1 > 0$  such that (EV;  $\lambda$ ) has no constant sign solutions in  $B_p(r_1) \setminus \{0\}$  provided  $\lambda > \lambda_1(a_0) + \varepsilon$ .*

*Proof.* By way of contradiction, we assume that there exist  $\varepsilon_0 > 0$ ,  $\{\lambda_n\}$  and  $\{u_n\}$  such that  $\lambda_n > \lambda_1(a_0) + \varepsilon_0$ ,  $0 \neq u_n \in B_p(1/n)$  and  $u_n$  is a constant sign solution of (EV;  $\lambda_n$ ). Because  $A$  is odd, we may suppose that  $u_n \geq 0$  by considering  $-u_n$  if necessary. Thus, by Remark 3(i)–(ii),  $u_n \in C^1(\bar{\Omega})$  and  $u_n > 0$  in  $\Omega$ . We note that  $\lambda_n \leq A_p \lambda_1(-\Delta_p)$  by Proposition 10, where  $\lambda_1(-\Delta_p)$  denotes the first eigenvalue of  $-\Delta_p$  (see (10) for the definition of  $A_p$ ), and so  $\{\lambda_n\}$  is bounded. Therefore, we may assume that  $\lambda_n$  converges to some  $\lambda_0$  by choosing a subsequence. In addition, by the same argument as in Theorem 12, we can show that  $u_n \rightarrow 0$  in  $C^1(\bar{\Omega})$ .

Set  $A_n(x, y) := A(x, \|u_n\|_p y) / \|u_n\|_p^{p-1}$  and  $f_n(x, t) := \lambda_n |t|^{p-2} t$ . Then  $A_n$  satisfies Assumption A(i)–(iv) with the same constants  $C_0$ ,  $C_1$ , and  $C_2$ . Moreover,  $|f_n(x, t)| \leq \lambda_n |t|^{p-1} \leq A_p \lambda_1(-\Delta_p) |t|^{p-1}$  for every  $t \in \mathbb{R}$ , a.e.  $x \in \Omega$ . Note also that we have the boundedness of  $\|v_n\|$  due to the inequality  $C_0 \|\nabla u_n\|_p^p / (p-1) \leq \int_{\Omega} A(x, \nabla u_n) \nabla u_n dx = \lambda_n \|u_n\|_p^p$ . Since  $v_n := u_n / \|u_n\|_p$  is a positive solution of

$$-\operatorname{div}(A_n(x, \nabla u)) = f_n(x, u) \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega,$$

Proposition 4 guarantees that  $\{v_n\}$  has a convergent subsequence in  $C^1(\bar{\Omega})$ . By choosing a subsequence, we may suppose that  $v_n \rightarrow v_0 \neq 0$  in  $C^1(\bar{\Omega})$  (note that  $\|v_0\|_p = 1$ ). Using that we obtain, for every  $w \in W_0^{1,p}(\Omega)$ , that

$$\int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u_n|) \nabla u_n}{\|u_n\|_p^{p-1}} \nabla w dx = \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u_n|) \nabla u_n}{|\nabla u_n|^{p-1}} \nabla w |\nabla v_n|^{p-1} dx \rightarrow 0$$

as  $n \rightarrow \infty$  in view of (AH0) and the convergence  $u_n \rightarrow 0$ . As a result, letting

$n \rightarrow \infty$  in the equality

$$\int_{\Omega} a_0(x) |\nabla v_n|^{p-2} \nabla v_n \nabla w \, dx + \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u_n|) \nabla u_n}{\|u_n\|_p^{p-1}} \nabla w \, dx = \lambda_n \int_{\Omega} |v_n|^{p-2} v_n w \, dx$$

for each  $w \in W_0^{1,p}(\Omega)$ , we see that  $v_0 \neq 0$  is a positive solution of (13) with  $\lambda = \lambda_0$  (see Remark 3(ii) for  $v_0 > 0$ ). This yields that  $\lambda_0 = \lambda_1(a_0)$ , because (13) has no positive solutions other than  $\lambda = \lambda_1(a_0)$ . Therefore we have a contradiction, because  $\lambda_0 = \lim_{n \rightarrow \infty} \lambda_n \geq \lambda_1(a_0) + \varepsilon_0$ .  $\square$

**Proposition 14.** *Assume (AH0). Then, for every  $\varepsilon > 0$ , there exists  $r_0 > 0$  such that*

$$\frac{\lambda_1(A, r)}{r^p} \geq \lambda_1(a_0) - \varepsilon \quad \text{for every } 0 < r < r_0.$$

*Proof.* Assume that there exist  $\varepsilon > 0$  and  $r_n > 0$  such that  $r_n \rightarrow 0$  as  $n \rightarrow \infty$  and  $\lambda_1(A, r_n)/r_n^p < \lambda_1(a_0) - \varepsilon$  for every  $n \in \mathbb{N}$ . Because of Proposition 5 and Lemma 6 (note that  $A$  is odd in the second variable), we can choose a positive function  $u_n \in (r_n S) \cap C^1(\bar{\Omega})$  satisfying

$$\int_{\Omega} A(x, \nabla u_n) \nabla u_n \, dx = \lambda_1(A, r_n), \quad \min_{v \in r_n S} \int_{\Omega} G(x, \nabla v) \, dx = \int_{\Omega} G(x, \nabla u_n) \, dx.$$

Note that

$$(15) \quad \frac{C_0}{p-1} \|\nabla u_n\|_p^p \leq \int_{\Omega} A(x, \nabla u_n) \nabla u_n \, dx = \lambda_1(A, r_n) < (\lambda_1(a_0) - \varepsilon) r_n^p \rightarrow 0,$$

and so  $u_n \rightarrow 0$  in  $W_0^{1,p}(\Omega)$ . Because  $u_n$  is a solution of (EV;  $\lambda$ ) with  $\lambda = \lambda_1(A, r_n)/r_n^p$  (see Proposition 5), by combining the inequality

$$\lambda_1(a_0) - \varepsilon > \frac{\lambda_1(A, r_n)}{r_n^p} = \int_{\Omega} a_0(x) |\nabla v_n|^p \, dx + \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u_n|) |\nabla u_n|^2}{\|u_n\|_p^p} \, dx$$

and an argument as in Theorem 12 with  $\lambda_n = \lambda_1(A, r_n)/r_n^p$ , we have a contradiction.  $\square$

**Proposition 15.** *Assume (AH0). Then, for every  $\varepsilon > 0$ , there exists  $r_1 > 0$  such that*

$$\frac{\bar{\lambda}_1(A, r)}{r^p} \leq \lambda_1(a_0) + \varepsilon \quad \text{for every } 0 < r < r_1.$$

*Proof.* Assume that there exist  $\varepsilon_0 > 0$  and  $r_n > 0$  such that  $r_n \rightarrow 0$  as  $n \rightarrow \infty$  and  $\bar{\lambda}_1(A, r_n)/r_n^p > \lambda_1(a_0) + \varepsilon_0$  for every  $n \in \mathbb{N}$ . According to Lemma 6 and Proposition 5, we can take a positive function  $u_n \in (r_n S) \cap C^1(\bar{\Omega})$  satisfying

$$\int_{\Omega} A(x, \nabla u_n) \nabla u_n \, dx = \bar{\lambda}_1(A, r_n), \quad \min_{v \in r_n S} \int_{\Omega} G(x, \nabla v) \, dx = \int_{\Omega} G(x, \nabla u_n) \, dx.$$

Noting that, with  $\varphi_{a_0}$  the positive eigenfunction corresponding to  $\lambda_1(a_0)$  satisfying

$\|\varphi_{a_0}\|_p = 1$ , we have

$$\frac{C_0}{p(p-1)} \|\nabla u_n\|_p^p \leq \int_{\Omega} G(x, \nabla u_n) dx \leq \int_{\Omega} G(x, r_n \nabla \varphi_{a_0}) dx \leq \frac{C_1 r_n^p}{p(p-1)} \|\nabla \varphi_{a_0}\|_p^p,$$

we see that  $u_n \rightarrow 0$  in  $C^1(\bar{\Omega})$  due to Proposition 4, because  $u_n$  is a positive solution of (EV;  $\lambda$ ) with  $\lambda = \bar{\lambda}_1(A, r_n)/r_n^p$  and  $(\lambda_1(a_0) + \varepsilon_0 <) \bar{\lambda}_1(A, r_n)/r_n^p \leq A_p \lambda_1(-\Delta_p)$  by Proposition 10, where  $\lambda_1(-\Delta_p)$  denotes the first eigenvalue of  $-\Delta_p$  (see (10) for the definition of  $A_p$ ). Therefore, by the same argument as in Theorem 13 with  $\lambda_n = \bar{\lambda}_1(A, r_n)/r_n^p$ , we have a contradiction.  $\square$

The following result follows from Propositions 14 and 15, (note  $\underline{\lambda}_1(A, r) \leq \bar{\lambda}_1(A, r)$  for every  $r > 0$ ).

**Corollary 16.** *Under (AH0), we have*

$$\lim_{r \rightarrow +0} \frac{\bar{\lambda}_1(A, r)}{r^p} = \lim_{r \rightarrow +0} \frac{\underline{\lambda}_1(A, r)}{r^p} = \lambda_1(a_0).$$

**Proposition 17.** *Under (AH0), we have*

$$\lim_{r \rightarrow +0} \frac{\mu_1(A, r)}{r^p} = \frac{\lambda_1(a_0)}{p}.$$

*Proof.* Due to Proposition 5, for every  $r > 0$ , there exists a positive solution  $u_r \in (rS) \cap C^1(\bar{\Omega})$  of (EV;  $\lambda$ ) with  $\lambda = \lambda_1(A, u_r)/r^p$  and  $\mu_1(A, r) = J(u_r)$ . Then we can prove that  $u_r \rightarrow 0$  in  $C^1(\bar{\Omega})$  as  $r \rightarrow +0$  and  $u_r/\|u_r\|_p$  is bounded in  $W_0^{1,p}(\Omega)$  as  $r \rightarrow +0$  by a similar reason to the one in Proposition 15 (note that  $\lambda_1(A, u_r)/r^p$  is bounded as  $r \rightarrow +0$  by the inequality below and Corollary 16).

Set  $\tilde{G}_0(x, y) := \int_0^{|y|} \tilde{a}_0(x, t) t dt$  for  $y \in \mathbb{R}^N$ . We point out that

$$\underline{\lambda}_1(A, r) \leq \lambda_1(A, u_r) \leq \bar{\lambda}_1(A, r)$$

and

$$\begin{aligned} \mu_1(A, r) &= \int_{\Omega} G(x, \nabla u_r) dx = \frac{1}{p} \int_{\Omega} a_0(x) |\nabla u_r|^p dx + \int_{\Omega} \tilde{G}_0(x, \nabla u_r) dx \\ &= \frac{\lambda_1(A, u_r)}{p} - \frac{1}{p} \int_{\Omega} \tilde{a}_0(x, |\nabla u|) |\nabla u_r|^2 dx + \int_{\Omega} \tilde{G}_0(x, \nabla u_r) dx. \end{aligned}$$

Thus, by Corollary 16 and  $r = \|u_r\|_p$ , it suffices to prove

$$\lim_{r \rightarrow +0} \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u|) |\nabla u_r|^2}{\|u_r\|_p^p} dx = 0 \quad \text{and} \quad \lim_{r \rightarrow +0} \int_{\Omega} \frac{\tilde{G}_0(x, \nabla u_r)}{\|u_r\|_p^p} dx = 0.$$

Now we fix any  $\varepsilon > 0$ . Then, by (AH0), there exists  $\delta > 0$  such that

$$|\tilde{a}_0(x, t)| \leq \varepsilon t^{p-2} \quad \text{and} \quad |\tilde{G}_0(x, y)| \leq \varepsilon |y|^p/p \quad \text{for every } 0 < t \leq \delta, |y| \leq \delta.$$

Because  $u_r \rightarrow 0$  in  $C^1(\bar{\Omega})$  as  $r \rightarrow +0$ , we may assume that  $\|u_r\|_{C^1(\bar{\Omega})} \leq \delta$  for sufficiently small  $r > 0$ . Therefore, we obtain

$$\left| \int_{\Omega} \frac{\tilde{a}_0(x, |\nabla u|) |\nabla u_r|^2}{\|u_r\|_p^p} dx \right| \leq \varepsilon \frac{\|\nabla u_r\|_p^p}{\|u_r\|_p^p}, \quad \left| \int_{\Omega} \frac{\tilde{G}_0(x, \nabla u_r)}{\|u_r\|_p^p} dx \right| \leq \varepsilon \frac{\|\nabla u_r\|_p^p}{p \|u_r\|_p^p}.$$

Since  $\|\nabla u_r\|_p / \|u_r\|_p$  is bounded as  $r \rightarrow +0$  and  $\varepsilon > 0$  is arbitrary, our conclusion holds. □

**3.2. Asymptotically homogeneous case near  $\infty$ .** In this subsection, we consider the case where  $A$  is asymptotically  $(p-1)$ -homogeneous near  $\infty$  in the following sense.

(AH) *There exist a positive function  $a_\infty \in C^1(\bar{\Omega}, \mathbb{R})$  and a continuous function  $\tilde{a}(x, t)$  on  $\bar{\Omega} \times \mathbb{R}$  such that*

$$A(x, y) = a_\infty(x) |y|^{p-2} y + \tilde{a}(x, |y|) y \quad \text{for every } x \in \Omega, y \in \mathbb{R}^N,$$

where

$$\lim_{t \rightarrow +\infty} \frac{\tilde{a}(x, t)}{t^{p-2}} = 0 \quad \text{uniformly in } x \in \bar{\Omega}.$$

For the weight function  $a_\infty$ , we define

$$(16) \quad \lambda_1(a_\infty) := \inf \left\{ \int_{\Omega} a_\infty(x) |\nabla u|^p dx : \|u\|_p = 1 \right\}.$$

Because  $0 < \min_{x \in \bar{\Omega}} a_\infty(x) \leq \max_{x \in \bar{\Omega}} a_\infty(x) < \infty$ , by the same argument as for the first eigenvalue of  $-\Delta_p$ , we can prove the following elementary results:

(i)  $\lambda_1(a_\infty)$  is the first eigenvalue of

$$(17) \quad -\operatorname{div}(a_\infty(x) |\nabla u|^{p-2} \nabla u) = \lambda |u|^{p-2} u \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega.$$

(ii)  $\lambda_1(a_\infty)$  has a positive eigenfunction  $\varphi_{a_\infty} \in C^1(\bar{\Omega})$  with  $\|\varphi_{a_\infty}\|_p = 1$  and it is simple.

(iii) If  $\lambda \neq \lambda_1(a_\infty)$ , then (17) has no constant sign solutions other than 0.

**Theorem 18.** *Assume (AH). For every  $\varepsilon > 0$  there exists  $R_0 > 0$  such that equation (EV;  $\lambda$ ) has no solutions in  $W_0^{1,p}(\Omega) \setminus B_p(R_0)$  provided  $\lambda < \lambda_1(a_\infty) - \varepsilon$ .*

To prove the theorem, we need the following result.

**Lemma 19.** *Assume (AH) and let  $\{u_n\} \subset W_0^{1,p}(\Omega)$  be a sequence satisfying  $\|u_n\|_p \rightarrow \infty$  as  $n \rightarrow \infty$ . If  $v_n := u_n / \|u_n\|_p$  is bounded in  $W_0^{1,p}(\Omega)$ , the following assertions hold:*

$$(i) \quad \lim_{n \rightarrow \infty} \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) |\nabla u_n|^2}{\|u_n\|_p^p} dx = 0.$$

(ii) For every  $w \in W_0^{1,p}(\Omega)$ ,

$$\lim_{n \rightarrow \infty} \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) \nabla u_n \nabla w}{\|u_n\|_p^{p-1}} dx = 0.$$

(iii)  $\lim_{n \rightarrow \infty} \int_{\Omega} \frac{\tilde{G}(x, \nabla u_n)}{\|u_n\|_p^p} dx = 0$ , where  $\tilde{G}(x, y) := \int_0^{|y|} \tilde{a}(x, t) t dt$  for  $y \in \mathbb{R}^N$ .

*Proof.* (i) Fix any  $\varepsilon > 0$ . By the property of the function  $\tilde{a}$ , there exist  $R > 0$  and  $C > 0$  such that

$$(18) \quad |\tilde{a}(x, t)| \leq \varepsilon |t|^{p-2} \text{ if } t \geq R \quad \text{and} \quad |\tilde{a}(x, t)| \leq C \text{ if } 0 \leq t \leq R.$$

Therefore, we obtain

$$\begin{aligned} \left| \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) |\nabla u_n|^2}{\|u_n\|_p^p} dx \right| &\leq \int_{|\nabla u_n| > R} \varepsilon |\nabla v_n|^p dx + \int_{|\nabla u_n| \leq R} \frac{C |\nabla u_n|^2}{\|u_n\|_p^p} dx \\ &\leq \varepsilon \|\nabla v_n\|_p^p + \frac{CR^2 |\Omega|}{\|u_n\|_p^p} \leq \varepsilon D^p + \frac{CR^2 |\Omega|}{\|u_n\|_p^p} \end{aligned}$$

by (18), where  $D := \sup_n \|\nabla v_n\|_p$ . Letting  $n \rightarrow \infty$ , we have

$$\limsup_{n \rightarrow \infty} \left| \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) |\nabla u_n|^2}{\|u_n\|_p^p} dx \right| \leq \varepsilon D^p,$$

because  $\|u_n\|_p \rightarrow \infty$  as  $n \rightarrow \infty$ . Thus, since  $\varepsilon > 0$  is arbitrary, our conclusion holds.

(ii) For any  $\varepsilon > 0$  and  $w \in W_0^{1,p}(\Omega)$ , we have

$$\begin{aligned} \left| \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) \nabla u_n \nabla w}{\|u_n\|_p^{p-1}} dx \right| &\leq \int_{|\nabla u_n| > R} \varepsilon |\nabla v_n|^{p-1} |\nabla w| dx + \int_{|\nabla u_n| \leq R} \frac{C |\nabla u_n| |\nabla w|}{\|u_n\|_p^{p-1}} dx \\ &\leq \varepsilon \|\nabla v_n\|_p^{p-1} \|\nabla w\|_p + \frac{CR \|\nabla w\|_p |\Omega|^{(p-1)/p}}{\|u_n\|_p^{p-1}} \end{aligned}$$

by Hölder's inequality and (18). By combining this inequality and a similar argument to that used in (i), our conclusion is shown.

(iii) It is easily shown that, for every  $\varepsilon > 0$ , there exists  $C > 0$  such that

$$|\tilde{G}(x, y)| \leq \varepsilon |y|^p + C \quad \text{for every } y \in \mathbb{R}^N.$$

Therefore,  $\left| \int_{\Omega} \tilde{G}(x, \nabla u_n) dx \right| \leq \varepsilon \|\nabla u_n\|_p^p + C |\Omega|$ . This implies our conclusion.  $\square$



*Proof of Theorem 18.* By way of contradiction, we assume that there exist  $\varepsilon_0 > 0$ ,  $\{\lambda_n\}$ , and  $\{u_n\}$  such that  $\lambda_n < \lambda_1(a_\infty) - \varepsilon_0$ ,  $\lim_{n \rightarrow \infty} \|u_n\|_p = \infty$ , and  $u_n$  is a solution of (EV;  $\lambda_n$ ). By taking  $u_n$  as a test function in (EV;  $\lambda_n$ ), we have

$$\frac{C_0}{p-1} \|\nabla u_n\|_p^p \leq \int_\Omega A(x, \nabla u_n) \nabla u_n \, dx = \lambda_n \|u_n\|_p^p \leq (\lambda_1(a_\infty) - \varepsilon_0) \|u_n\|_p^p;$$

refer to Remark 1(iii). Therefore,  $v_n := u_n / \|u_n\|_p$  is bounded in  $W_0^{1,p}(\Omega)$ .

Again by taking  $u_n / \|u_n\|_p^p$  as a test function in (EV;  $\lambda_n$ ), we obtain

$$\begin{aligned} \lambda_1(a_\infty) - \varepsilon_0 > \lambda_n &= \int_\Omega \frac{a_\infty(x) |\nabla u_n|^p}{\|u_n\|_p^p} \, dx + \int_\Omega \frac{\tilde{a}(x, |\nabla u_n|) |\nabla u_n|^2}{\|u_n\|_p^p} \, dx \\ &= \int_\Omega a_\infty(x) |\nabla v_n|^p \, dx + \int_\Omega \frac{\tilde{a}(x, |\nabla u_n|) |\nabla u_n|^2}{\|u_n\|_p^p} \, dx \\ &\geq \lambda_1(a_\infty) + o(1), \end{aligned}$$

using the definition of  $\lambda_1(a_\infty)$  and Lemma 19(i). This is a contradiction. □

**Theorem 20.** Assume (AH). For every  $\varepsilon > 0$  there exists  $R_1 > 0$  such that (EV;  $\lambda$ ) has no constant sign solutions in  $W_0^{1,p}(\Omega) \setminus B_p(R_1)$  provided  $\lambda > \lambda_1(a_\infty) + \varepsilon$ .

*Proof.* By way of contradiction, we assume that there exist  $\varepsilon_0 > 0$ ,  $\{\lambda_n\}$ , and  $\{u_n\}$  such that  $\lambda_n > \lambda_1(a_\infty) + \varepsilon_0$ ,  $\lim_{n \rightarrow \infty} \|u_n\|_p = \infty$ , and  $u_n$  is a constant sign solution of (EV;  $\lambda_n$ ). Because  $A$  is odd, we may suppose that  $u_n \geq 0$  by considering  $-u_n$  if necessary. Thus, by Remark 3,  $u_n \in C^1(\bar{\Omega})$  and  $u_n > 0$  in  $\Omega$ . Here we note that  $\lambda_n \leq A_p \lambda_1(-\Delta_p)$  by Proposition 10, where  $\lambda_1(-\Delta_p)$  denotes the first eigenvalue of  $-\Delta_p$  (see (10) for the definition of  $A_p$ ), and so  $\{\lambda_n\}$  is bounded. Hence we may assume, by taking a subsequence, that  $\lambda_n$  converges to some

$$\lambda_0 \in [\lambda_1(a_\infty) + \varepsilon_0, A_p \lambda_1(-\Delta_p)].$$

In addition, we know that  $v_n := u_n / \|u_n\|_p$  is bounded in  $W_0^{1,p}(\Omega)$

$$\frac{C_0}{p-1} \|\nabla u_n\|_p^p \leq \int_\Omega A(x, \nabla u_n) \, dx = \lambda_n \|u_n\|_p^p,$$

where we take  $u_n$  as a test function in (EV;  $\lambda_n$ ). Thus, by choosing a subsequence, we may suppose that  $v_n$  converges to some  $v$  weakly in  $W_0^{1,p}(\Omega)$  and strongly in  $L^p(\Omega)$ .

We claim that  $v$  is a positive solution of

$$(19) \quad -\operatorname{div}(a_\infty(x) |\nabla v|^{p-2} \nabla v) = \lambda_0 |v|^{p-2} v \quad \text{in } \Omega, \quad v = 0 \quad \text{on } \partial\Omega,$$

that is,  $v$  is a positive eigenfunction corresponding to  $\lambda_0$ . If our claim holds, then  $\lambda_0 = \lambda_1(a_\infty)$  occurs because (17) has no positive solutions in the case of  $\lambda \neq \lambda_1(a_\infty)$ . Hence this contradicts  $\lambda_1(a_\infty) + \varepsilon_0 \leq \lim_{n \rightarrow \infty} \lambda_n = \lambda_0$ .

We now prove our claim. First, we show that  $v_n$  converges to  $v$  strongly in  $W_0^{1,p}(\Omega)$ . Indeed, by taking  $(v_n - v)/\|u_n\|_p^{p-1}$  as a test function in (EV;  $\lambda_n$ ), we have

$$\begin{aligned} & \lambda_n \int_{\Omega} v_n^{p-1} (v_n - v) \, dx \\ &= \int_{\Omega} a_{\infty}(x) |\nabla v_n|^{p-2} \nabla v_n \nabla (v_n - v) \, dx + \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) \nabla u_n}{\|u_n\|_p^{p-1}} \nabla (v_n - v) \, dx \\ &= \int_{\Omega} a_{\infty}(x) |\nabla v_n|^{p-2} \nabla v_n \nabla (v_n - v) \, dx + o(1) \end{aligned}$$

as  $n \rightarrow \infty$  due to Lemma 19(i)–(ii). Since  $v_n \rightarrow v$  in  $L^p(\Omega)$ , this implies that  $\int_{\Omega} a_{\infty}(x) |\nabla v_n|^{p-2} \nabla v_n \nabla (v_n - v) \, dx$  converges to 0 as  $n \rightarrow \infty$ . Noting that

$$\begin{aligned} o(1) &= \int_{\Omega} a_{\infty}(x) (|\nabla v_n|^{p-2} \nabla v_n - |\nabla v|^{p-2} \nabla v) \nabla (v_n - v) \, dx \\ &\geq \min_{\bar{\Omega}} a_{\infty} \int_{\Omega} (|\nabla v_n|^{p-2} \nabla v_n - |\nabla v|^{p-2} \nabla v) \nabla (v_n - v) \, dx \\ &\geq \min_{\bar{\Omega}} a_{\infty} (\|\nabla v_n\|_p^{p-1} - \|\nabla v\|_p^{p-1}) (\|\nabla v_n\|_p - \|\nabla v\|_p) \geq 0, \end{aligned}$$

we have  $v_n \rightarrow v$  in  $W_0^{1,p}(\Omega)$  (note  $0 < \min_{\bar{\Omega}} a_{\infty} \leq \max_{\bar{\Omega}} a_{\infty} < \infty$ ). As a result,  $v$  is a solution of (19) by letting  $n \rightarrow \infty$  in the equality

$$\int_{\Omega} a_{\infty}(x) |\nabla v_n|^{p-2} \nabla v_n \nabla w \, dx + \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) \nabla u_n \nabla w}{\|u_n\|_p^{p-1}} \, dx = \lambda_n \int_{\Omega} v_n^{p-1} w \, dx$$

for every  $w \in W_0^{1,p}(\Omega)$ ; note that, by Lemma 19(ii), the second term converges to zero. Since  $v_n = u_n/\|u_n\|_p > 0$  in  $\Omega$ ,  $v$  is nonnegative, and so  $v$  is positive by Remark 3(i) and  $\|v\|_p = 1$ . Thus our claim is shown.  $\square$

**Proposition 21.** *Assume (AH). Then, for every  $\varepsilon > 0$ , there exists  $R_0 > 0$  such that*

$$\frac{\underline{\lambda}_1(A, r)}{r^p} \geq \lambda_1(a_{\infty}) - \varepsilon \quad \text{for every } r > R_0.$$

*Proof.* Assume that there exist  $\varepsilon_0 > 0$  and  $r_n > 0$  such that  $r_n \rightarrow \infty$  as  $n \rightarrow \infty$  and  $\underline{\lambda}_1(A, r_n)/r_n^p < \lambda_1(a_{\infty}) - \varepsilon_0$  for every  $n \in \mathbb{N}$ . Because of Proposition 5 and Lemma 6, we can choose a positive function  $u_n \in (r_n S) \cap C^1(\bar{\Omega})$  satisfying

$$\int_{\Omega} A(x, \nabla u_n) \nabla u_n \, dx = \underline{\lambda}_1(A, r_n), \quad \min_{v \in r_n S} \int_{\Omega} G(x, \nabla v) \, dx = \int_{\Omega} G(x, \nabla u_n) \, dx.$$

Note that

$$\frac{C_0}{p-1} \|\nabla u_n\|_p^p \leq \int_{\Omega} A(x, \nabla u_n) \nabla u_n \, dx = \underline{\lambda}_1(A, r_n) < (\lambda_1(a_{\infty}) - \varepsilon_0) r_n^p,$$

and so  $u_n/r_n = u_n/\|u_n\|_p$  is bounded in  $W_0^{1,p}(\Omega)$ . Because  $u_n$  is a solution of (EV;  $\lambda$ ) with  $\lambda = \underline{\lambda}_1(A, r_n)/r_n^p$  (see Proposition 5), by the same argument as in Theorem 18 with  $\lambda_n = \underline{\lambda}_1(A, r_n)/r_n^p$ , we have a contradiction.  $\square$

**Proposition 22.** Assume (AH). Then, for every  $\varepsilon > 0$ , there exists  $R_1 > 0$  such that

$$\frac{\bar{\lambda}_1(A, r)}{r^p} \leq \lambda_1(a_\infty) + \varepsilon \quad \text{for every } r > R_1.$$

*Proof.* Assume that there exist  $\varepsilon_0 > 0$  and  $r_n > 0$  such that  $r_n \rightarrow \infty$  as  $n \rightarrow \infty$  and  $\bar{\lambda}_1(A, r_n)/r_n^p > \lambda_1(a_\infty) + \varepsilon_0$  for every  $n \in \mathbb{N}$ . According to Lemma 6 and Proposition 5, we can take a positive function  $u_n \in (r_n S) \cap C^1(\bar{\Omega})$  satisfying

$$\int_\Omega A(x, \nabla u_n) \nabla u_n \, dx = \bar{\lambda}_1(A, r_n), \quad \min_{v \in r_n S} \int_\Omega G(x, \nabla v) \, dx = \int_\Omega G(x, \nabla u_n) \, dx.$$

Note that, with  $\varphi_{a_\infty}$  as in item (ii) of page 165, we have

$$\frac{C_0}{p(p-1)} \|\nabla u_n\|_p^p \leq \int_\Omega G(x, \nabla u_n) \, dx \leq \int_\Omega G(x, r_n \nabla \varphi_{a_\infty}) \, dx \leq \frac{C_1 r_n^p}{p(p-1)} \|\nabla \varphi_{a_\infty}\|_p^p.$$

Hence  $u_n/r_n = u_n/\|u_n\|_p$  is bounded in  $W_0^{1,p}(\Omega)$ . Since  $u_n$  is a positive solution of (EV;  $\lambda$ ) with  $\lambda = \bar{\lambda}_1(A, r_n)/r_n^p$ , by the same argument as in Theorem 20 with  $\lambda_n = \bar{\lambda}_1(A, r_n)/r_n^p$ , we have a contradiction.  $\square$

By Propositions 21 and 22, we have the following result.

**Corollary 23.** Under (AH), we have

$$\lim_{r \rightarrow +\infty} \frac{\bar{\lambda}_1(A, r)}{r^p} = \lim_{r \rightarrow +\infty} \frac{\underline{\lambda}_1(A, r)}{r^p} = \lambda_1(a_\infty).$$

**Proposition 24.** Under (AH), we have

$$\lim_{r \rightarrow +\infty} \frac{\mu_1(A, r)}{r^p} = \frac{\lambda_1(a_\infty)}{p}.$$

*Proof.* Due to Proposition 5, for every  $r > 0$ , there exists a positive solution  $u_r \in (r S) \cap C^1(\bar{\Omega})$  of (EV;  $\lambda$ ) with  $\lambda = \lambda_1(A, u_r)/r^p$  and  $\mu_1(A, r) = J(u_r)$ . Then  $u_r/\|u_r\|_p = u_r/r$  is bounded in  $W_0^{1,p}(\Omega)$ , as seen from

$$\frac{C_0}{p(p-1)} \|\nabla u_r\|_p^p \leq \int_\Omega G(x, \nabla u_r) \, dx \leq \int_\Omega G(x, r \nabla w) \, dx \leq \frac{r^p C_1}{p(p-1)} \|\nabla w\|_p^p$$

for any  $w \in W_0^{1,p}(\Omega)$  with  $\|w\|_p = 1$ .

Set

$$\tilde{G}(x, y) := \int_0^{|y|} \tilde{a}(x, t) t \, dx \quad \text{for } y \in \mathbb{R}^N.$$

Note that

$$\underline{\lambda}_1(A, r) \leq \lambda_1(A, u_r) \leq \bar{\lambda}_1(A, r)$$

and

$$\begin{aligned}\mu_1(A, r) &= \int_{\Omega} G(x, \nabla u_r) dx = \frac{1}{p} \int_{\Omega} a_{\infty}(x) |\nabla u_r|^p dx + \int_{\Omega} \tilde{G}(x, \nabla u_r) dx \\ &= \frac{\lambda_1(A, u_r)}{p} - \frac{1}{p} \int_{\Omega} \tilde{a}(x, |\nabla u|) |\nabla u_r|^2 dx + \int_{\Omega} \tilde{G}(x, \nabla u_r) dx.\end{aligned}$$

According to Corollary 23 and Lemma 19(i) and (iii) (note  $\|u_r\|_p = r \rightarrow +\infty$ ), our conclusion is achieved.  $\square$

#### 4. Existence of a positive solution

In this section, we provide the existence of a positive solution to the equation

$$(P) \quad \begin{cases} -\operatorname{div} A(x, \nabla u) = f(x, u) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where the nonlinear term  $f$  satisfies Assumption (f).

**Theorem 25.** *Assume (AH0), (AH), and (f). Let  $\lambda_1(a_0)$  and  $\lambda_1(a_{\infty})$  be the first eigenvalues of, respectively, (13) and (17) (see the discussion there). If one of the following conditions holds, (P) has at least one positive solution.*

- (i)  $\alpha_0 > \lambda_1(a_0)$  and  $\alpha < \lambda_1(a_{\infty})$ .
- (ii)  $\alpha_0 < \lambda_1(a_0)$  and  $\alpha > \lambda_1(a_{\infty})$ .

This addresses the existence of an eigenvalue for our operator because we can apply Theorem 25 to  $f(x, u) = \lambda|u|^{p-2}u$ .

**Corollary 26.** *Assume (AH0), (AH), and  $\lambda_1(a_0) \neq \lambda_1(a_{\infty})$ . Then, for every  $\lambda$  between  $\lambda_1(a_0)$  and  $\lambda_1(a_{\infty})$ , (EV;  $\lambda$ ) has a nontrivial (positive) solution. Therefore  $\lambda$  is an eigenvalue of  $A$*

To show the existence of a positive solution, we define a  $C^1$  functional  $I$  on  $W_0^{1,p}(\Omega)$  by

$$I(u) := \int_{\Omega} G(x, \nabla u) dx - \int_{\Omega} F_+(x, u) dx \quad \text{for } u \in W_0^{1,p}(\Omega),$$

where  $F_+(x, u) := \int_0^u f_+(x, t) dt$ , with  $f_+(x, t)$  given by  $f(x, t)$  if  $t \geq 0$  and 0 if  $t \leq 0$ .

**Remark 27.** If  $u \in W_0^{1,p}(\Omega)$  is a nontrivial critical point of  $I$ , then  $u$  is a positive solution of (P).

Indeed, by taking  $-u_-$  as a test function, we obtain

$$\begin{aligned}0 = \langle I'(u), -u_- \rangle &= \int_{\Omega} A(x, \nabla u)(-\nabla u_-) dx - \int_{\Omega} f_+(x, u)(-u_-) dx \\ &= \int_{\Omega} A(x, \nabla u)(-\nabla u_-) dx \geq \frac{C_0}{p-1} \|\nabla u_-\|_p^p.\end{aligned}$$

Thus  $u \geq 0$ . By Remark 3(ii) (note that  $u \not\equiv 0$ ), we see that  $u$  is a positive solution of (P) (note that  $f_+(x, u) = f(x, u)$ ).

**Convention.** From now on, let Assumption (f) be satisfied.

**Lemma 28.** *If  $\alpha \neq \lambda_1(a_\infty)$ , then  $I$  satisfies the Palais–Smale condition.*

*Proof.* Let  $\{u_n\}$  be a Palais–Smale sequence of  $I$ , which means that

$$I(u_n) \rightarrow c \quad \text{and} \quad \|I'(u_n)\|_{W_0^{1,p}(\Omega)^*} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

for some  $c \in \mathbb{R}$ . In view of Proposition 2 and the compactness of the embedding  $W_0^{1,p}(\Omega) \hookrightarrow L^p(\Omega)$ , it is sufficient to prove the boundedness of  $\{u_n\}$  in  $W_0^{1,p}(\Omega)$ . Then, in view of the inequality

$$(20) \quad \frac{C_0}{p(p-1)} \|\nabla u_n\|_p^p \leq \int_\Omega G(x, \nabla u_n) dx = I(u_n) + \int_\Omega F_+(x, u_n) dx \leq I(u_n) + C\|u_n\|_p^p,$$

it is sufficient to prove the boundedness of  $\{u_n\}$  in  $L^p(\Omega)$ . By way of contradiction we may assume that  $\|u_n\|_p \rightarrow \infty$  as  $n \rightarrow \infty$  by choosing a subsequence if necessary. Set  $v_n := u_n/\|u_n\|_p$ . The inequality (20) ensures that  $\{v_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ . Hence, by choosing a subsequence, we may suppose that  $v_n \rightharpoonup v_0$  in  $W_0^{1,p}(\Omega)$  and  $v_n \rightarrow v_0$  in  $L^p(\Omega)$  for some  $v_0$ .

First, we see that  $v_0 \geq 0$  for a.e.  $x \in \Omega$ . Indeed, by taking  $-(u_n)_-$  as a test function, we have

$$\begin{aligned} o(1)\|\nabla(u_n)_-\|_p &= \langle I'(u_n), -(u_n)_- \rangle \\ &= \int_\Omega A(x, \nabla u_n)(-\nabla(u_n)_-) dx \geq \frac{C_0}{p-1} \|\nabla(u_n)_-\|_p^p. \end{aligned}$$

Because  $p > 1$ , we have  $\|\nabla(u_n)_-\|_p \rightarrow 0$  as  $n \rightarrow \infty$ . Thus  $(v_n)_- \rightarrow 0$  in  $W_0^{1,p}(\Omega)$ , and hence  $(v_0)_- = 0$  for a.e.  $x \in \Omega$ .

Now we prove that

$$(21) \quad \lim_{n \rightarrow \infty} \frac{\|f_+(\cdot, u_n) - \alpha(u_n)_+^{p-1}\|_{p'}}{\|u_n\|_p^{p-1}} = 0,$$

where  $p' = p/(p-1)$ . Fix an arbitrary  $\varepsilon > 0$ . It follows from condition (ii) of Assumption (f) that there exists a  $C_\varepsilon > 0$  such that

$$|f(x, u) - \alpha u^{p-1}| \leq \varepsilon |u|^{p-1} + C_\varepsilon \quad \text{for every } u \geq 0, \text{ a.e. } x \in \Omega.$$

Then we obtain

$$\int_\Omega |f_+(x, u_n) - \alpha(u_n)_+^{p-1}|^{p'} dx \leq 2^{p'-1}(\varepsilon^{p'-1} \|(u_n)_+\|_p^p + C_\varepsilon^{p'-1} |\Omega|).$$

Since we are assuming that  $\|u_n\|_p \rightarrow \infty$  as  $n \rightarrow \infty$ , this shows that

$$\lim_{n \rightarrow \infty} \|f_+(\cdot, u_n) - \alpha(u_n)_+^{p-1}\|_{p'} / \|u_n\|_p^{p-1} = 0,$$

because  $\varepsilon > 0$  is arbitrary.

Here we recall the following result proved in Lemma 19:

$$(22) \quad \lim_{n \rightarrow \infty} \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) \nabla u_n}{\|u_n\|_p^{p-1}} \nabla(v_n - v_0) dx = \lim_{n \rightarrow \infty} \int_{\Omega} \frac{\tilde{a}(x, |\nabla u_n|) \nabla u_n}{\|u_n\|_p^{p-1}} \nabla \varphi dx = 0$$

for every  $\varphi \in W_0^{1,p}(\Omega)$ . Thus, by considering

$$o(1) = \frac{\langle I'(u_n), v_n - v_0 \rangle}{\|u_n\|_p^{p-1}} = \int_{\Omega} a_{\infty}(x) |\nabla v_n|^{p-2} \nabla v_n \nabla(v_n - v_0) dx + o(1),$$

and using Proposition 2, we see that  $v_n$  converges strongly to  $v_0$  in  $W_0^{1,p}(\Omega)$ . Hence, by passing to the limit in  $o(1) = \langle I'(u_n), \varphi \rangle / \|u_n\|_p^{p-1}$  for any  $\varphi \in W_0^{1,p}(\Omega)$  and by noting (21) and (22), we infer that  $v_0$  is a nontrivial solution of

$$-\operatorname{div}(a_{\infty} |\nabla u|^{p-2} \nabla u) = \alpha |u|^{p-2} u \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega$$

(note that  $\|v_0\|_p = 1$  and  $v_0 \geq 0$  for a.e.  $x \in \Omega$ ). Since  $v_0 \geq 0$  for a.e.  $x \in \Omega$ ,  $v$  is a positive solution of (17) with  $\lambda = \alpha$  (see Remark 3). This implies that  $\alpha = \lambda_1(a_{\infty})$ , because (17) has no positive solutions if  $\lambda \neq \lambda_1(a_{\infty})$ . It contradicts the hypothesis  $\alpha \neq \lambda_1(a_{\infty})$ . Hence  $\|u_n\|_p$  is bounded, which completes the proof.  $\square$

**Lemma 29.** *Assume (AH) and  $\alpha < \lambda_1(a_{\infty})$ . Then  $I$  is coercive, bounded from below and weakly lower semicontinuous (wlsc) on  $W_0^{1,p}(\Omega)$ .*

*Proof.* Because  $\alpha < \lambda_1(a_{\infty})$ , we can take sufficiently small constants  $\varepsilon > 0$  and  $0 < \delta < 1$  satisfying

$$(23) \quad (1 - \delta)(\lambda_1(a_{\infty}) - \varepsilon) > \alpha + \varepsilon.$$

By condition (ii) of Assumption (f), there exists a  $C > 0$  such that

$$|F_+(x, u)| \leq (\alpha + \varepsilon) \frac{u^p}{p} + C$$

for every  $u \geq 0$  and a.e.  $x \in \Omega$ . Due to Proposition 24 and the definition of  $\mu_1(A, r)$ , there exists an  $R > 0$  such that, for every  $u \in W_0^{1,p}(\Omega)$  with  $\|u\|_p \geq R$ ,

$$(24) \quad \int_{\Omega} G(x, \nabla u) dx \geq \mu_1(A, \|u\|_p) \geq \frac{\lambda_1(a_{\infty}) - \varepsilon}{p} \|u\|_p^p.$$

Hence, for every  $u \in W_0^{1,p}(\Omega)$  with  $\|u\|_p \geq R$ , we obtain

$$\begin{aligned}
 I(u) &\geq \frac{(1-\delta)(\lambda_1(a_\infty) - \varepsilon)}{p} \|u\|_p^p + \frac{\delta C_0}{p(p-1)} \|\nabla u\|_p^p - \frac{\alpha + \varepsilon}{p} \|u_+\|_p^p - C|\Omega| \\
 &\geq \frac{\delta C_0}{p(p-1)} \|\nabla u\|_p^p - C|\Omega|
 \end{aligned}$$

by (2), (23), and (24), where  $u_+ := \max\{0, u\}$ . This yields that  $I$  is coercive. Moreover, because  $I$  is bounded from below on  $B_p(R)$ , we see that  $I$  is bounded from below on  $W_0^{1,p}(\Omega)$ . Since  $J$  is wisc (see the proof of Proposition 5) and  $W_0^{1,p}(\Omega) \hookrightarrow L^p(\Omega)$  is compact,  $I$  is wisc on  $W_0^{1,p}(\Omega)$ .  $\square$

**Lemma 30.** *Assume (AH0) and  $\alpha_0 < \lambda_1(a_0)$ . Let  $p < q \leq p^*$ , where  $p^* = Np/(N - p)$  if  $N > p$  and  $p^* = +\infty$  if  $N \leq p$ . Then there exists  $\rho_0 > 0$  such that*

$$\inf\{I(u) : \|u\|_q = \rho\} > 0 \quad \text{for every } 0 < \rho < \rho_0.$$

*Proof.* Because  $\alpha_0 < \lambda_1(a_0)$ , we can take some sufficiently small  $\varepsilon > 0$  and  $0 < \delta < 1$  satisfying

$$(25) \quad (1 - \delta)(\lambda_1(a_0) - \varepsilon) > \alpha_0 + \varepsilon.$$

According to Proposition 17, there exists an  $r_0 > 0$  such that

$$(26) \quad \frac{\mu_1(A, r)}{r^p} \geq \frac{\lambda_1(a_0) - \varepsilon}{p} \quad \text{for every } 0 < r < r_0.$$

In addition, Assumption (f) guarantees the existence of  $D_q > 0$  satisfying

$$(27) \quad F_+(x, u) \leq \frac{\alpha_0 + \varepsilon}{p} u^p + D_q u^q \quad \text{for every } u \geq 0, \text{ a.e. } x \in \Omega.$$

Because  $W_0^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$  is continuous, we can take a positive constant  $C_q$  such that  $\|u\|_q \leq C_q \|\nabla u\|_p$  for every  $W_0^{1,p}(\Omega)$ . We choose a positive constant  $\rho$  satisfying

$$(28) \quad \rho < \min \left\{ r_0 |\Omega|^{1/q-1/p}, \left( \frac{\delta C_0}{2p(p-1)D_q C_q^p} \right)^{1/(q-p)} \right\} =: \rho_0.$$

Note that  $\|u\|_p < r_0$  if  $\|u\|_q = \rho$ , by Hölder’s inequality and (28). Therefore, for every  $\|u\|_q = \rho$ , we have

$$\begin{aligned}
 I(u) &= (1 - \delta) \int_\Omega G(x, \nabla u) dx + \delta \int_\Omega G(x, \nabla u) dx - \int_\Omega F_+(x, u) dx \\
 &\geq (1 - \delta) \frac{\mu_1(A, \|u\|_p)}{\|u\|_p^p} \|u\|_p^p + \frac{\delta C_0}{p(p-1)} \|\nabla u\|_p^p - \frac{\alpha_0 + \varepsilon}{p} \|u_+\|_p^p - D_q \|u_+\|_q^q \\
 &\geq \frac{1}{p} \{ (1 - \delta)(\lambda_1(a_0) - \varepsilon) - \alpha_0 - \varepsilon \} \|u\|_p^p + \left( \frac{\delta C_0}{p(p-1)C_q^p} - D_q \|u\|_q^{q-p} \right) \|u\|_q^p
 \end{aligned}$$

$$\geq \frac{\delta C_0}{2p(p-1)C_q^p} \rho^p,$$

by the definition of  $\mu_1(A, r)$ , (2), (27), (26), (25), and (28). This ensures our conclusion.  $\square$

*Proof of Theorem 25.* (i) Lemma 29 guarantees the existence of a global minimizer of  $I$ . Thus it suffices to prove that  $\min_{W_0^{1,p}(\Omega)} I < 0$  to show the existence of a nontrivial critical point of  $I$ . Choose a positive constant  $\varepsilon > 0$  such that  $\alpha_0 > \lambda_1(a_0) + 2\varepsilon$ . Let  $\varphi_{a_0} \in C^1(\bar{\Omega})$  be a positive eigenfunction corresponding to  $\lambda_1(a_0)$  with  $\|\varphi_{a_0}\|_p = 1$  (refer to the text below (13) and note that (13) is a homogeneous equation). It is easily seen that  $\int_{\Omega} \tilde{G}_0(x, r\nabla\varphi_{a_0}) dx/r^p \rightarrow 0$  as  $r \rightarrow +0$  (refer to the proof of Proposition 17 with  $\|r\varphi_{a_0}\|_p = r$ ). Hence there exists  $r_0 > 0$  such that

$$(29) \quad \int_{\Omega} G(x, r\nabla\varphi_{a_0}) dx = \frac{r^p}{p} \int_{\Omega} a_0(x) |\nabla\varphi_{a_0}|^p dx + r^p \int_{\Omega} \frac{\tilde{G}_0(x, r\nabla\varphi_{a_0})}{r^p} dx \\ \leq \frac{\lambda_1(a_0) + \varepsilon}{p} r^p = \frac{\lambda_1(a_0) + \varepsilon}{p} \|r\varphi_{a_0}\|_p^p$$

for every  $0 < r < r_0$ . On the other hand, it follows from part (i) of Assumption (f) that there exists a  $\delta > 0$  such that

$$(30) \quad F_+(x, u) \geq \frac{\alpha_0 - \varepsilon}{p} u^p \quad \text{for every } u \in [0, \delta], \text{ a.e. } x \in \Omega.$$

Therefore, for every  $0 < r < \min\{r_0, \delta/\|\varphi_{a_0}\|_{\infty}\}$ , we have

$$I(ru_0) \leq \frac{r^p}{p} (\lambda_1(a_0) + 2\varepsilon - \alpha_0) \|\varphi_{a_0}\|_p^p < 0,$$

by (29) and (30) (note  $\lambda_1(a_0) + 2\varepsilon - \alpha_0 < 0$ ), whence  $\min_{W_0^{1,p}(\Omega)} I < 0$ .

(ii) Let  $p < q \leq p^*$ . Then, by Lemma 30, we obtain  $\rho > 0$  satisfying

$$\delta_0 := \inf\{I(u) : \|u\|_q = \rho\} > 0.$$

Now we claim the existence of  $w \in W_0^{1,p}(\Omega)$  such that

$$(31) \quad \|w\|_q > \rho \quad \text{and} \quad I(w) < \delta_0.$$

Admitting this claim, we define

$$c := \inf_{\gamma \in \Gamma} \max_{t \in [0,1]} I(\gamma(t)), \quad \Gamma := \{\gamma \in C([0,1], W_0^{1,p}(\Omega)) : \gamma(0) = 0, \gamma(1) = w\}.$$

It is obvious that  $\Gamma \neq \emptyset$  and  $\gamma([0,1]) \cap \{u \in W_0^{1,p}(\Omega) : \|u\|_q = \rho\} \neq \emptyset$  for every  $\gamma \in \Gamma$ , since  $W_0^{1,p}(\Omega) \hookrightarrow L^q(\Omega)$  is continuous. Thus the mountain pass theorem guarantees that  $c(\geq \delta_0)$  is a nontrivial critical value of  $I$  because  $I$  satisfies the Palais–Smale condition by Lemma 28.



Finally, we prove the existence of  $w$  satisfying (31). Because  $\alpha > \lambda_1(a_\infty)$ , we can choose a positive constant  $\varepsilon_0 > 0$  such that

$$(32) \quad \alpha > \lambda_1(a_\infty) + 2\varepsilon_0.$$

Using item (ii) on page 165, we can take  $\varphi_{a_\infty} \in C^1(\bar{\Omega})$  be a positive eigenfunction corresponding to  $\lambda_1(a_\infty)$  with  $\|\varphi_{a_\infty}\|_p = 1$ . It follows from Lemma 19(iii) that

$$\int_{\Omega} \tilde{G}(x, r\nabla\varphi_{a_\infty}) dx / r^p \rightarrow 0$$

as  $r \rightarrow +\infty$  (note that  $\|r\varphi_{a_\infty}\|_p = r$ ). Hence there exists  $R_0 > 0$  such that

$$(33) \quad \int_{\Omega} G(x, r\nabla\varphi_{a_\infty}) dx = \frac{r^p}{p} \int_{\Omega} a_\infty(x) |\nabla\varphi_{a_\infty}|^p dx + r^p \int_{\Omega} \frac{\tilde{G}_0(x, r\nabla\varphi_{a_\infty})}{r^p} dx \\ \leq \frac{\lambda_1(a_\infty) + \varepsilon_0}{p} r^p = \frac{\lambda_1(a_\infty) + \varepsilon_0}{p} \|r\varphi_{a_\infty}\|_p^p$$

for every  $r \geq R_0$ . In addition, it follows from condition (ii) of Assumption (f) that there exists  $D > 0$  such that

$$(34) \quad F_+(x, u) \geq \frac{\alpha - \varepsilon_0}{p} u^p - D \quad \text{for every } u \geq 0, \text{ a.e. } x \in \Omega.$$

Consequently, by (32), (33), and (34), we obtain

$$I(r\varphi_{a_0}) \leq \frac{r^p}{p} (\lambda_1(a_\infty) + 2\varepsilon_0 - \alpha) \|\varphi_{a_0}\|_p^p + D|\Omega| \rightarrow -\infty$$

as  $t \rightarrow +\infty$ . This implies the existence of  $w$  satisfying (31). □

**4.1. Resonant cases.** To consider the resonant cases, we introduce the following hypotheses for

$$\tilde{G}(x, y) := \int_0^{|y|} \tilde{a}(x, t)t dt \quad \text{and} \quad \tilde{G}_0(x, y) := \int_0^{|y|} \tilde{a}_0(x, t)t dt,$$

where  $\tilde{a}$  and  $\tilde{a}_0$  are as in (AH) and (AH0).

(H+) There exist  $1 \leq q < p$  and  $H_0 > 0$  such that

$$\lim_{|y| \rightarrow \infty} \frac{p\tilde{G}(x, y) - \tilde{a}(x, |y|)|y|^2}{|y|^q} = +\infty \quad \text{for a.e. } x \in \Omega, \\ p\tilde{G}(x, y) - \tilde{a}(x, |y|)|y|^2 \geq -H_0(1 + |y|^q) \quad \text{for a.e. } x \in \Omega, \text{ every } y \in \mathbb{R}^N, \\ f(x, t)t - pF(x, t) \geq -H_0(1 + t^q) \quad \text{for a.e. } x \in \Omega, \text{ every } t \geq 0.$$

(H−) There exist  $1 \leq q < p$  and  $H_0 > 0$  such that

$$\begin{aligned} \lim_{|y| \rightarrow \infty} \frac{p\tilde{G}(x, y) - \tilde{a}(x, |y|)|y|^2}{|y|^q} &= -\infty && \text{for a.e. } x \in \Omega, \\ p\tilde{G}(x, y) - \tilde{a}(x, |y|)|y|^2 &\leq H_0(1 + |y|^q) && \text{for a.e. } x \in \Omega, \text{ every } y \in \mathbb{R}^N, \\ f(x, t)t - pF(x, t) &\leq H_0(t^q + 1) && \text{for a.e. } x \in \Omega, \text{ every } t \geq 0. \end{aligned}$$

(HF+) There exist  $1 \leq q < p$  and  $H_0 > 0$  such that

$$\begin{aligned} p\tilde{G}(x, y) - \tilde{a}(x, |y|)|y|^2 &\geq -H_0(1 + |y|^q) && \text{for a.e. } x \in \Omega, \text{ every } y \in \mathbb{R}^N, \\ f(x, t)t - pF(x, t) &\geq -H_0(1 + t^q) && \text{for every } t \geq 0, \text{ a.e. } x \in \Omega, \\ \lim_{t \rightarrow +\infty} \frac{f(x, t)t - pF(x, t)}{t^q} &= +\infty && \text{for a.e. } x \in \Omega. \end{aligned}$$

(HF−) There exist  $1 \leq q < p$  and  $H_0 > 0$  such that

$$\begin{aligned} p\tilde{G}(x, y) - \tilde{a}(x, |y|)|y|^2 &\leq H_0(1 + |y|^q) && \text{for a.e. } x \in \Omega, \text{ every } y \in \mathbb{R}^N, \\ f(x, t)t - pF(x, t) &\leq H_0(1 + t^q) && \text{for every } t \geq 0, \text{ a.e. } x \in \Omega, \\ \lim_{t \rightarrow +\infty} \frac{f(x, t)t - pF(x, t)}{t^q} &= -\infty && \text{for a.e. } x \in \Omega. \end{aligned}$$

(H0+) There exist  $p \leq r < p^*$  and  $H_0 > 0$  such that

$$\begin{aligned} \lim_{|y| \rightarrow 0} \frac{p\tilde{G}_0(x, y) - \tilde{a}_0(x, |y|)|y|^2}{|y|^r} &= +\infty && \text{for a.e. } x \in \Omega, \\ p\tilde{G}_0(x, y) - \tilde{a}_0(x, |y|)|y|^2 &\geq -H_0|y|^r && \text{for a.e. } x \in \Omega, \text{ every } |y| \leq 1, \\ f(x, t)t - pF(x, t) &\geq -H_0t^r && \text{for a.e. } x \in \Omega, \text{ every } t \in [0, 1]. \end{aligned}$$

(H0−) There exist  $p \leq r < p^*$  and  $H_0 > 0$  such that

$$\begin{aligned} \lim_{|y| \rightarrow 0} \frac{p\tilde{G}_0(x, y) - \tilde{a}_0(x, |y|)|y|^2}{|y|^r} &= -\infty && \text{for a.e. } x \in \Omega, \\ p\tilde{G}_0(x, y) - \tilde{a}_0(x, |y|)|y|^2 &\leq H_0|y|^r && \text{for a.e. } x \in \Omega, \text{ every } |y| \leq 1, \\ f(x, t)t - pF(x, t) &\leq H_0t^r && \text{for a.e. } x \in \Omega, \text{ every } t \in [0, 1]. \end{aligned}$$

(HF0+) There exist  $p \leq r < p^*$  and  $H_0 > 0$  such that

$$\begin{aligned} p\tilde{G}_0(x, y) - \tilde{a}_0(x, |y|)|y|^2 &\geq -H_0|y|^r && \text{for a.e. } x \in \Omega, \text{ every } |y| \leq 1, \\ f(x, t)t - pF(x, t) &\geq -H_0t^r && \text{for every } t \in [0, 1], \text{ a.e. } x \in \Omega, \\ \lim_{t \rightarrow +0} \frac{f(x, t)t - pF(x, t)}{t^r} &= +\infty && \text{for a.e. } x \in \Omega. \end{aligned}$$

(HF0−) There exist  $p \leq r < p^*$  and  $H_0 > 0$  such that

$$\begin{aligned} p\tilde{G}_0(x, y) - \tilde{a}_0(x, |y|)|y|^2 &\leq H_0|y|^r \quad \text{for a.e. } x \in \Omega, \text{ every } |y| \leq 1, \\ f(x, t)t - pF(x, t) &\leq H_0t^r \quad \text{for every } t \in [0, 1], \text{ a.e. } x \in \Omega, \\ \lim_{t \rightarrow +0} \frac{f(x, t)t - pF(x, t)}{t^r} &= -\infty \quad \text{for a.e. } x \in \Omega. \end{aligned}$$

**Theorem 31.** *Let Assumption (f), (AH0), and (AH) hold. If any of the following conditions is satisfied, (P) has at least one positive solution.*

- (i)  $\alpha_0 > \lambda_1(a_0)$ ,  $\alpha = \lambda_1(a_\infty)$ , and (HF+) or (H+).
- (ii)  $\alpha_0 < \lambda_1(a_0)$ ,  $\alpha = \lambda_1(a_\infty)$ , and (HF−) or (H−).
- (iii)  $\alpha_0 = \lambda_1(a_0)$ ,  $\alpha < \lambda_1(a_\infty)$ , and (HF0+) or (H0+).
- (iv)  $\alpha_0 = \lambda_1(a_0)$ ,  $\alpha > \lambda_1(a_\infty)$ , and (HF0−) or (H0−).
- (v)  $\alpha_0 = \lambda_1(a_0)$ ,  $\alpha = \lambda_1(a_\infty)$ , (HF0+) or (H0+), and (HF+) or (H+).
- (vi)  $\alpha_0 = \lambda_1(a_0)$ ,  $\alpha = \lambda_1(a_\infty)$ , (HF0−) or (H0−), and (HF−) or (H−).

The rest of this section is devoted to the proof of this theorem, which involves some preparatory steps.

*The singly resonant case.* Set  $f_{\pm n}(x, t) := f(x, t) \pm \frac{p}{n}|t|^{p-2}t$  and define approximate functionals on  $W_0^{1,p}(\Omega)$  by

$$I_{\pm n}(u) := \int_{\Omega} G(x, \nabla u) \, dx - \int_{\Omega} (F_{\pm n})_+(x, u) \, dx = I(u) \mp \frac{1}{n} \|u_+\|_p^p.$$

From now on, assume  $f$  satisfies Assumption (f). Take first the case  $\alpha = \lambda_1(a_\infty)$ .

**Lemma 32.** *If either (H+) or (HF+) (resp. either (H−) or (HF−)) hold and  $\{u_n\}$  satisfies*

$$\begin{aligned} \sup_{n \in \mathbb{N}} I_{\pm n}(u_n) < +\infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \|I'_{\pm n}(u_n)\|_{W_0^{1,p}(\Omega)^*} = 0 \\ (\text{resp. } \inf_{n \in \mathbb{N}} I_{\pm n}(u_n) > -\infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \|I'_{\pm n}(u_n)\|_{W_0^{1,p}(\Omega)^*} = 0), \end{aligned}$$

then  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ .

*Proof.* The boundedness of  $\|u_n\|_p$  guarantees that  $\|u_n\|$  is bounded, since

$$o(1)\|u_n\| = \langle I'_{\pm n}(u_n), u_n \rangle \geq \frac{C_0}{p-1} \|u_n\|^p - C(1 + \|u_n\|_p^p) \mp \frac{1}{n} \|(u_n)_+\|_p^p$$

for some  $C > 0$  independent of  $n$ . So, by way of contradiction, we assume that  $\|u_n\|_p \rightarrow \infty$  as  $n \rightarrow \infty$ . Then, by the same argument as in Lemma 28, we see that  $v_n := u_n/\|u_n\|_p$  has a subsequence strongly converging to a positive solution  $v_0$  of

$$(35) \quad -\operatorname{div}(a_\infty |\nabla u|^{p-2} \nabla u) = \alpha |u|^{p-2} u \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega.$$

If  $\alpha \neq \lambda_1(a_\infty)$ , we have a contradiction, because (35) does not have a positive solution except when  $\lambda = \lambda_1(a_\infty)$ . So we may assume that  $\alpha = \lambda_1(a_\infty)$  and  $v_0 = \varphi_{a_\infty}$  (note  $\|v_0\|_p = 1$ ). For simplicity, we still denote the subsequence under discussion by  $\{v_n\}$ . Thus  $u_n(x) \rightarrow \infty$  as  $n \rightarrow \infty$  for a.e.  $x \in \Omega$  (note  $v_0 = \varphi_{a_\infty} > 0$  in  $\Omega$ ).

Assume (HF+) or (HF−). We show that

$$(36) \quad I := \int_{\Omega} \frac{f_+(x, u_n)u_n - pF_+(x, u_n)}{\|u_n\|_p^q} dx \rightarrow \pm\infty,$$

where the sign on  $\infty$  matches (HF±) and  $q$  is a constant as in (HF±). Indeed, it follows from (HF+) that  $(f_+(x, t)t - pF_+(x, t))/t^q$  is bounded from below on  $\Omega \times [1, +\infty)$ . Therefore, since  $u_n(x) \rightarrow \infty$  for a.e.  $x \in \Omega$ , we have (36) if (HF+) holds, by applying Fatou's lemma to the inequality

$$I \geq \int_{u_n(x) \geq 1} \frac{f_+(x, u_n)u_n - pF_+(x, u_n)}{u_n^q} v_n^q dx - \frac{2H_0}{\|u_n\|_p^p} |\Omega|,$$

where  $H_0 > 0$  is a constant as in (HF+). The case of (HF−) is handled by the same argument, with  $-f$  instead of  $f$ . This shows (36).

Furthermore, by Hölder's inequality, we have

$$(37) \quad \begin{aligned} II &:= \int_{\Omega} \frac{p\tilde{G}(x, \nabla u_n) - \tilde{a}(x, |\nabla u_n|)|\nabla u_n|^2}{\|u_n\|_p^q} dx \\ &\leq H_0 \int_{\Omega} (|\nabla v_n|^q + \frac{1}{\|u_n\|_p^q}) dx \leq H_0 \|\nabla v_n\|_p^q |\Omega|^{(p-q)/p} + o(1) \\ &\leq H_0 \|\nabla v_0\|_p^q |\Omega|^{(p-q)/p} + o(1) \end{aligned}$$

in the case of (HF−), because  $v_n \rightarrow v_0$  in  $W_0^{1,p}(\Omega)$ , where  $q \in [1, p)$  and  $H_0 > 0$  are constants as in (HF−). Similarly, we obtain

$$(38) \quad II \geq -H_0 \|\nabla v_0\|_p^q |\Omega|^{(p-q)/p} + o(1)$$

in the case of (HF+).

Hence we have a contradiction because of (36), (37) or (38) by taking the limit inferior or superior in the equality

$$\frac{pI_{\pm n}(u_n) - \langle I'_{\pm n}(u_n), u_n \rangle}{\|u_n\|_p^q} = II + I.$$

Assume (H+) or (H−). Because  $v_0$  is a positive solution of (35), we have  $|\nabla u_n(x)| \rightarrow \infty$  as  $n \rightarrow \infty$  for a.e.  $x \in \Omega_0 := \{x' \in \Omega : |\nabla v_0(x')| \neq 0\}$ . Because  $|\Omega_0| > 0$ , we can show, by an argument similar to the one used for  $f$ , that

$$\int_{\Omega} \frac{p\tilde{G}(x, \nabla u_n) - \tilde{a}(x, |\nabla u_n|)|\nabla u_n|^2}{\|u_n\|_p^q} dx \rightarrow \pm\infty,$$

where again the sign matches that of  $(H_{\pm})$ . In addition, we easily obtain that

$$\pm \int_{\Omega} \frac{f_{\pm}(x, u_n)u_n - pF_{\pm}(x, u_n)}{\|u_n\|_p^q} dx \geq -H_0\|v_n\|_q^q + o(1) = -H_0\|v_0\|_q^q + o(1)$$

(again, the sign matches). Hence we have a contradiction by considering the limit of  $(pI_{\pm n}(u_n) - \langle I'_{\pm n}(u_n), u_n \rangle) / \|u_n\|_p^q$ .  $\square$

*Proof of Theorem 31(i).* Because  $\alpha_0 > \lambda_1(a_0)$ , there exists an  $n_0 \in \mathbb{N}$  such that  $\alpha_0 - p/n_0 > \lambda_1(a_0)$ . Note that  $f_{-n}(x, t)/t^{p-1} \rightarrow \alpha_0 - p/n > \lambda_1(a_0)$  as  $t \rightarrow +0$  for  $n \geq n_0$  and  $f_{-n}(x, t)/t^{p-1} \rightarrow \alpha - p/n = \lambda_1(a_{\infty}) - p/n < \lambda_1(a_{\infty})$  as  $t \rightarrow +\infty$ . Hence, by using the proof of Theorem 25(i) to  $f_{-n}$ , we can find a global minimizer  $u_n$  of  $I_{-n}$  with  $I_{-n}(u_n) < 0$  for each  $n \geq n_0$ . Here we remark that  $\sup_{n \geq n_0} I_{-n}(u_n) < 0$ . In fact, for every  $n \geq n_0$ , we have

$$I_{-n}(u_n) \leq I_{-n}(u_{n_0}) = I(u_{n_0}) + \frac{1}{n}\|u_{n_0}\|_p^p \leq I(u_{n_0}) + \frac{1}{n_0}\|u_{n_0}\|_p^p = I_{-n_0}(u_{n_0}) < 0,$$

where, in the first inequality, we use the fact that  $u_n$  is a global minimizer of  $I_{-n}$ . Now, due to Lemma 32, we see that  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ . Therefore,

$$\|I'(u_n)\|_{W_0^{1,p}(\Omega)^*} = \|I'(u_n) - I'_{-n}(u_n)\|_{W_0^{1,p}(\Omega)^*} \leq \frac{p}{n\lambda_1(-\Delta_p)} \|u_n\|^{p-1} \rightarrow 0$$

as  $n \rightarrow \infty$ , where  $\lambda_1(-\Delta_p)$  is the first eigenvalue of  $-\Delta_p$ . Since  $I$  is bounded on a bounded set, we may assume that  $\{u_n\}$  is a bounded Palais–Smale sequence of  $I$ . Because  $I$  satisfies the bounded Palais–Smale condition (see Proposition 2),  $u_n$  has a subsequence converging to some  $v_0$  in  $W_0^{1,p}(\Omega)$ . It is clear that  $I(v_0) \leq \sup_{n \geq n_0} I_{-n}(u_n) = I_{-n_0}(u_{n_0}) < 0$ , and so  $v_0$  is a nontrivial critical point of  $I$ .  $\square$

*Proof of Theorem 31(ii).* Using Lemma 30 and  $\alpha_0 < \lambda_1(a_0)$ , we can choose  $q_0 \in (p, p^*]$  and  $\rho > 0$  such that  $\inf\{I(u) : \|u\|_{q_0} = \rho\} > 0$ . Since  $I_{+n}(u) \geq I(u) - \|u\|_{q_0}^p |\Omega|^{1-p/q_0}/n$  for every  $u \in W_0^{1,p}(\Omega)$ , we can take  $n_0 \in \mathbb{N}$  such that  $\alpha_0 + p/n_0 < \lambda_1(a_0)$  and  $\delta_0 := \inf\{I_{+n_0}(u) : \|u\|_{q_0} = \rho\} > 0$ . Hence, for every  $n \geq n_0$ , we have  $\inf\{I_{+n}(u) : \|u\|_{q_0} = \rho\} \geq \delta_0$ , because  $I_{+n}(u) \geq I_{+n_0}(u)$  for every  $n \geq n_0$  and  $u \in W_0^{1,p}(\Omega)$ . By noting that  $f_{+n}(x, t)/t^{p-1} \rightarrow \alpha + p/n > \alpha = \lambda_1(a_{\infty})$  as  $t \rightarrow +\infty$ , and applying Lemma 28 to  $f_{+n}$  instead of  $f$ ,  $I_{+n}$  satisfies the Palais–Smale condition. Therefore, the proof of Theorem 25(ii) implies that, for every  $n \geq n_0$ , there exists a critical point  $u_n \in W_0^{1,p}(\Omega)$  of  $I_{+n}$  such that  $I_{+n}(u_n) \geq \delta_0$ . According to Lemma 32,  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ . Thus, because we have a bounded Palais–Smale sequence of  $I$  due to a similar reason as in the case of (i), we can obtain a nontrivial critical point of  $I$  (note that  $\inf_{n \geq n_0} I(u_n) \geq \inf_{n \geq n_0} I_{+n}(u_n) \geq \delta_0 > 0$ ).  $\square$

We next turn to the case where  $\alpha_0 = \lambda_1(a_0)$ .

**Lemma 33.** *Assume (H0−) or (HF0−) (resp. (H0+) or (HF0+)). Let  $u_n \neq 0$  be an element of  $W_0^{1,p}(\Omega)$  satisfying  $I'_{\pm n}(u_n) = 0$  for every  $n \in \mathbb{N}$  and  $\inf_n I_{\pm n}(u_n) \geq 0$  (resp.  $\sup_n I_{\pm n}(u_n) \leq 0$ ). Then  $\liminf_{n \rightarrow \infty} \|u_n\|_p > 0$ .*

*Proof.* By way of contradiction, we assume that  $\lim_{n \rightarrow \infty} \|u_n\|_p = 0$  by choosing a subsequence. Note that the boundedness of  $\|u_n\|_p$  yields that  $\|u_n\|$  and  $\|u_n\|/\|u_n\|_p$  are bounded in view of

$$(39) \quad o(1)\|u_n\| = \langle I'_{\pm n}(u_n), u_n \rangle \geq \frac{C_0}{p-1} \|u_n\|^p - C(1 + \|(u_n)_+\|_p^p) \mp \frac{p}{n} \|(u_n)_+\|_p^p$$

for some  $C > 0$  independent of  $n$ . Then, since  $u_n$  is a positive solution of

$$-\operatorname{div}(A(x, \nabla u)) = f_{\pm n}(x, u_n) \quad \text{in } \Omega$$

(refer to Remarks 3 and 27), it follows from Proposition 4 that  $u_n \rightarrow 0$  in  $C^1(\bar{\Omega})$  (note that  $|(f_{\pm n})_+(x, t)| \leq Ct_+^{p-1}$  (see Assumption (f)) and  $u_n \rightarrow 0$  in  $L^p(\Omega)$ ). Therefore, we may assume that  $\|u_n\|_{C^1(\bar{\Omega})} \leq 1$  by considering a sufficiently large  $n$ . Since  $|f_{\pm n}(x, \|u_n\|_p t)/\|u_n\|_p^{p-1}| \leq Ct^p$  for every  $t \geq 0$ , a.e.  $x \in \Omega$  ( $C > 0$  independent of  $n$ ; see Assumption (f) and (39)), by a similar argument to Theorem 13, we see that  $v_n := u_n/\|u_n\|_p$  has a subsequence converging to a positive solution  $v_0$  in  $C^1(\bar{\Omega})$  of

$$(40) \quad -\operatorname{div}(a_0(x)|\nabla u|^{p-2}\nabla u) = \alpha_0|u|^{p-2}u \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega.$$

If  $\alpha_0 \neq \lambda_1(a_0)$ , we have a contradiction because (13) does not have a positive solution unless  $\lambda = \lambda_1(a_0)$ . So we may assume that  $\alpha_0 = \lambda_1(a_0)$  and  $v_0 = \varphi_{a_0}$  (note  $\|v_0\|_p = 1$ ). For simplicity, we still denote the subsequence under discussion by  $\{v_n\}$ .

Assume (H0+) or (H0−). Then we can prove that

$$(41) \quad I := \int_{\Omega} \frac{p\tilde{G}_0(x, \nabla u_n) - \tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2}{\|u_n\|_p^r} dx \rightarrow \pm\infty$$

(signs match), where  $r \in [p, p^*)$  is a constant as in (H0+) or (H0−). Indeed, because  $\|\nabla v_0\|_p > 0$ , we can choose a constant  $\varepsilon_0 > 0$  such that  $|\{x \in \Omega : |\nabla v_0| > 2\varepsilon_0\}| > 0$ . With this  $\varepsilon_0$ , we have under assumption (H0+)

$$\begin{aligned} I &\geq \int_{|\nabla v_n| > \varepsilon_0} \frac{p\tilde{G}_0(x, \nabla u_n) - \tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2}{|\nabla u_n|^r} |\nabla v_n|^r dx - \int_{|\nabla v_n| \leq \varepsilon_0} H_0 |\nabla v_n|^r dx \\ &\geq \int_{|\nabla v_n| > \varepsilon_0} \frac{p\tilde{G}_0(x, \nabla u_n) - \tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2}{|\nabla u_n|^r} |\nabla v_n|^r dx - \varepsilon_0^r H_0 |\Omega|, \end{aligned}$$

where  $H_0$  is a positive constant as in (H0+). Hence, applying Fatou's lemma, our claim is shown, because the Lebesgue measure of  $\{x \in \Omega : |\nabla v_0| > 2\varepsilon_0\}$  is positive. Similarly, by considering  $\tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2 - p\tilde{G}_0(x, \nabla u_n)$ , we can prove (41) under (H0−).

On the other hand, by using (H0+) or (H0−), we obtain

$$(42) \quad \begin{aligned} \pm II &:= \pm \int_{\Omega} \frac{f_+(x, u_n)u_n - pF_+(x, u_n)}{\|u_n\|_p^r} dx \geq -H_0 \int_{\Omega} (v_n)_+^r dx \\ &\geq -H_0 \|v_n\|_r^r = -H_0 \|v_0\|_r^r + o(1) \end{aligned}$$

(note that  $\|u_n\|_{C^1(\bar{\Omega})} \leq 1$  and  $v_n \rightarrow v_0$  in  $C^1(\bar{\Omega})$ ). Now set  $\Psi_n = I_{\pm n}$ . Since

$$(43) \quad \pm(I + II) = \pm \frac{p\Psi_n(u_n) - \langle \Psi_n'(u_n), u_n \rangle}{\|u_n\|_p^r} = \pm \frac{p\Psi_n(u_n)}{\|u_n\|_p^r} \leq 0$$

if  $\sup_n (\pm I_{\pm}(u_n)) \leq 0$  (where the signs match throughout), we obtain a contradiction with (41) and (42) by taking the limit superior or inferior in (43).

Assume (HF0+) or (HF0−). As in the argument for  $I$  in the case of (H0±), we can show that

$$\int_{\Omega} \frac{f_+(x, u_n)u_n - pF_+(x, u_n)}{\|u_n\|_p^r} dx = \int_{v_n > 0} \frac{f_+(x, u_n)u_n - pF_+(x, u_n)}{(u_n)_+^r} (v_n)_+^r dx \rightarrow \pm\infty,$$

the sign matching that of (HF0±). Moreover, it is easily seen that

$$\pm \int_{\Omega} \frac{p\tilde{G}_0(x, \nabla u_n) - \tilde{a}_0(x, |\nabla u_n|)|\nabla u_n|^2}{\|u_n\|_p^r} dx \geq \mp H_0 \|\nabla v_n\|_r^r = \mp H_0 \|\nabla v_0\|_r^r + o(1).$$

(Note that  $\|u_n\|_{C^1(\bar{\Omega})} \leq 1$  and  $v_n \rightarrow v_0$  in  $C^1(\bar{\Omega})$ .) Our conclusion follows from a similar argument as before.  $\square$

*Proof of Theorem 31(iii).* Let  $n_0 \in \mathbb{N}$  such that  $\alpha + p/n_0 < \lambda_1(a_\infty)$ . The proof of Theorem 25(i) guarantees that, for every  $n \geq n_0$ ,  $I_{+n}$  has a global minimizer  $u_n$  such that  $I_{+n}(u_n) < 0$ , because  $f_{+n}(x, t)/t^{p-1} \rightarrow \alpha + p/n > \alpha = \lambda_1(a_0)$  as  $t \rightarrow +0$  and  $f_{+n}(x, t)/t^{p-1} \rightarrow \alpha + p/n < \lambda_1(a_\infty)$  as  $t \rightarrow +\infty$  if  $n \geq n_0$ . Noting that  $I_{+n}(u) \geq I_{+n_0}(u)$  for every  $u \in W_0^{1,p}(\Omega)$  and  $n \geq n_0$ ,  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$  since  $I_{+n_0}$  is coercive on  $W_0^{1,p}(\Omega)$  by Lemma 29. Thus  $\{u_n\}$  is a bounded Palais–Smale sequence of  $I$  by the same argument as in (i). Therefore,  $\{u_n\}$  has a convergent subsequence to some  $u_0$  in  $W_0^{1,p}(\Omega)$  because  $I$  satisfies the bounded Palais–Smale condition. On the other hand, Lemma 33 guarantees that  $u_0 \neq 0$  (note  $\sup_{n \geq n_0} I_{+n}(u_n) \leq 0$ ). Therefore  $u_0$  is a nontrivial critical point of  $I$ .  $\square$

*Proof of Theorem 31(iv).* Let  $n_0 \in \mathbb{N}$  be such that  $\alpha - p/n_0 > \lambda_1(a_\infty)$ . Applying Lemma 30 to  $f_{-n}$  for  $n \geq n_0$  (and since  $\alpha_0 - p/n < \lambda_1(a_0)$ ), we can choose  $q_0 \in (p, p^*]$  and  $\rho_n > 0$  such that  $\delta_n := \inf\{I_{-n}(u) : \|u\|_{q_0} = \rho_n\} > 0$ . By noting that  $f_{-n}(x, t)/t^{p-1} \rightarrow \alpha - p/n > \lambda_1(a_\infty)$  as  $t \rightarrow +\infty$  for every  $n \geq n_0$ , and applying Lemma 28 to  $f_{-n}$  instead of  $f$ , we see that  $I_{-n}$  satisfies the Palais–Smale condition. Therefore, the proof of Theorem 25(ii) implies that, for every  $n \geq n_0$ , there exists

a critical point  $u_n \in W_0^{1,p}(\Omega)$  of  $I_{-n}$  such that  $I_{-n}(u_n) \geq \delta_n > 0$ . By Lemma 32,  $\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ . Thus, by arguing as in case (i), we find a subsequence  $\{u_n\}$  converging to some  $u_0$  in  $W_0^{1,p}(\Omega)$ . Also, Lemma 33 yields  $u_0 \neq 0$  (note that  $\inf_{n \geq n_0} I_{-n}(u_n) \geq 0$ ). This shows that  $u_0$  is a nontrivial critical point of  $I$ .  $\square$

*The doubly resonant case.* Choose smooth nonnegative functions  $\varphi$  and  $\psi$  on  $[0, +\infty)$  satisfying  $\varphi(t) = 1$  if  $0 \leq t \leq 2$ ,  $\varphi(t) = 0$  if  $t \geq 4$ ,  $\psi(t) = 0$  if  $t \leq 5$ , and  $\psi(t) = 1$  if  $t \geq 10$ . Define approximate functionals on  $W_0^{1,p}(\Omega)$  by

$$\tilde{I}_{\pm n}(u) := I(u) \mp \frac{1}{n} \psi(\|u\|_p^p) \|u_+\|_p^p \pm \frac{1}{n} \varphi(\|u\|_p^p) \|u_+\|_p^p.$$

Because  $\tilde{I}_{\pm n}(u) = I_{\mp n}(u)$  provided  $\|u\|_p \leq 2$ , the following result can be proved by the same argument as in Lemma 33. We omit the proof.

**Lemma 34.** *Assume (H0−) or (HF0−) (resp. (H0+) or (HF0+)). Let  $u_n \neq 0$  be an element of  $W_0^{1,p}(\Omega)$  satisfying  $(\tilde{I}_{\pm n})'(u_n) = 0$  for every  $n \in \mathbb{N}$  and  $\inf_n \tilde{I}_{\pm n}(u_n) \geq 0$  (resp.  $\sup_n \tilde{I}_{\pm n}(u_n) \leq 0$ ). Then  $\liminf_{n \rightarrow \infty} \|u_n\|_p > 0$ .*

**Lemma 35.** *If  $\alpha \pm p/n \neq \lambda_1(a_\infty)$ , then  $\tilde{I}_{\pm n}$  (with the matching sign) satisfies the Palais–Smale condition.*

*Proof.* Let  $\{u_m\}$  be a Palais–Smale sequence of  $\tilde{I}_{+n}$  or  $\tilde{I}_{-n}$ . If  $\|u_m\|_p \rightarrow \infty$  occurs, then  $\tilde{I}_{\pm n}(u_m) = I_{\pm n}(u_m)$  for sufficiently large  $m$ . So, by applying Lemma 28 to  $f_{\pm n}$  (note that  $\alpha \pm p/n \neq \lambda_1(a_\infty)$ ), we have a contradiction if  $\|u_m\|_p \rightarrow \infty$ . Consequently, we see that  $\|u_m\|_p$  is bounded. Then, by the same reason as in Lemma 28,  $\{u_m\}$  has a convergent subsequence in  $W_0^{1,p}(\Omega)$ .  $\square$

Because  $\tilde{I}_{\pm n}(u) = I_{\pm n}(u)$  provided  $\|u\|_p \geq 10$ , the following result can be proved by the same argument as in Lemma 32. We omit the proof.

**Lemma 36.** *If either (H+) or (HF+) (resp. either (H−) or (HF−)) and  $\{u_n\}$  satisfies*

$$\begin{aligned} \sup_{n \in \mathbb{N}} \tilde{I}_{\pm n}(u_n) < +\infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \|(\tilde{I}_{\pm n})'(u_n)\|_{W_0^{1,p}(\Omega)^*} = 0 \\ (\text{resp. } \inf_{n \in \mathbb{N}} \tilde{I}_{\pm n}(u_n) > -\infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \|(\tilde{I}_{\pm n})'(u_n)\|_{W_0^{1,p}(\Omega)^*} = 0), \end{aligned}$$

*$\{u_n\}$  is bounded in  $W_0^{1,p}(\Omega)$ .*

*Proof of Theorem 31(v).* Note that  $\tilde{I}_{-n}(u) = I_{-n}(u)$  provided  $\|u\|_p \geq 10$  and  $\tilde{I}_{-n}(u) = I_{+n}(u)$  if  $\|u\|_p \leq 2$ . So, by a similar argument to that in (i),  $\tilde{I}_{-n}$  has a global minimizer  $u_n$ . Moreover, by a similar argument to that in (iii) (note that  $f_{+n}(x, t)/t^{p-1} \rightarrow \alpha_0 + p/n > \lambda_1(a_0)$  as  $t \rightarrow +0$  and  $f_{-n}(x, t)/t^{p-1} \rightarrow \alpha - p/n < \lambda_1(a_\infty)$  as  $t \rightarrow +\infty$ ), we have  $\tilde{I}_{-n}(u_n) < 0$ , whence  $u_n \neq 0$ . Because Lemma 36 implies the boundedness of  $\|u_n\|$ , by the same argument as in (i), we see that  $\{u_n\}$



is a bounded Palais–Smale sequence of  $I$ . Therefore, we may assume that  $u_n$  converges to some  $u_0$  in  $W_0^{1,p}(\Omega)$  by choosing a subsequence. On the other hand, Lemma 33 yields  $\liminf_{n \rightarrow \infty} \|u_n\|_p > 0$ . Hence  $u_0 \neq 0$ . This means that  $u_0$  is a nontrivial critical point of  $I$ .  $\square$

*Proof of Theorem 31(vi).* Note that  $\tilde{I}_{+n}(u) = I_{+n}(u)$  provided  $\|u\|_p \geq 10$  and  $\tilde{I}_{+n}(u) = I_{-n}(u)$  if  $\|u\|_p \leq 2$ . So, because  $f_{-n}(x, t)/t^{p-1} \rightarrow \alpha_0 - p/n < \lambda_1(a_0)$  as  $t \rightarrow +0$  and  $f_{+n}(x, t)/t^{p-1} \rightarrow \alpha + p/n > \lambda_1(a_\infty)$  as  $t \rightarrow +\infty$ , by a similar argument to those in (ii) and (iv), for each  $n$ , we have a nontrivial critical point  $u_n$  of  $\tilde{I}_{+n}$  with  $\tilde{I}_{+n}(u_n) > 0$ . As a result, by a similar reasoning as in (v), we can obtain a nontrivial critical point of  $I$ .  $\square$

### Acknowledgements

The second author would like to express her sincere thanks to Professor Shizuo Miyajima for helpful comments and encouragement.

### References

- [Anane 1987] A. Anane, *Etude des valeurs propres et de la résonance pour l'opérateur  $p$ -Laplacien*, Ph.D. thesis, Université Libre de Bruxelles, 1987.
- [Cuesta et al. 1999] M. Cuesta, D. de Figueiredo, and J.-P. Gossez, “The beginning of the Fučík spectrum for the  $p$ -Laplacian”, *J. Differential Equations* **159**:1 (1999), 212–238. MR 2001f:35308 Zbl 0947.35068
- [Damascelli 1998] L. Damascelli, “Comparison theorems for some quasilinear degenerate elliptic operators and applications to symmetry and monotonicity results”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **15**:4 (1998), 493–516. MR 99e:35081 Zbl 0911.35009
- [Dancer and Perera 2001] N. Dancer and K. Perera, “Some remarks on the Fučík spectrum of the  $p$ -Laplacian and critical groups”, *J. Math. Anal. Appl.* **254**:1 (2001), 164–177. MR 2001k:35238 Zbl 0970.35056
- [Deimling 1985] K. Deimling, *Nonlinear functional analysis*, Springer, Berlin, 1985. MR 86j:47001 Zbl 0559.47040
- [Fukagai and Narukawa 2007] N. Fukagai and K. Narukawa, “On the existence of multiple positive solutions of quasilinear elliptic eigenvalue problems”, *Ann. Mat. Pura Appl.* (4) **186**:3 (2007), 539–564. MR 2008d:35052 Zbl 1223.35132
- [García-Huidobro et al. 1995] M. García-Huidobro, R. Manásevich, and P. Ubilla, “Existence of positive solutions for some Dirichlet problems with an asymptotically homogeneous operator”, *Electron. J. Differential Equations* **1995**:10 (1995), 1–22. MR 96f:35053 Zbl 0823.35057
- [Kim 2009] Y.-H. Kim, “A global bifurcation for nonlinear equations with nonhomogeneous part”, *Nonlinear Anal.* **71**:12 (2009), 738–743. MR 2011d:35048 Zbl 1238.35009
- [Kim and Kim 2010] I.-S. Kim and Y.-H. Kim, “Global bifurcation for equations involving non-homogeneous operators in an unbounded domain”, *Nonlinear Anal.* **73**:4 (2010), 1057–1064. MR 2011e:35023 Zbl 1194.35492
- [Kyritsi et al. 2010] S. T. Kyritsi, D. O’Regan, and N. S. Papageorgiou, “Existence of multiple solutions for nonlinear Dirichlet problems with a nonhomogeneous differential operator”, *Adv. Nonlinear Stud.* **10**:3 (2010), 631–657. MR 2011e:35116 Zbl 1216.35037

- [Lieberman 1988] G. M. Lieberman, “Boundary regularity for solutions of degenerate elliptic equations”, *Nonlinear Anal.* **12**:11 (1988), 1203–1219. MR 90a:35098 Zbl 0675.35042
- [Mawhin and Willem 1989] J. Mawhin and M. Willem, *Critical point theory and Hamiltonian systems*, Applied Mathematical Sciences **74**, Springer, New York, 1989. MR 90e:58016 Zbl 0676.58017
- [Miyajima et al. 2012] S. Miyajima, D. Motreanu, and M. Tanaka, “Multiple existence results of solutions for the Neumann problems via super- and sub-solutions”, *J. Funct. Anal.* **262**:4 (2012), 1921–1953. MR 2873865 Zbl pre06012152
- [Motreanu and Papageorgiou 2011] D. Motreanu and N. S. Papageorgiou, “Multiple solutions for nonlinear Neumann problems driven by a nonhomogeneous differential operator”, *Proc. Amer. Math. Soc.* **139**:10 (2011), 3527–3535. MR 2012c:35166 Zbl 1226.35021
- [Motreanu et al. 2011] D. Motreanu, V. V. Motreanu, and N. S. Papageorgiou, “Multiple constant sign and nodal solutions for nonlinear Neumann eigenvalue problems”, *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)* **10**:3 (2011), 729–755. MR 2012m:35087 Zbl 1234.35169
- [Prado and Ubilla 1998] H. Prado and P. Ubilla, “Existence of nonnegative solutions for generalized  $p$ -Laplacians”, pp. 289–298 in *Reaction diffusion systems* (Trieste, 1995), edited by G. Caristi and E. Mitidieri, Lecture Notes in Pure and Appl. Math. **194**, Dekker, New York, 1998. MR 98j:35069 Zbl 0913.35044
- [Robinson 2004] S. B. Robinson, “On the second eigenvalue for nonhomogeneous quasi-linear operators”, *SIAM J. Math. Anal.* **35**:5 (2004), 1241–1249. MR 2005a:35219 Zbl 1061.35071
- [Tanaka 2009] M. Tanaka, “Existence of constant sign solutions for the  $p$ -Laplacian problems in the resonant case with respect to Fučík spectrum”, *SUT J. Math.* **45**:2 (2009), 149–166. MR 2011d:35207 Zbl 1196.35115
- [Tanaka 2012a] M. Tanaka, “The antimaximum principle and the existence of a solution for the generalized  $p$ -Laplace equations with indefinite weight”, *Differ. Equ. Appl.* **4**:4 (2012), 581–613.
- [Tanaka 2012b] M. Tanaka, “Existence of the Fučík type spectrums for the generalized  $p$ -Laplace operators”, *Nonlinear Anal.* **75**:7 (2012), 3407–3435. MR 2891177 Zbl 1241.35091

Received June 19, 2012.

DUMITRU MOTREANU  
DÉPARTAMENT DE MATHÉMATIQUES  
UNIVERSITÉ DE PERPIGNAN  
52 AVENUE PAUL ALDUY  
66860 PERPIGNAN  
FRANCE  
motreanu@univ-perp.fr

MIEKO TANAKA  
DEPARTMENT OF MATHEMATICS  
TOKYO UNIVERSITY OF SCIENCE  
KAGURAZAKA 1-3  
SHINJYUKU-KU  
TOKYO 162-8601  
JAPAN  
tanaka@ma.kagu.tus.ac.jp

## WEIGHTED RICCI CURVATURE ESTIMATES FOR HILBERT AND FUNK GEOMETRIES

SHIN-ICHI OHTA

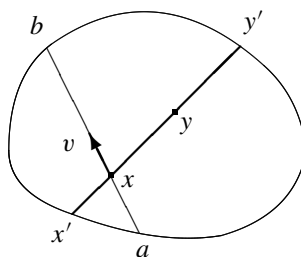
**We consider Hilbert and Funk geometries on a strongly convex domain in Euclidean space. We show that, with respect to the Lebesgue measure on the domain, the Hilbert and Funk metrics have bounded and constant negative weighted Ricci curvature, respectively. As a corollary, these metric measure spaces satisfy the curvature-dimension condition in the sense of Lott, Sturm and Villani.**

### 1. Introduction

Hilbert [1895] introduced the distance function  $d_{\mathcal{H}}$  on a bounded convex domain  $D \subset \mathbb{R}^n$ , related to his fourth problem. Given distinct points  $x, y \in D$ , denoting by  $x' = x + s(y - x)$  and  $y' = x + t(y - x)$  the intersections of the boundary  $\partial D$  and the line passing through  $x$  and  $y$  with  $s < 0 < t$  (see figure), Hilbert's distance  $d_{\mathcal{H}}$  is given by

$$d_{\mathcal{H}}(x, y) = \frac{1}{2} \log \frac{|x' - y| \cdot |x - y'|}{|x' - x| \cdot |y - y'|},$$

where  $|\cdot|$  stands for the Euclidean norm. This is indeed a distance function on  $D$ , and satisfies the interesting property that line segments between any points are minimizing. In the particular case where  $D$  is the unit ball,  $(D, d_{\mathcal{H}})$  coincides with the Klein model of hyperbolic space. The structure of  $(D, d_{\mathcal{H}})$  has been investigated




---

Supported in part by the Grant-in-Aid for Young Scientists (B) 23740048.

*MSC2010:* primary 53C60; secondary 53C23.

*Keywords:* Hilbert geometry, Funk geometry, Ricci curvature, curvature-dimension condition.

from geometric and dynamical aspects; see, for example, [Egloff 1997; Benoist 2003; Colbois and Verovic 2004]. For instance,  $(D, d_{\mathcal{H}})$  is known to be Gromov hyperbolic under mild smoothness and convexity assumptions on  $D$ .

Funk [1929] introduced a nonsymmetrization of  $d_{\mathcal{H}}$ , namely

$$d_{\mathcal{F}}(x, y) = \log \frac{|x - y'|}{|y - y'|}.$$

Note that  $d_{\mathcal{F}}(x, y) \neq d_{\mathcal{F}}(y, x)$ , while the triangle inequality

$$d_{\mathcal{F}}(x, z) \leq d_{\mathcal{F}}(x, y) + d_{\mathcal{F}}(y, z)$$

still holds. Clearly we have  $2d_{\mathcal{H}}(x, y) = d_{\mathcal{F}}(x, y) + d_{\mathcal{F}}(y, x)$ , and line segments are minimizing also with respect to Funk's distance.

If  $\partial D$  is smooth and  $D$  is strongly convex (in other words,  $\partial D$  is positively curved; see Definition 2.1), then  $d_{\mathcal{H}}$  and  $d_{\mathcal{F}}$  are realized by the smooth Finsler structures

$$(1-1) \quad \begin{aligned} F_{\mathcal{H}}(x, v) &= \frac{|v|}{2} \left( \frac{1}{|x-a|} + \frac{1}{|x-b|} \right), \\ F_{\mathcal{F}}(x, v) &= \frac{|v|}{|x-b|} \quad \text{for } v \in T_x D = \mathbb{R}^n, \end{aligned}$$

respectively (cf. [Shen 2001a, §2.3]), where  $a = x + sv$  and  $b = x + tv$  denote the intersections of  $\partial D$  and the line passing through  $x$  in the direction  $v$  with  $s < 0 < t$  (see figure on page 185). Note that  $2F_{\mathcal{H}}(x, v) = F_{\mathcal{F}}(x, v) + F_{\mathcal{F}}(x, -v)$ . A remarkable feature of these metrics is that they have the constant negative flag curvatures  $-1$  and  $-\frac{1}{4}$ , respectively; see [Okada 1983, Theorem 1; Shen 2001a, Theorem 12.2.11], provided that  $n \geq 2$  as a matter of course. The flag curvature is a generalization of the sectional curvature in Riemannian geometry, so it is natural that  $(D, d_{\mathcal{H}})$  and  $(D, d_{\mathcal{F}})$  enjoy properties of negatively curved spaces.

Recently, the theory of the *weighted Ricci curvature* (see Definition 2.2) for Finsler manifolds equipped with arbitrary measures has been developed in connection with optimal transport theory. It turned out that the weighted Ricci curvature is a natural quantity and quite useful in the study of geometry and analysis on Finsler manifolds; see [Ohta 2009a; 2012; Ohta and Sturm 2009; 2011]. The aim of this article is to show that the weighted Ricci curvature for Hilbert and Funk geometries admits uniform bounds with respect to the Lebesgue measure  $m_L$  restricted on  $D$ .

**Theorem 1.1** (Funk case). *Let  $D \subset \mathbb{R}^n$  with  $n \geq 2$  be a strongly convex domain such that  $\partial D$  is smooth. Then  $(D, F_{\mathcal{F}}, m_L)$  has constant negative weighted Ricci curvature: specifically, for any unit vector  $v \in TD$ ,*

$$\text{Ric}_{\infty}(v) = -\frac{n-1}{4}, \quad \text{Ric}_N(v) = -\frac{n-1}{4} - \frac{(n+1)^2}{4(N-n)} \quad \text{for } N \in (n, \infty).$$

**Theorem 1.2** (Hilbert case). *Let  $D \subset \mathbb{R}^n$  with  $n \geq 2$  be a strongly convex domain such that  $\partial D$  is smooth. Then the weighted Ricci curvature of  $(D, F_{\mathcal{H}}, m_L)$  is bounded; specifically, for any unit vector  $v \in TD$ ,*

$$\text{Ric}_{\infty}(v) \in (-(n-1), 2], \quad \text{Ric}_N(v) \in \left( -(n-1) - \frac{(n+1)^2}{N-n}, 2 \right] \quad \text{for } N \in (n, \infty).$$

We stress that our estimates are independent of the choice of the domain  $D$ . There are several applications (Corollaries 5.1, 5.2) via the theory of the weighted Ricci curvature.

The article is organized as follows. After preliminaries for Finsler geometry and the weighted Ricci curvature, we prove Theorem 1.1 in Section 3 and Theorem 1.2 in Section 4. We finally discuss applications and remarks in Section 5.

## 2. Preliminaries

We very briefly review the necessary notions in Finsler geometry; we refer to [Bao et al. 2000; Shen 2001a; 2001b] for further reading. Let  $M$  be a connected,  $n$ -dimensional  $\mathcal{C}^\infty$ -manifold without boundary such that  $n \geq 2$ . Given a local coordinate  $(x^i)_{i=1}^n$  on an open set  $\Omega \subset M$ , we always use the coordinate  $(x^i, v^j)_{i,j=1}^n$  of  $T\Omega$  such that

$$v = \sum_{j=1}^n v^j \frac{\partial}{\partial x^j} \Big|_x \in T_x M \quad \text{for } x \in \Omega.$$

**Definition 2.1** (Finsler structures). A nonnegative function  $F : TM \rightarrow [0, \infty)$  is called a  $\mathcal{C}^\infty$ -Finsler structure of  $M$  if the following three conditions hold.

- (1) (*Regularity*)  $F$  is  $\mathcal{C}^\infty$  on  $TM \setminus 0$ , where  $0$  stands for the zero section.
- (2) (*Positive 1-homogeneity*) It holds  $F(cv) = cF(v)$  for all  $v \in TM$  and  $c > 0$ .
- (3) (*Strong convexity*) The  $n \times n$  matrix

$$(2-1) \quad (g_{ij}(v))_{i,j=1}^n := \left( \frac{1}{2} \frac{\partial^2(F^2)}{\partial v^i \partial v^j}(v) \right)_{i,j=1}^n$$

is positive definite for all  $v \in TM \setminus 0$ .

For  $x, y \in M$ , we can define the *distance* from  $x$  to  $y$  in a natural way by

$$d(x, y) := \inf_{\eta} \int_0^1 F(\dot{\eta}(t)) dt,$$

where the infimum is taken over all  $\mathcal{C}^1$ -curves  $\eta : [0, 1] \rightarrow M$  with  $\eta(0) = x$  and  $\eta(1) = y$ . This distance can be *nonsymmetric* (namely  $d(y, x) \neq d(x, y)$ ), since  $F$  is only positively homogeneous. A  $\mathcal{C}^\infty$ -curve  $\eta$  on  $M$  is called a *geodesic* if it is locally minimizing and has a constant speed (i.e.,  $F(\dot{\eta})$  is constant).

Given  $v \in T_x M$ , if there is a geodesic  $\eta : [0, 1] \rightarrow M$  with  $\dot{\eta}(0) = v$ , then we define the *exponential map* by  $\exp_x(v) := \eta(1)$ . We say that  $(M, F)$  is *forward complete* if the exponential map is defined on whole  $TM$ . If the *reverse* Finsler manifold  $(M, \overleftarrow{F})$  with  $\overleftarrow{F}(v) := F(-v)$  is forward complete, then  $(M, F)$  is said to be *backward complete*. We remark that  $(D, F_{\mathcal{F}})$  is both forward and backward complete (they are indeed equivalent since  $\overleftarrow{F}_{\mathcal{F}} = F_{\mathcal{F}}$ ), while  $(D, F_{\mathcal{F}})$  is only forward complete.

For each  $v \in T_x M \setminus 0$ , the positive definite matrix  $(g_{ij}(v))_{i,j=1}^n$  in (2-1) induces the Riemannian structure  $g_v$  of  $T_x M$  as

$$(2-2) \quad g_v \left( \sum_{i=1}^n a_i \frac{\partial}{\partial x^i} \Big|_x, \sum_{j=1}^n b_j \frac{\partial}{\partial x^j} \Big|_x \right) := \sum_{i,j=1}^n a_i b_j g_{ij}(v).$$

Note that  $g_{cv} = g_v$  for  $c > 0$ . This inner product is regarded as the best Riemannian approximation of  $F|_{T_x M}$  in the direction  $v$ , in the sense that the unit sphere of  $g_v$  is tangent to that of  $F|_{T_x M}$  at  $v/F(v)$  up to the second order. In particular, we have  $g_v(v, v) = F(v)^2$ .

The *Ricci curvature* (as the trace of the *flag curvature*) for a Finsler manifold is defined by using the Chern connection. Instead of giving the precise definition in coordinates, we explain a useful interpretation due to Shen [2001b, §6.2; 1997, Lemma 2.4]. Given a unit vector  $v \in T_x M \cap F^{-1}(1)$ , we extend it to a nonvanishing  $\mathcal{C}^\infty$ -vector field  $V$  on a neighborhood of  $x$  in such a way that every integral curve of  $V$  is geodesic, and consider the Riemannian structure  $g_v$  induced from (2-2). Then the Ricci curvature  $\text{Ric}(v)$  of  $v$  with respect to  $F$  coincides with the Ricci curvature of  $v$  with respect to  $g_v$  (in particular, it is independent of the choice of  $V$ ).

Let us fix a positive  $\mathcal{C}^\infty$ -measure  $m$  on  $M$ . Inspired by the above interpretation of the Finsler Ricci curvature and the theory of weighted Riemannian manifolds, the weighted Ricci curvature for the triple  $(M, F, m)$  was introduced in [Ohta 2009a] as follows.

**Definition 2.2** (weighted Ricci curvature). Given a unit vector  $v \in T_x M \cap F^{-1}(1)$ , let  $\eta : (-\varepsilon, \varepsilon) \rightarrow M$  be the geodesic such that  $\dot{\eta}(0) = v$ . We decompose  $m$  along  $\eta$  using the Riemannian volume measure  $\text{vol}_{g_\eta}$  of  $g_\eta$  as  $m = e^{-\Psi} \text{vol}_{g_\eta}$ , where  $\Psi : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}$ . Then we define the *weighted Ricci curvature* involving a parameter  $N \in [n, \infty]$  by

- (1)  $\text{Ric}_n(v) := \begin{cases} \text{Ric}(v) + \Psi''(0) & \text{if } \Psi'(0) = 0, \\ -\infty & \text{if } \Psi'(0) \neq 0, \end{cases}$
- (2)  $\text{Ric}_N(v) := \text{Ric}(v) + \Psi''(0) - \frac{\Psi'(0)^2}{N-n}$  for  $N \in (n, \infty)$ ,
- (3)  $\text{Ric}_\infty(v) := \text{Ric}(v) + \Psi''(0)$ .

We also set  $\text{Ric}_N(cv) := c^2 \text{Ric}_N(v)$  for  $c \geq 0$ .

We will say that  $\text{Ric}_N \geq K$  holds for some  $K \in \mathbb{R}$  if  $\text{Ric}_N(v) \geq KF(v)^2$  for all  $v \in TM$ . Observe that  $\text{Ric}_N(v) \leq \text{Ric}_{N'}(v)$  for  $N < N'$ , and that for the scaled space  $M' = (M, F, am)$  with  $a > 0$  we have  $\text{Ric}_N^{M'}(v) = \text{Ric}_N^M(v)$ . It was shown in [Ohta 2009a, Theorem 1.2] that  $\text{Ric}_N \geq K$  is equivalent to Lott, Sturm and Villani’s *curvature-dimension condition*  $\text{CD}(K, N)$ . (Roughly speaking, the curvature-dimension condition is a convexity condition of an entropy functional on the space of probability measures; we refer to [Sturm 2006a; 2006b; Lott and Villani 2007; 2009; Villani 2009, Part III] for details and further theories.) This equivalence extends the corresponding result on (weighted) Riemannian manifolds, and has many analytic and geometric applications; see [Ohta 2009a].

### 3. The Funk case

We turn to the proof of Theorem 1.1. For brevity, we denote the Funk metric simply by  $F$ , and we consider the standard coordinate of  $D \subset \mathbb{R}^n$ . The following lemma enables us to translate all the vertical derivatives ( $\partial/\partial v^i$ ) into horizontal derivatives ( $\partial/\partial x^i$ ).

**Lemma 3.1** [Okada 1983, Proposition 1; Shen 2001a, Lemma 2.3.1]. *For any  $v \in TD \setminus 0$  and  $i = 1, 2, \dots, n$ , we have*

$$\frac{\partial F}{\partial x^i}(v) = F(v) \frac{\partial F}{\partial v^i}(v).$$

*Proof of Theorem 1.1.* On  $TD \setminus 0$ ,

$$(3-1) \quad \frac{1}{2} \frac{\partial^2(F^2)}{\partial v^i \partial v^j} = \frac{\partial}{\partial v^i} \left( \frac{\partial F}{\partial x^j} \right) = \frac{\partial}{\partial x^j} \left( \frac{1}{F} \frac{\partial F}{\partial x^i} \right) = \frac{1}{F} \frac{\partial^2 F}{\partial x^i \partial x^j} - \frac{1}{F^2} \frac{\partial F}{\partial x^i} \frac{\partial F}{\partial x^j}.$$

Now, we fix a unit vector  $v \in T_x D \cap F^{-1}(1)$  and choose a coordinate such that  $x$  is the origin,  $v = \partial/\partial x^n$  and  $g_{in}(v) = 0$  for all  $i = 1, 2, \dots, n - 1$ . Such a coordinate exchange multiplies the Lebesgue measure merely by a positive constant, so the weighted Ricci curvature does not change. Put  $V := \partial/\partial x^n$  on  $D$  and recall that the all integral curves of  $V$  are minimizing (and hence reparametrizations of geodesics). Therefore it suffices to calculate the weighted Ricci curvature of  $(D, g_V, m_L)$ .

We can represent  $\partial D \cap \{x \in \mathbb{R}^n \mid x^n > 0\}$  as the graph of a  $\mathcal{C}^\infty$ -function  $h : U \rightarrow (0, \infty)$  for a sufficiently small neighborhood  $U \subset \mathbb{R}^{n-1}$  of 0, namely

$$(3-2) \quad \partial D \cap \{(z, t) \in \mathbb{R}^{n-1} \times \mathbb{R} \mid z \in U, t > 0\} = \{(z, h(z)) \mid z \in U\}.$$

Then (1-1) yields

$$F(V(z, t)) = \frac{1}{h(z) - t} \quad \text{for } (z, t) \in D \subset \mathbb{R}^{n-1} \times \mathbb{R}.$$

Putting  $\partial_i := \partial/\partial x^i$  for simplicity, we deduce from (3-1) that

$$\begin{aligned} g_{ij}(V) &= (h-t)\partial_i\partial_j\left(\frac{1}{h-t}\right) - (h-t)^2\partial_i\left(\frac{1}{h-t}\right)\partial_j\left(\frac{1}{h-t}\right) \\ &= (h-t)\left(-\frac{\partial_i\partial_j(h-t)}{(h-t)^2} + \frac{2\partial_i(h-t)\partial_j(h-t)}{(h-t)^3}\right) - \frac{\partial_i(h-t)\partial_j(h-t)}{(h-t)^2} \\ &= -\frac{\partial_i\partial_j(h-t)}{h-t} + \frac{\partial_i(h-t)\partial_j(h-t)}{(h-t)^2}, \end{aligned}$$

where the evaluations at  $(z, t) \in D$  were omitted. We remark that, for  $i, j \neq n$ ,

$$g_{ij}(V) = -\frac{\partial_i\partial_j h}{h-t} + \frac{\partial_i h \partial_j h}{(h-t)^2}, \quad g_{in}(V) = -\frac{\partial_i h}{(h-t)^2}, \quad g_{nn}(V) = \frac{1}{(h-t)^2}.$$

Hence, when differentiating  $g_{ij}(V(z, t))$  by  $t$ , we need to take only the denominators into account. Thus we find

$$\begin{aligned} \frac{\partial[g_{ij}(V)]}{\partial t} &= -\frac{\partial_i\partial_j(h-t)}{(h-t)^2} + \frac{2\partial_i(h-t)\partial_j(h-t)}{(h-t)^3} \\ &= \frac{1}{h-t} \left( g_{ij}(V) + \frac{\partial_i(h-t)\partial_j(h-t)}{(h-t)^2} \right). \end{aligned}$$

Decomposing  $m_L$  as

$$m_L = e^{-\Psi} \sqrt{\det(g_{ij}(V))} dx^1 dx^2 \cdots dx^n$$

along the curve  $\eta(t) = (0, t) \in D$ , we observe

$$\Psi(t) = \frac{1}{2} \log \det(g_{ij}(t)), \quad \Psi'(t) = \frac{1}{2} \text{trace}[(g^{ij}(t)) \cdot (g'_{ij}(t))],$$

where we abbreviated as  $g_{ij}(t) := g_{ij}(V(0, t))$  and  $(g^{ij}(t))$  stands for the inverse matrix of  $(g_{ij}(t))$ . Dividing  $\Psi'(t)$  by the speed  $F(\dot{\eta}(t)) = F(V(0, t)) = (h(0)-t)^{-1}$ , we obtain

$$(h(0)-t)\Psi'(t) = \frac{1}{2} \text{trace} \left[ (g^{ij}(t)) \cdot \left( g_{ij}(t) + \frac{\partial_i(h(0)-t)\partial_j(h(0)-t)}{(h(0)-t)^2} \right) \right] \equiv \frac{n+1}{2},$$

where the second equality follows from the fact that  $g_{in}(t) = -\partial_i h(0)/(h(0)-t)^2 = 0$  for  $i \neq n$ , guaranteed by  $g_{in}(v) = 0$ . As  $(D, F)$  has constant flag curvature  $-\frac{1}{4}$ , we therefore conclude that

$$\text{Ric}_\infty(v) = -\frac{n-1}{4}, \quad \text{Ric}_N(v) = -\frac{n-1}{4} - \frac{(n+1)^2}{4(N-n)}. \quad \square$$

#### 4. The Hilbert case

We next consider the Hilbert case, where the calculation is similar but more involved. Now  $F$  will denote the Hilbert metric of  $D$ .



*Proof of Theorem 1.2.* Given a unit vector  $v \in T_x D \cap F^{-1}(1)$ , similarly to the previous section, we choose a coordinate such that  $x$  is the origin,  $v = \partial/\partial x^n$  and that  $g_{in}(v) = 0$  for all  $i = 1, 2, \dots, n-1$ . Put  $V := \partial/\partial x^n$  again. In addition to  $h : U \rightarrow (0, \infty)$  as in (3-2), we introduce the function  $b : U \rightarrow (-\infty, 0)$  such that

$$\partial D \cap \{(z, t) \in \mathbb{R}^{n-1} \times \mathbb{R} \mid z \in U, t < 0\} = \{(z, b(z)) \mid z \in U\}.$$

Using the Funk metric  $F_+$  of  $D$  and its reverse  $F_-(v) := F_+(-v)$ , and recalling (1-1), we can write  $F(V)$  as

$$F(V(z, t)) = \frac{F_+(V(z, t)) + F_-(V(z, t))}{2} = \frac{1}{2} \left( \frac{1}{h(z)-t} + \frac{1}{t-b(z)} \right).$$

It follows from Lemma 3.1 and  $F_-(v) = F_+(-v)$  that

$$\frac{\partial F_-}{\partial x^i} = -F_- \frac{\partial F_-}{\partial v^i}.$$

This yields

$$\begin{aligned} 2 \frac{\partial^2(F^2)}{\partial v^i \partial v^j} &= \frac{1}{2} \frac{\partial^2}{\partial v^i \partial v^j} (F_+^2 + 2F_+F_- + F_-^2) \\ &= \frac{1}{2} \frac{\partial^2(F_+^2)}{\partial v^i \partial v^j} + \frac{1}{2} \frac{\partial^2(F_-^2)}{\partial v^i \partial v^j} - \frac{\partial_i F_+}{F_+} \frac{\partial_j F_-}{F_-} - \frac{\partial_j F_+}{F_+} \frac{\partial_i F_-}{F_-} \\ &\quad + \left( \frac{\partial_i \partial_j F_+}{F_+^2} - \frac{2\partial_i F_+ \partial_j F_+}{F_+^3} \right) F_- + \left( \frac{\partial_i \partial_j F_-}{F_-^2} - \frac{2\partial_i F_- \partial_j F_-}{F_-^3} \right) F_+. \end{aligned}$$

By (3-1) we have, omitting the evaluations at  $(z, t) \in D$ ,

$$\begin{aligned} 4g_{ij}(V) &= -\frac{\partial_i \partial_j (h-t)}{h-t} + \frac{\partial_i (h-t) \partial_j (h-t)}{(h-t)^2} - \frac{\partial_i \partial_j (t-b)}{t-b} + \frac{\partial_i (t-b) \partial_j (t-b)}{(t-b)^2} \\ &\quad - \left( \frac{\partial_i (h-t)}{h-t} \frac{\partial_j (t-b)}{t-b} + \frac{\partial_j (h-t)}{h-t} \frac{\partial_i (t-b)}{t-b} \right) \\ &\quad - \frac{\partial_i \partial_j (h-t)}{t-b} - \frac{\partial_i \partial_j (t-b)}{h-t} \\ &= -(\partial_i \partial_j (h-t) + \partial_i \partial_j (t-b)) \left( \frac{1}{h-t} + \frac{1}{t-b} \right) \\ &\quad + \left( \frac{\partial_i (h-t)}{h-t} - \frac{\partial_i (t-b)}{t-b} \right) \left( \frac{\partial_j (h-t)}{h-t} - \frac{\partial_j (t-b)}{t-b} \right). \end{aligned}$$

Note that the assumption  $g_{in}(v) = 0$  implies

$$(4-1) \quad \frac{\partial_i h(0)}{h(0)} - \frac{\partial_i b(0)}{b(0)} = 0 \quad \text{for } i = 1, 2, \dots, n-1.$$

We also observe for later convenience that, for  $i, j \neq n$ ,

$$4g_{ij}(v) = -(\partial_i \partial_j h(0) - \partial_i \partial_j b(0)) \left( \frac{1}{h(0)} - \frac{1}{b(0)} \right), \quad 4g_{nn}(v) = \left( \frac{1}{h(0)} - \frac{1}{b(0)} \right)^2.$$

By the same reasoning as the Funk case, the numerators can be neglected when one differentiates  $g_{ij}(V)$  with respect to  $t$ . Thus we find

$$\begin{aligned} 4 \frac{\partial g_{ij}(V)}{\partial t} &= -(\partial_i \partial_j (h-t) + \partial_i \partial_j (t-b)) \left( \frac{1}{(h-t)^2} - \frac{1}{(t-b)^2} \right) \\ &\quad + \left( \frac{\partial_i (h-t)}{(h-t)^2} + \frac{\partial_i (t-b)}{(t-b)^2} \right) \left( \frac{\partial_j (h-t)}{h-t} - \frac{\partial_j (t-b)}{t-b} \right) \\ &\quad + \left( \frac{\partial_i (h-t)}{h-t} - \frac{\partial_i (t-b)}{t-b} \right) \left( \frac{\partial_j (h-t)}{(h-t)^2} + \frac{\partial_j (t-b)}{(t-b)^2} \right). \end{aligned}$$

We further calculate

$$\begin{aligned} 4 \frac{\partial^2 [g_{ij}(V)]}{\partial t^2} &= -(\partial_i \partial_j (h-t) + \partial_i \partial_j (t-b)) \left( \frac{2}{(h-t)^3} + \frac{2}{(t-b)^3} \right) \\ &\quad + \left( \frac{2\partial_i (h-t)}{(h-t)^3} - \frac{2\partial_i (t-b)}{(t-b)^3} \right) \left( \frac{\partial_j (h-t)}{h-t} - \frac{\partial_j (t-b)}{t-b} \right) \\ &\quad + \left( \frac{\partial_i (h-t)}{h-t} - \frac{\partial_i (t-b)}{t-b} \right) \left( \frac{2\partial_j (h-t)}{(h-t)^3} - \frac{2\partial_j (t-b)}{(t-b)^3} \right) \\ &\quad + 2 \left( \frac{\partial_i (h-t)}{(h-t)^2} + \frac{\partial_i (t-b)}{(t-b)^2} \right) \left( \frac{\partial_j (h-t)}{(h-t)^2} + \frac{\partial_j (t-b)}{(t-b)^2} \right). \end{aligned}$$

We abbreviate as  $g_{ij}(t) := g_{ij}(V(0, t))$  and deduce from (4-1) that, for  $i, j \neq n$ ,

$$\begin{aligned} 4g'_{ij}(0) &= 4g_{ij}(0) \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right), \\ 4g'_{in}(0) &= - \left( \frac{\partial_i h(0)}{h(0)^2} - \frac{\partial_i b(0)}{b(0)^2} \right) \left( \frac{1}{h(0)} - \frac{1}{b(0)} \right), \\ 4g'_{nn}(0) &= 8g_{nn}(0) \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right). \end{aligned}$$

We also obtain, for  $i, j \neq n$ ,

$$\begin{aligned} 4g''_{ij}(0) &= 8g_{ij}(0) \left( \frac{1}{h(0)^2} + \frac{1}{h(0)b(0)} + \frac{1}{b(0)^2} \right) \\ &\quad + 2 \left( \frac{\partial_i h(0)}{h(0)^2} - \frac{\partial_i b(0)}{b(0)^2} \right) \left( \frac{\partial_j h(0)}{h(0)^2} - \frac{\partial_j b(0)}{b(0)^2} \right), \\ 4g''_{nn}(0) &= 8g_{nn}(0) \left( 2 \left( \frac{1}{h(0)^2} + \frac{1}{h(0)b(0)} + \frac{1}{b(0)^2} \right) + \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 \right). \end{aligned}$$

Put  $\Psi(t) = 2^{-1} \log(\det(g_{ij}(t)))$  and observe

$$\begin{aligned}\Psi'(t) &= \frac{1}{2} \operatorname{trace}[(g^{ij}(t)) \cdot (g'_{ij}(t))], \\ \Psi''(t) &= \frac{1}{2} \operatorname{trace}[(g^{ij}(t)) \cdot (g''_{ij}(t)) - ((g^{ij}(t)) \cdot (g'_{ij}(t)))^2].\end{aligned}$$

Comparing  $g_{ij}(0)$  and  $g'_{ij}(0)$ , we have

$$\Psi'(0) = \frac{1}{2} \left( (n-1) \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right) + 2 \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right) \right) = \frac{n+1}{2} \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right).$$

Similarly,

$$\begin{aligned}\frac{1}{2} \operatorname{trace} [(g^{ij}(0)) \cdot (g''_{ij}(0))] &= (n-1) \left( \frac{1}{h(0)^2} + \frac{1}{h(0)b(0)} + \frac{1}{b(0)^2} \right) \\ &\quad + \frac{1}{4} \sum_{i,j=1}^{n-1} g^{ij}(0) \left( \frac{\partial_i h(0)}{h(0)^2} - \frac{\partial_i b(0)}{b(0)^2} \right) \left( \frac{\partial_j h(0)}{h(0)^2} - \frac{\partial_j b(0)}{b(0)^2} \right) \\ &\quad + 2 \left( \frac{1}{h(0)^2} + \frac{1}{h(0)b(0)} + \frac{1}{b(0)^2} \right) + \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 \\ &= (n+1) \left( \frac{1}{h(0)^2} + \frac{1}{h(0)b(0)} + \frac{1}{b(0)^2} \right) + \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 \\ &\quad + \frac{1}{4} \sum_{i,j=1}^{n-1} g^{ij}(0) \left( \frac{\partial_i h(0)}{h(0)^2} - \frac{\partial_i b(0)}{b(0)^2} \right) \left( \frac{\partial_j h(0)}{h(0)^2} - \frac{\partial_j b(0)}{b(0)^2} \right).\end{aligned}$$

Combining this with

$$\begin{aligned}\operatorname{trace} [((g^{ij}(0)) \cdot (g'_{ij}(0)))^2] &= (n-1) \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 + 4 \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 \\ &\quad + \frac{g^{nn}(0)}{8} \sum_{i,j=1}^{n-1} g^{ij}(0) \left( \frac{\partial_i h(0)}{h(0)^2} - \frac{\partial_i b(0)}{b(0)^2} \right) \left( \frac{\partial_j h(0)}{h(0)^2} - \frac{\partial_j b(0)}{b(0)^2} \right) \left( \frac{1}{h(0)} - \frac{1}{b(0)} \right)^2 \\ &= (n+3) \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 + \frac{1}{2} \sum_{i,j=1}^{n-1} g^{ij}(0) \left( \frac{\partial_i h(0)}{h(0)^2} - \frac{\partial_i b(0)}{b(0)^2} \right) \left( \frac{\partial_j h(0)}{h(0)^2} - \frac{\partial_j b(0)}{b(0)^2} \right),\end{aligned}$$

we obtain

$$\begin{aligned}\Psi''(0) &= (n+1) \left( \frac{1}{h(0)^2} + \frac{1}{h(0)b(0)} + \frac{1}{b(0)^2} \right) - \frac{n+1}{2} \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 \\ &= \frac{n+1}{2} \left( \frac{1}{h(0)^2} + \frac{1}{b(0)^2} \right).\end{aligned}$$

Therefore we have, as  $F(v) = (h(0)^{-1} - b(0)^{-1})/2 = 1$ ,

$$\frac{d}{dt} \left[ \frac{\Psi'(t)}{F(V(0, t))} \right]_{t=0} = \Psi''(0) - \frac{\Psi'(0)}{2} \left( \frac{1}{h(0)^2} - \frac{1}{b(0)^2} \right) = -\frac{n+1}{h(0)b(0)}.$$

Since

$$0 < -\frac{1}{h(0)b(0)} \leq \frac{1}{4} \left( \frac{1}{h(0)} - \frac{1}{b(0)} \right)^2 = 1,$$

this yields  $\text{Ric}_\infty(v) \in (-(n-1), 2]$ . Moreover,

$$\Psi'(0)^2 = \frac{(n+1)^2}{4} \left( \frac{1}{h(0)} + \frac{1}{b(0)} \right)^2 = (n+1)^2 \left( 1 + \frac{1}{h(0)b(0)} \right) \in [0, (n+1)^2)$$

shows that

$$\text{Ric}_N(v) \in \left( -(n-1) - \frac{(n+1)^2}{N-n}, 2 \right]. \quad \square$$

## 5. Applications and remarks

As mentioned in Section 2,  $\text{Ric}_N \geq K$  is equivalent to the curvature-dimension condition  $\text{CD}(K, N)$ . Spaces satisfying  $\text{CD}(K, N)$  enjoy a number of properties similar to Riemannian manifolds of  $\text{Ric} \geq K$  and  $\dim \leq N$ . Since  $\text{CD}(K, N)$  (between compactly supported measures) is preserved under the *pointed measured Gromov–Hausdorff convergence* of locally compact, complete metric measure spaces [Villani 2009, Theorem 29.25], we can deal with merely bounded, convex domains  $D$ .

**Corollary 5.1.** *Let  $D \subset \mathbb{R}^n$  be a bounded convex domain with  $n \geq 2$ . Then the metric measure spaces  $(D, d_{\mathcal{F}}, m_L)$  and  $(D, d_{\mathcal{H}}, m_L)$  satisfy  $\text{CD}(K, N)$  for  $N \in (n, \infty]$  with*

$$K = -\frac{n-1}{4} - \frac{(n+1)^2}{4(N-n)}, \quad K = -(n-1) - \frac{(n+1)^2}{N-n},$$

respectively, where we read  $K = -(n-1)/4$  and  $K = -(n-1)$  when  $N = \infty$ . In particular, they satisfy

- the Brunn–Minkowski inequality by  $\text{CD}(K, N)$  with  $N \in (n, \infty]$ ,
- the Bishop–Gromov volume comparison by  $\text{CD}(K, N)$  with  $N \in (n, \infty)$ .

See [Sturm 2006b, Proposition 2.1, Theorem 2.3] (and, for  $N = \infty$ , also [Villani 2009, Theorem 30.7; Ohta 2010, Theorem 6.1]) for the precise statements of the Brunn–Minkowski inequality and the Bishop–Gromov volume comparison. Beyond the general theory of the curvature-dimension condition, the weighted Ricci curvature bound implies the following.

**Corollary 5.2.** *Let  $D \subset \mathbb{R}^n$  with  $n \geq 2$  be a strongly convex domain such that  $\partial D$  is smooth. For  $K$  as in Corollary 5.1,  $(D, F_{\mathcal{F}}, m_L)$  and  $(D, F_{\mathcal{H}}, m_L)$  satisfy*

- *the Laplacian comparison for  $N \in (n, \infty)$ ,*
- *the Bochner–Weitzenböck inequality for  $N \in (n, \infty)$ .*

See [Ohta and Sturm 2009, Theorem 5.2] for the Laplacian comparison, and [Ohta and Sturm 2011, Theorems 3.3, 3.6] for the Bochner–Weitzenböck formula (by the Bochner–Weitzenböck inequality we meant the inequality given by plugging the weighted Ricci curvature bound into the Bochner–Weitzenböck formula).

We conclude the article with remarks on possible improvements of the estimates in Theorems 1.1, 1.2. Our estimates on  $\text{Ric}_N$  with respect to  $m_L$  are independent of the shape of  $D$ . In particular, Theorem 1.2 provides the same (far from optimal) estimates even for the Klein model of the hyperbolic spaces. Thus there would be a better choice of a measure depending on the shape of  $D$ . Then, as an arbitrary measure is represented by  $e^{-\psi} m_L$ , its weighted Ricci curvature is calculated by combining Theorems 1.1, 1.2 and the convexity of  $\psi$ . One may think of the squared distance function from some point as a candidate of  $\psi$ , however, in order to estimate its convexity along geodesics, we need to bound not only the flag curvature but also the *uniform convexity* as well as the *tangent curvature* (also called the *S-curvature*; see [Ohta 2009b, Theorem 5.1]). The uniform convexity is measured by the constant

$$\mathbf{C} = \sup_{x \in M} \sup_{v, w \in T_x M \setminus \{0\}} \frac{F(w)}{g_v(w, w)^{1/2}},$$

and it is infinite for Funk metrics. As for Hilbert geometry, one could bound  $\mathbf{C}$  by the convexity of  $\partial D$  (but this seems unclear; see [Egloff 1997, Remark 2.1]). The author has no idea about the tangent curvature, which measures how the tangent spaces are distorted as one moves in  $M$ .

There are several natural constructive measures  $m$  on  $D$ , and it is interesting to consider the corresponding weighted Ricci curvature  $\text{Ric}_N^m(V)$ . Then, however, it seems not easy (at least more difficult than  $m_L$ ) to calculate  $\text{Ric}_N^m(V)$  because  $m$  should depend on the shape of whole  $\partial D$ , while  $g_V$  is induced only from the behavior of  $F_{\mathcal{F}}$  or  $F_{\mathcal{H}}$  near the direction  $V$ .

We also remark that, in Hilbert geometry (which is both forward and backward complete),  $\text{Ric}_N$  with  $N < \infty$  cannot be nonnegative for any measure. Otherwise,  $g_V$  splits isometrically, which is a contradiction [Ohta 2012, Proposition 4.3]. Due

to the same reasoning,  $\text{Ric}_\infty$  can be nonnegative only when  $\sup \Psi = \infty$ .

### References

- [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
- [Benoist 2003] Y. Benoist, “Convexes hyperboliques et fonctions quasisymétriques”, *Publ. Math. Inst. Hautes Études Sci.* **97** (2003), 181–237. MR 2005g:53066 Zbl 1049.53027
- [Colbois and Verovic 2004] B. Colbois and P. Verovic, “Hilbert geometry for strictly convex domains”, *Geom. Dedicata* **105** (2004), 29–42. MR 2005e:53111 Zbl 1078.52002
- [Egloff 1997] D. Egloff, “Uniform Finsler Hadamard manifolds”, *Ann. Inst. H. Poincaré Phys. Théor.* **66**:3 (1997), 323–357. MR 98c:53079 Zbl 0919.53020
- [Funk 1929] P. Funk, “Über Geometrien, bei denen die Geraden die Kürzesten sind”, *Math. Ann.* **101**:1 (1929), 226–237. MR 1512527 JFM 55.1043.01
- [Hilbert 1895] D. Hilbert, “Über die gerade Linie als kürzeste Verbindung zweier Punkte”, *Math. Ann.* **46** (1895), 91–96. JFM 26.0540.02
- [Lott and Villani 2007] J. Lott and C. Villani, “Weak curvature conditions and functional inequalities”, *J. Funct. Anal.* **245**:1 (2007), 311–333. MR 2008f:53039 Zbl 1119.53028
- [Lott and Villani 2009] J. Lott and C. Villani, “Ricci curvature for metric-measure spaces via optimal transport”, *Ann. of Math. (2)* **169**:3 (2009), 903–991. MR 2010i:53068 Zbl 1178.53038
- [Ohta 2009a] S.-i. Ohta, “Finsler interpolation inequalities”, *Calc. Var. Partial Differential Equations* **36**:2 (2009), 211–249. MR 2011m:58027 Zbl 1175.49044
- [Ohta 2009b] S.-i. Ohta, “Uniform convexity and smoothness, and their applications in Finsler geometry”, *Math. Ann.* **343**:3 (2009), 669–699. MR 2009m:53199 Zbl 1160.53033
- [Ohta 2010] S. Ohta, “Ricci curvature, entropy and optimal transport”, 2010, <http://www.math.kyoto-u.ac.jp/sohta/papers/Grenoble.pdf>.
- [Ohta 2012] S. Ohta, “Splitting theorems for Finsler manifolds of nonnegative Ricci curvature”, preprint, 2012. To appear in *J. Reine Angew. Math.* arXiv 1203.0079
- [Ohta and Sturm 2009] S.-I. Ohta and K.-T. Sturm, “Heat flow on Finsler manifolds”, *Comm. Pure Appl. Math.* **62**:10 (2009), 1386–1433. MR 2010j:58058 Zbl 1176.58012
- [Ohta and Sturm 2011] S. Ohta and K.-T. Sturm, “Bochner-Weitzenböck formula and Li–Yau estimates on Finsler manifolds”, preprint, 2011. arXiv 1104.5276
- [Okada 1983] T. Okada, “On models of projectively flat Finsler spaces of constant negative curvature”, *Tensor (N.S.)* **40**:2 (1983), 117–124. MR 87c:53124 Zbl 0558.53022
- [Shen 1997] Z. Shen, “Curvature, distance and volume in Finsler geometry”, preprint M-97-48, IHES, 1997, <http://www.math.iupui.edu/zshen/Research/papers/cdv9710.dvi>.
- [Shen 2001a] Z. Shen, *Differential geometry of spray and Finsler spaces*, Kluwer Academic Publishers, Dordrecht, 2001. MR 2003k:53090 Zbl 1009.53004
- [Shen 2001b] Z. Shen, *Lectures on Finsler geometry*, World Scientific Publishing Co., Singapore, 2001. MR 2002f:53032 Zbl 0974.53002
- [Sturm 2006a] K.-T. Sturm, “On the geometry of metric measure spaces, I”, *Acta Math.* **196**:1 (2006), 65–131. MR 2007k:53051a Zbl 1105.53035
- [Sturm 2006b] K.-T. Sturm, “On the geometry of metric measure spaces, II”, *Acta Math.* **196**:1 (2006), 133–177. MR 2007k:53051b Zbl 1106.53032

[Villani 2009] C. Villani, *Optimal transport: Old and new*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] **338**, Springer, Berlin, 2009.  
MR 2010f:49001 Zbl 1156.53003

Received July 24, 2012.

SHIN-ICHI OHTA  
DEPARTMENT OF MATHEMATICS  
KYOTO UNIVERSITY  
KYOTO 606-8502  
JAPAN  
sohta@math.kyoto-u.ac.jp





## ON GENERALIZED WEIGHTED HILBERT MATRICES

EMMANUEL PREISSMANN AND OLIVIER LÉVÊQUE

**We study spectral properties of generalized weighted Hilbert matrices. In particular, we establish results on the spectral norm, the determinant, and various relations between the eigenvalues and eigenvectors of such matrices. We also study the asymptotic behavior of the spectral norm of the classical Hilbert matrix.**

### 1. Introduction

The classical infinite Hilbert matrices

$$(1) \quad T_\infty = \begin{pmatrix} \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \ddots & 0 & -1 & -\frac{1}{2} & -\frac{1}{3} & \ddots \\ \ddots & 1 & 0 & -1 & -\frac{1}{2} & \ddots \\ \ddots & \frac{1}{2} & 1 & 0 & -1 & \ddots \\ \ddots & \frac{1}{3} & \frac{1}{2} & 1 & 0 & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \quad \text{and} \quad H_\infty = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \ddots \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \ddots \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \ddots \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}$$

have been widely studied in the mathematical literature, for a variety of good reasons (see [Choi 1983] for a nice survey of their astonishing properties). In this paper, we present results and conjectures on spectral properties of these matrices and related types of matrices. We first review known results in Section 2, and then introduce new results in Section 3 on generalized weighted Hilbert matrices of the form

$$(2) \quad b_{m,n}(\mathbf{x}, \mathbf{c}) = \begin{cases} 0 & \text{if } m = n, \\ \frac{c_m c_n}{x_m - x_n} & \text{if } m \neq n. \end{cases}$$

Our results can be summarized as follows. Theorem 1 states a surprising property of these matrices: Their spectral norm depends monotonically on the absolute values of their entries, a property known a priori only for matrices with positive entries. Theorem 2 says that the determinants of such matrices are polynomials in the square

*MSC2010:* 26D15.

*Keywords:* Hilbert matrices, Hilbert inequalities, eigenvalue-eigenvector relations.

of their entries. In Lemma 5, we prove a key relation between the eigenvalues and eigenvectors of these matrices, which leads to a chain of nice consequences, including Corollaries 1 and 2. Our work finds its roots in [Montgomery and Vaughan 1973], a seminal paper that initiated the study of generalized Hilbert matrices.

**Notation.** Let  $p > 1$ . In what follows,  $\|\mathbf{y}\|_p$  denotes the  $\ell^p$ -norm of the vector  $\mathbf{y} \in \mathbb{C}^S$ :

$$\|\mathbf{y}\|_p := \left( \sum_{k=1}^S |y_k|^p \right)^{1/p}.$$

For an  $S \times S$  matrix  $M$ ,  $\|M\|_p$  denotes the matrix norm induced by this vector norm:

$$\|M\|_p := \sup_{\|\mathbf{y}\|_p=1} \|M\mathbf{y}\|_p.$$

In the particular case  $p = 2$ , the following simplified notation will be adopted:

$$\|\mathbf{y}\|_2 = \|\mathbf{y}\| \text{ (Euclidean norm)} \quad \text{and} \quad \|M\|_2 = \|M\|.$$

When  $M$  is normal (i.e., when  $MM^* = M^*M$ , where  $M^*$  stands for the complex-conjugate transpose of the matrix  $M$ ), the above norm is equal to the spectral norm of  $M$ :

$$\|M\| = \sup\{|\lambda| : \lambda \in \text{Spec}(M)\}.$$

## 2. A survey of classical results and conjectures

**2.1. Hilbert's inequalities.** The infinite-dimensional matrices presented in (1) are two different versions of the classical Hilbert matrix. Notice first that  $T_\infty$  is a Toeplitz matrix (i.e., a matrix whose entry  $n, m$  depends only on the difference  $m - n$ ), while  $H_\infty$  is a Hankel matrix (i.e., a matrix whose entry  $n, m$  depends only on the sum  $n + m$ ). The Hilbert inequalities state (see [Hardy et al. 1952, p. 212]) that

$$\left| \sum_{m,n \in \mathbb{Z}} u_m (T_\infty)_{m,n} v_n \right| \leq \pi \quad \text{for } \mathbf{u}, \mathbf{v} \in \ell^2(\mathbb{Z}; \mathbb{C}) \text{ with } \|\mathbf{u}\| = \|\mathbf{v}\| = 1$$

and

$$\left| \sum_{m,n \in \mathbb{N}} u_m (H_\infty)_{m,n} v_n \right| \leq \pi \quad \text{for } \mathbf{u}, \mathbf{v} \in \ell^2(\mathbb{N}; \mathbb{C}) \text{ with } \|\mathbf{u}\| = \|\mathbf{v}\| = 1;$$

here  $\pi$  cannot be replaced by a smaller constant.<sup>1</sup> This is saying that  $T_\infty$  and  $H_\infty$  are bounded operators in  $\ell^2(\mathbb{Z}; \mathbb{C})$  and  $\ell^2(\mathbb{N}; \mathbb{C})$ , respectively, with norm  $\pi$ .

Titchmarsh [1926] proved that  $\|T_\infty\|_p < \infty$ . Hardy, Littlewood and Pólya [1952,

---

<sup>1</sup>Hilbert originally proved these inequalities with  $2\pi$  instead of  $\pi$ ; the optimal constant was found later by Schur.

p. 227] obtained the explicit expression

$$\|H_\infty\|_p = \frac{\pi}{\sin(\pi/p)} \quad \text{for all } p > 1.$$

It is clear that  $\|T_\infty\|_p \geq \|H_\infty\|_p$ , as  $H_\infty$  may be seen as the lower left corner of  $T_\infty$  (up to a column permutation), but no exact value is known for it (except in the case where  $p = 2^n$  or  $p = 2^n/(2^n - 1)$  for some integer  $n \geq 1$ ; see [Laeng 2007; 2009] for a review of the subject).

Consider the corresponding  $R \times R$  matrices  $T_R$  and  $H_R$ , defined by

$$(T_R)_{m,n} = \begin{cases} 0 & \text{if } m = n, \\ \frac{1}{m-n} & \text{if } m \neq n, \end{cases} \quad (H_R)_{m,n} = \frac{1}{m+n-1} \quad \text{for } 1 < m, n < R.$$

The Hilbert inequalities imply that for every integer  $R \geq 1$ ,

$$(3) \quad \|T_R\| < \pi \quad \text{and} \quad \|H_R\| < \pi.$$

Clearly also  $\|T_R\|$  and  $\|H_R\|$  increase as  $R$  increases, and

$$\lim_{R \rightarrow \infty} \|T_R\| = \lim_{R \rightarrow \infty} \|H_R\| = \pi.$$

A question of interest is the convergence speed of  $\|H_R\|$  and  $\|T_R\|$  toward their common limiting value  $\pi$ . Up to a column permutation,  $H_R$  can be seen as the lower left corner of  $T_{2R+1}$ , so  $\|H_R\| \leq \|T_{2R+1}\|$  for every integer  $R \geq 1$ . This hints at a slower convergence speed for the matrices  $H_R$  than for the matrices  $T_R$ . Indeed, Wilf and de Bruijn (see [Wilf 1970]) have shown that

$$\pi - \|H_R\| \sim \frac{\pi^5}{2(\log R)^2} \quad \text{as } R \rightarrow \infty,$$

whereas there exist  $a, b > 0$  such that

$$(4) \quad \frac{a}{R} < \pi - \|T_R\| < \frac{b \log R}{R} \quad \text{for } R \geq 2.$$

We will prove these inequalities at the end of this paper. The lower bound has been proved by Montgomery (see [Matthews 2002]), and it has been conjectured in [Preissmann 1985], and independently by Montgomery, that the upper bound in the previous inequality is tight, i.e., that

$$\pi - \|T_R\| \sim \frac{c \log R}{R} \quad \text{as } R \rightarrow \infty.$$

We also provide some numerical evidence for this conjecture at the end of the paper.

**2.2. Toeplitz matrices and Grenander–Szegő’s theorem.** We review the theory developed by Grenander and Szegő [1958] to analyze the asymptotic spectrum of Toeplitz matrices. In particular, we cite their result on the convergence speed of the spectral norm of such matrices.

Let  $(c_r)_{r \in \mathbb{Z}}$  be a sequence of complex numbers such that

$$(5) \quad \sum_{r \in \mathbb{Z}} |c_r| < \infty,$$

and let us define the corresponding function, or *symbol*:

$$f(x) = \sum_{r \in \mathbb{Z}} c_r \exp(irx) \quad \text{for } x \in [0, 2\pi].$$

Because of the assumption made on the Fourier coefficients  $c_r$ , the function  $f$  is continuous, and of course  $f(0) = f(2\pi)$ . Equivalently,  $f$  can be viewed as a continuous  $2\pi$ -periodic function on  $\mathbb{R}$ .

Now let  $C_R$  be the  $R \times R$  matrix defined by

$$(C_R)_{m,n} = c_{m-n} \quad \text{for } 1 \leq m \text{ and } n \leq R.$$

One checks by direct computation that, for any vector  $\mathbf{u} \in \mathbb{C}^R$  with  $\|\mathbf{u}\|^2 := \sum_{1 \leq n \leq R} |u_n|^2 = 1$ , we have

$$(6) \quad \mathbf{u}^* C_R \mathbf{u} = \int_0^{2\pi} f(x) |\phi(x)|^2 dx,$$

where

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \sum_{1 \leq n \leq R} u_n \exp(i(n-1)x).$$

Let us now assume that  $C_R$  is a normal matrix ( $C_R C_R^* = C_R^* C_R$ ); this is the case, for example, when  $f$  is a real-valued function (in which case  $C_R$  is Hermitian:  $C_R^* = C_R$ ). As  $\|\mathbf{u}\| = 1$ , we also have  $\int_0^{2\pi} |\phi(x)|^2 dx = 1$ , which implies that

$$\|C_R\| \leq \sup_{x \in [0, 2\pi]} |f(x)| =: M$$

for any integer  $R \geq 1$ . Grenander and Szegő [1958, p. 72] proved the following refined statement on the convergence speed of the spectral norm. If  $f$  is twice continuously differentiable, admits a unique maximum in  $x_0$  and is such that  $f''(x_0) \neq 0$ , then

$$M - \|C_R\| \sim f(x_0) - f\left(x_0 + \frac{\pi}{R}\right) \sim \frac{\pi^2 |f''(x_0)|}{2R^2} \quad \text{as } R \rightarrow \infty.$$

This result does not apply to Hilbert matrices of the form  $T_R$ : Since the harmonic

series diverges, condition (5) is not satisfied. Correspondingly, the symbol associated with these matrices is the function

$$f(x) = \sum_{r \geq 1} \frac{-\exp(irx) + \exp(-irx)}{r} = -2i \sum_{r \geq 1} \frac{\sin(rx)}{r} = i(x - \pi)$$

for  $x \in ]0, 2\pi[$ , while by Dirichlet’s theorem  $f(0) = f(2\pi) = 0$ . The function  $f$  is therefore discontinuous, but relation (6) still holds in this case and allows us to deduce Hilbert’s inequality:

$$\|T_R\| \leq \sup_{x \in [0, 2\pi]} |f(x)| = \pi.$$

However, relation (6) alone does not allow us to draw conclusions on the convergence speed toward  $\pi$ .

Evaluating the convergence speed of the spectral norm is a difficult problem when  $f$  attains its maximum at a point of discontinuity. An interesting matrix of this type was studied in detail in [Slepian 1978];<sup>2</sup> known as the *prolate matrix*, it is defined as

$$(P_R)_{m,n} = p_{m-n} \quad \text{for } 1 \leq m \text{ and } n \leq R, \quad \text{where } p_r = \begin{cases} \frac{\sin(2\pi wr)}{r} & \text{if } r \neq 0, \\ 2\pi w & \text{if } r = 0, \end{cases}$$

for  $0 < w < \frac{1}{2}$  a fixed parameter. Here, again, we see that condition (5) is not satisfied. The symbol associated with this matrix is the function

$$\begin{aligned} f_w(x) &= \sum_{r \in \mathbb{Z}} p_r \exp(irx) = 2\pi w + 2 \sum_{r \geq 1} \frac{\sin(2\pi wr)}{r} \cos(rx) \\ &= \pi \mathbf{1}_{[0, 2\pi w] \cup [2\pi(1-w), 2\pi]}(x) \end{aligned}$$

for all  $x \in [0, 2\pi] \setminus \{2\pi w, 2\pi(1-w)\}$ . In this case, we again have for any integer  $R \geq 1$

$$\|P_R\| < \sup_{x \in [0, 2\pi]} |f_w(x)| = \pi \quad \text{and} \quad \lim_{R \rightarrow \infty} \|P_R\| = \pi.$$

It is moreover shown in [Slepian 1978] that for all  $0 < \omega < \frac{1}{2}$ , there exist  $c_w, d_w > 0$  (given explicitly in [Varah 1993]) such that

$$\pi - \|P_R\| \sim c_w \sqrt{R} \exp(-d_w R).$$

We see here that although the function  $f_w$  is discontinuous, the convergence speed is exponential, not polynomial (as is the case with a smooth symbol). Of course, the

---

<sup>2</sup>See also [Varah 1993] for a recent exposition of the problem; we are thankful to Ben Adcock for pointing out this interesting reference to us.

situation here is quite particular, as the function  $f_w$  has a plateau at its maximum value, which is not the case for the Hilbert matrix  $T_R$ .

**2.3. Generalized weighted Hilbert matrices.** Let  $\mathbf{x} = (x_1, \dots, x_R)$  be a vector of *distinct* real numbers and  $\mathbf{c} = (c_1, \dots, c_R)$  any vector of real numbers. We define the  $R \times R$  matrix  $B(\mathbf{x}, \mathbf{c})$  by formula (2). We also set

$$(7) \quad A(\mathbf{x}) = B(\mathbf{x}, \mathbf{1}), \quad \text{where } \mathbf{1} = (1, \dots, 1).$$

If there is no risk of confusion, we write  $A$  and  $B$  instead of  $A(\mathbf{x})$  and  $B(\mathbf{x}, \mathbf{c})$ .

Thus  $A(\mathbf{x})$  is the classical Hilbert matrix. To motivate the study of the generalization  $B(\mathbf{x}, \mathbf{c})$ , we mention that Montgomery and Vaughan [1973] proved that

$$\|A(\mathbf{x})\| \leq \frac{\pi}{\delta}, \quad \text{with } \delta = \inf_{\substack{1 \leq m, n \leq R \\ m \neq n}} |x_n - x_m|,$$

and that

$$(8) \quad \|B(\mathbf{x}, \mathbf{c})\| \leq \frac{3\pi}{2}, \quad \text{with } c_n = \sqrt{\min_{\substack{1 \leq m \leq R \\ m \neq n}} |x_m - x_n|}.$$

They also conjectured that the tightest upper bound is  $\|B(\mathbf{x}, \mathbf{c})\| \leq \pi$ . Montgomery and Vaughan’s result was improved in [Preissmann 1984] to  $\|B(\mathbf{x}, \mathbf{c})\| \leq 4\pi/3$ , but the conjecture remains open so far.

We conclude this section with some applications.

*Large sieve inequalities.* Suppose the real numbers  $x_1, \dots, x_R$  are distinct modulo 1. Let  $\|t\|$  denote the distance from a real number  $t$  to the closest integer, and let

$$\delta := \min_{r, s, r \neq s} \|x_r - x_s\| \quad \text{and} \quad \delta_r := \min_{s, s \neq r} \|x_r - x_s\|.$$

For an arbitrary sequence of complex numbers  $(a_n)_{M+1 \leq n \leq M+N}$ , we write

$$S(x) := \sum_{M+1 \leq n \leq M+N} a_n \exp(2\pi i n x).$$

A large sieve inequality has the generic form

$$\sum_{1 \leq r \leq R} |S(x_r)|^2 \leq \Delta(N, \delta) \sum_{M+1 \leq n \leq M+N} |a_n|^2.$$

Using Hilbert’s inequality (3), one can show that the previous inequality holds with  $\Delta(N, \delta) = N + \delta^{-1} - 1$ . Equivalently, this says that if

$$B := \{\exp(2\pi i n x_r)\}_{M+1 \leq n \leq M+N, 1 \leq r \leq R}$$

then

$$\|B\|^2 \leq \Delta(N, \delta).$$

Generalized Hilbert inequalities of type (8) are particularly useful when studying irregularly spaced  $x_r$  (such as Farey sequences), as they allow us to prove the following refined large sieve inequality:

$$\sum_{1 \leq r \leq R} \left(N + \frac{3}{2} \delta_r^{-1}\right)^{-1} |S(x_r)|^2 \leq \sum_{M+1 \leq n \leq M+N} |a_n|^2.$$

This last result is useful for arithmetic applications. It allows us to show, for example, that  $\pi(M + N) - \pi(M) \leq 2\pi(N)$ , where  $\pi(N)$  is the number of primes smaller than or equal to  $N$  (see [Montgomery and Vaughan 1973]). By contrast, the inequality  $\pi(M + N) - \pi(M) \leq \pi(N)$  stands as a conjecture so far.

The Bombieri–Vinogradov theorem, which is related to various conjectures on the distribution of primes, is another important application of large sieve inequalities (see [Bombieri et al. 1986], for instance).

*Other Hilbert inequalities.* Montgomery and Vaughan [1974] studied variants of Hilbert’s inequality (with, for instance,  $1/(x_r - x_s)$  replaced by  $\csc(x_r - x_s)$ ), which allow them to show that if  $\sum_{n \geq 1} n|a_n|^2 < \infty$ , then

$$\int_0^T \left| \sum_{n \geq 1} a_n n^{-it} \right|^2 dt = \sum_{n \geq 1} |a_n|^2 (T + O(n)).$$

The key idea behind the proof of the main result in their paper is the identity

$$\csc(x_k - x_l) \csc(x_l - x_m) = \csc(x_k - x_m) (\cot(x_k - x_l) + \cot(x_l - x_m)),$$

which is of the same type as our relation (10) below. A further generalization of Hilbert’s inequalities has been built on this in [Preissmann 1987], where we solved the functional equations

$$\frac{1}{\theta(x)\theta(y)} = \Psi(x) - \Psi(y) + \frac{\phi(x - y)}{\theta(x - y)}$$

and

$$\frac{1}{\theta(x)\theta(y)} = \frac{\sigma(x) - \sigma(y)}{\theta(x - y)} + \tau(x)\tau(y) \quad \text{with } \tau(0) = 0.$$

### 3. New results

**3.1. Spectral norm of  $B(x, c)$ .** In this subsection we state and prove our first main result, on the monotonicity of the spectral norm of matrices  $B(x, c)$ .

**Theorem 1.** *If  $x, x', c$  and  $c'$  are vectors of real numbers such that*

$$|b_{m,n}(x, c)| \leq |b_{m,n}(x', c')| \quad \text{for } 1 \leq m \text{ and } n \leq R,$$

*then*

$$(9) \quad \|B(x, c)\| \leq \|B(x', c')\|.$$

**Remark.** For matrices  $Y$  and  $Z$  with positive entries, if  $0 \leq y_{m,n} \leq z_{m,n}$  for all  $m$  and  $n$ , then  $\|Y\| \leq \|Z\|$ . Indeed, consider the normalized eigenvector  $\mathbf{u}$  corresponding to the largest eigenvalue of  $Y^*Y$ : Since  $Y^*Y$  has positive entries,  $\mathbf{u}$  is also positive, so  $\|Y\| = \|Y\mathbf{u}\| \leq \|Z\mathbf{u}\| \leq \|Z\|$ . The above result states that a similar result holds for matrices of the form  $B(\mathbf{x}, \mathbf{c})$ , even though these do not have positive entries.

We decompose the proof of Theorem 1 into a sequence of lemmas. We will use several times the relation

$$(10) \quad a_{k,l} a_{l,m} = a_{k,m} (a_{k,l} + a_{l,m}) \quad \text{for } k, l, m \text{ distinct,}$$

where  $a_{m,n} = 1/(x_m - x_n)$ .

**Lemma 1.** *If  $k$  is a positive integer and  $1 \leq n \leq R$ , then, denoting by  $B_{-n}$  the matrix  $B$  with the  $n$ -th row and column removed, we have*

$$(11) \quad S := \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n, l \neq m}} b_{n,l} b_{m,n} (B_{-n}^k)_{l,m} = 0.$$

*Proof.* Using (10), we obtain

$$\begin{aligned} S &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n, l \neq m}} c_l c_m c_n^2 a_{m,n} a_{n,l} (B_{-n}^k)_{l,m} \\ &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n, l \neq m}} c_l c_m c_n^2 a_{m,l} (a_{m,n} + a_{n,l}) (B_{-n}^k)_{l,m} =: S_1 + S_2, \end{aligned}$$

where

$$\begin{aligned} S_1 &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n, l \neq m}} c_l c_m c_n^2 a_{m,l} a_{m,n} (B_{-n}^k)_{l,m} \\ &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n, l \neq m}} c_n^2 b_{m,l} a_{m,n} (B_{-n}^k)_{l,m} = \sum_{\substack{1 \leq m \leq R \\ m \neq n}} c_n^2 a_{m,n} (B_{-n}^{k+1})_{m,m} \end{aligned}$$

and

$$\begin{aligned} S_2 &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n, l \neq m}} c_l c_m c_n^2 a_{m,l} a_{n,l} (B_{-n}^k)_{l,m} \\ &= \sum_{\substack{1 \leq l \leq R \\ l \neq n}} c_n^2 a_{n,l} (B_{-n}^{k+1})_{l,l} = - \sum_{\substack{1 \leq l \leq R \\ l \neq n}} c_n^2 a_{l,n} (B_{-n}^{k+1})_{l,l} = -S_1, \end{aligned}$$

since  $A$  is antisymmetric. □



**Lemma 2.** *Let  $1 \leq n \leq R$  and  $k \geq 2$  be an integer. Then*

$$\begin{aligned} (B^k)_{n,n} &= \sum_{0 \leq r \leq k-2} \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n}} b_{n,l} (B_{-n}^r)_{l,m} b_{m,n} (B^{k-r-2})_{n,n} \\ &= - \sum_{0 \leq r \leq k-2} \sum_{1 \leq l \leq R} b_{n,l}^2 (B_{-n}^r)_{l,l} (B^{k-r-2})_{n,n}. \end{aligned}$$

*Proof.* Notice first that

$$(B^k)_{n,n} = \sum_{1 \leq n_1, \dots, n_{k-1} \leq R} b_{n,n_1} b_{n_1,n_2} \dots b_{n_{k-2},n_{k-1}} b_{n_{k-1},n}.$$

As  $b_{n,n} = 0$ , we may consider  $n_1, n_{k-1} \neq n$  in this sum. For each  $(n_1, \dots, n_{k-1})$ , define

$$s = \inf\{t \in \{2, \dots, k\} \mid n_1 \neq n, \dots, n_{t-1} \neq n, n_t = n\},$$

where, by convention,  $n_k = n$ . Ordering the terms in the above sum according to the value of  $s$ , we obtain

$$\begin{aligned} (B^k)_{n,n} &= \sum_{2 \leq s \leq k} \sum_{n_1, n_{s-1} \neq n} b_{n,n_1} (B_{-n}^{s-2})_{n_1, n_{s-1}} b_{n_{s-1},n} (B^{k-s})_{n,n} \\ &= \sum_{0 \leq r \leq k-2} \sum_{n_1, n_{r+1} \neq n} b_{n,n_1} (B_{-n}^r)_{n_1, n_{r+1}} b_{n_{r+1},n} (B^{k-r-2})_{n,n}, \end{aligned}$$

which is the first equality in the lemma. The second follows from (11) and the fact that  $B$  is antisymmetric. □

**Lemma 3.** *Let  $1 \leq n \leq R$  and let  $k \geq 2$  be an integer.*

- *If  $k$  is odd, then  $(B^k)_{n,n} = 0$ .*
- *If  $k$  is even, then  $(-1)^{k/2} (B^k)_{n,n}$  is a polynomial in the  $b_{l,m}^2, 1 \leq l < m \leq R$ , with positive coefficients.*

*Proof.* Since  $B$  is antisymmetric, the first statement is obvious. The second follows by induction from Lemma 2. □

*Proof of Theorem 1.* Observe that since the matrix  $iB$  is Hermitian, it has  $R$  real eigenvalues  $\mu_1, \dots, \mu_R$  corresponding to an orthonormal basis of eigenvectors, so

$$\|B\| = \max_{1 \leq r \leq R} |\mu_r|.$$

And for a positive integer  $k$ , we have

$$\text{Tr}(B^{2k}) = \sum_{1 \leq r \leq R} (-1)^k \mu_r^{2k}.$$

Therefore, we obtain

$$\|B\| = \lim_{k \rightarrow \infty} \left( (-1)^k \operatorname{Tr}(B^{2k}) \right)^{1/2k},$$

and the theorem follows from Lemma 3.  $\square$

**3.2. Determinant of  $B(\mathbf{x}, \mathbf{c})$ .** Our next result shows that the determinant of  $B(\mathbf{x}, \mathbf{c})$  is a polynomial in the  $b_{l,m}^2$ .

**Theorem 2.** *If  $R$  is odd, then  $\det(B(\mathbf{x}, \mathbf{c})) = 0$ . If  $R = 2T$  is even, then*

$$(12) \quad \det(B(\mathbf{x}, \mathbf{c})) = \prod_{k=1}^R c_k^2 \sum_{(m_i, n_i)_1^T \in E} \prod_{i=1}^T a_{m_i, n_i}^2 = \sum_{(m_i, n_i)_1^T \in E} \prod_{i=1}^T b_{m_i, n_i}^2,$$

where

$$E := \left\{ (m_i, n_i)_1^T \mid \bigcup_{i=1}^T \{m_i, n_i\} = \{1, \dots, R\}, m_i < n_i \text{ for all } i, \text{ and } m_1 < \dots < m_T \right\}.$$

**Lemma 4.** *Let  $l$  be an integer, with  $3 \leq l \leq R$ . Denoting by  $\mathcal{S}_l$  the set of permutations of  $\{1, \dots, l\}$ , we have*

$$(13) \quad S := \sum_{\sigma \in \mathcal{S}_l} a_{\sigma(1), \sigma(2)} a_{\sigma(2), \sigma(3)} \dots a_{\sigma(l-1), \sigma(l)} a_{\sigma(l), \sigma(1)} = 0.$$

*Proof.* We define

$$S_1 := \sum_{\sigma \in \mathcal{S}_l} a_{\sigma(1), \sigma(2)} a_{\sigma(2), \sigma(3)} \dots a_{\sigma(l-1), \sigma(1)} a_{\sigma(l-1), \sigma(l)},$$

$$S_2 := \sum_{\sigma \in \mathcal{S}_l} a_{\sigma(1), \sigma(2)} a_{\sigma(2), \sigma(3)} \dots a_{\sigma(l-1), \sigma(1)} a_{\sigma(l), \sigma(1)}.$$

By (10), we have  $S = S_1 + S_2$ . Now let  $\tau \in \mathcal{S}_l$  be the permutation defined by  $\tau(1) = l-1$ ,  $\tau(2) = 1$ ,  $\tau(3) = 2$ ,  $\dots$ ,  $\tau(l-1) = l-2$ ,  $\tau(l) = l$ . We obtain

$$\begin{aligned} S_2 &= \sum_{\sigma \in \mathcal{S}_l} a_{\sigma\tau(1), \sigma\tau(2)} a_{\sigma\tau(2), \sigma\tau(3)} \dots a_{\sigma\tau(l-1), \sigma\tau(1)} a_{\sigma\tau(l), \sigma\tau(1)} \\ &= \sum_{\sigma \in \mathcal{S}_l} a_{\sigma(l-1), \sigma(1)} a_{\sigma(1), \sigma(2)} \dots a_{\sigma(l-2), \sigma(l-1)} a_{\sigma(l), \sigma(l-1)} = -S_1, \end{aligned}$$

which completes the proof.  $\square$

*Proof of Theorem 2.* By definition,

$$\det(B) = \sum_{\sigma \in \mathcal{S}_R} \varepsilon(\sigma) \prod_{1 \leq n \leq R} a_{n, \sigma(n)} c_n^2.$$

Every permutation  $\sigma$  is a product of  $k$  cycles, with  $1 \leq k \leq n$ . We denote by  $F_1, \dots, F_k$  the supports of these cycles and by  $n_1, n_2, \dots, n_k$  their cardinalities, and we set

$$S(F_i) := \frac{1}{n_i} \sum_{s_1, s_2, \dots, s_{n_i} | \{s_1, s_2, \dots, s_{n_i}\} = F_i} a_{s_1, s_2} a_{s_2, s_3} \dots a_{s_{n_i-1}, s_{n_i}} a_{s_{n_i}, s_1}.$$

In the above expression for  $\det(B)$ , the contribution of the permutations having  $F_1, \dots, F_k$  as supports for their cycles is of the

$$(-1)^{n_1+n_2+\dots+n_k-k} \prod_{i=1}^k S(F_i) \prod_{r=1}^R c_r^2.$$

Hence, by (13) and the fact that the main diagonal is zero, a nonzero contribution can only occur when all cycles are of cardinality 2, which proves the theorem.  $\square$

**Remark.** The above statement allows us to recover part of the conclusion of Lemma 3. First notice that by Theorem 2 and for all  $J \subset \{1, \dots, R\}$ ,  $\det(B_J)$ , where  $B_J = (b_{l,m})_{l,m \in J}$ , is also a polynomial in the  $b_{l,m}^2$ . Define

$$\sigma_k = \sum_{\substack{J \subset \{1, \dots, R\} \\ |J|=k}} \prod_{i \in J} \lambda_i,$$

where  $\lambda_1, \dots, \lambda_R$  are the eigenvalues of  $B$ . Notice that

$$(14) \quad \sigma_k = \sum_{\substack{J \subset \{1, \dots, R\} \\ |J|=k}} \det(B_J).$$

Indeed, let  $P$  be the polynomial defined as  $P(x) = \prod_{1 \leq i \leq R} (x - \lambda_i)$ . We observe that, on one hand, the matrix-valued version of this polynomial is given by

$$P(x) = \prod_{1 \leq i \leq R} (x - \lambda_i I) = x^R + \sum_{k=1}^R x^{R-k} (-1)^k \sum_{\substack{J \subset \{1, \dots, R\} \\ |J|=k}} \prod_{i \in J} \lambda_i = x^R + \sum_{1 \leq k \leq R} x^{R-k} (-1)^k \sigma_k,$$

while, on the other hand,

$$P(x) = \prod_{i=1}^R (x - \lambda_i) = \det(xI - B) = x^R + \sum_{k=1}^R x^{R-k} (-1)^k \sum_{\substack{J \subset \{1, \dots, R\} \\ |J|=k}} \det(B_J),$$

so by identifying the coefficients we obtain equality (14). This implies that  $\sigma_k$  is also a polynomial in the  $b_{l,m}^2$ . Finally, for  $s_l = \sum_{1 \leq i \leq R} \lambda_i^l$ , we have the following recursion, also known as Newton–Girard’s formula:

$$s_l = \sum_{1 \leq i \leq l-1} (-1)^{i-1} \sigma_i s_{l-i} + (-1)^{l-1} l \sigma_l.$$

For example,  $s_0 = n$ ,  $s_1 = \sigma_1$ ,  $s_2 = s_1 \sigma_1 - 2\sigma_2$ ,  $s_3 = s_2 \sigma_1 - s_1 \sigma_2 + 3\sigma_3$ , etc. We therefore find by induction that for all  $k$ ,  $(-1)^k \text{Tr}(B^{2k}) = (-1)^k s_{2k}$  is also

a polynomial in the  $b_{l,m}^2$ , but this alone does not guarantee the positivity of the coefficients obtained in Lemma 3 above.

### 3.3. Formulas regarding the eigenvalues and eigenvectors of $A(\mathbf{x})$ and $B(\mathbf{x}, \mathbf{c})$ .

We first state the following lemma, which has important consequences for the eigenvalues of the matrices  $A(\mathbf{x})$  and  $B(\mathbf{x}, \mathbf{c})$ , as we will see. The approach taken below generalizes the method initiated by Montgomery and Vaughan [1973].

**Lemma 5.** (a) *Let  $\mathbf{u} = (u_1, \dots, u_R)^T$  be an eigenvector of  $A(\mathbf{x})$  for the eigenvalue  $i\mu$ . Then for  $1 \leq n \leq R$ , we have*

$$(15) \quad \mu^2 |u_n|^2 = \sum_{1 \leq m \leq R} a_{m,n}^2 (|u_m|^2 + 2 \Re(u_n \bar{u}_m)).$$

(b) *Let  $\mathbf{u} = (u_1, \dots, u_R)^T$  be an eigenvector of  $B(\mathbf{x}, \mathbf{c})$  for the eigenvalue  $i\mu$ . Then for  $1 \leq n \leq R$ , we have*

$$(16) \quad \mu^2 |u_n|^2 = \sum_{1 \leq m \leq R} a_{m,n}^2 (c_n^2 c_m^2 |u_m|^2 + 2 c_n^3 c_m \Re(u_n \bar{u}_m)).$$

*Proof.* We prove (16), from which (15) follows by specializing to the case  $\mathbf{c} = \mathbf{1}$ .

Our starting assumption is  $B\mathbf{u} = i\mu\mathbf{u}$ , i.e.,  $\sum_{1 \leq m \leq R} b_{n,m} u_m = i\mu u_n$ . Taking the modulus square on both sides, we obtain

$$\mu^2 |u_n|^2 = \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n}} b_{n,m} b_{n,l} u_m \bar{u}_l.$$

(Notice that the sum can be taken over  $l \neq n$  and  $m \neq n$ , as  $b_{n,n} = 0$ .) Therefore,

$$(17) \quad \mu^2 |u_n|^2 = c_n^2 \sum_{\substack{1 \leq l, m \leq R \\ l \neq n, m \neq n}} c_l c_m a_{n,m} a_{n,l} u_m \bar{u}_l = c_n^2 (S_1 + S_2),$$

where  $S_1$  corresponds to the terms in the sum with  $l = m$  and  $S_2$  is its complement:

$$(18) \quad S_1 = \sum_{\substack{1 \leq m \leq R \\ m \neq n}} c_m^2 a_{m,n}^2 |u_m|^2, \quad S_2 = \sum_{\substack{1 \leq l, m \leq R \\ l \neq m, l \neq n, m \neq n}} c_l c_m a_{n,m} a_{n,l} u_m \bar{u}_l.$$

As  $l$ ,  $m$ , and  $n$  are all distinct in this last sum, we can use (10) and the antisymmetry of  $A$  to obtain

$$a_{n,m} a_{n,l} = a_{l,m} a_{n,l} + a_{m,l} a_{n,m},$$

so

$$(19) \quad S_2 = \sum_{\substack{1 \leq l, m \leq R \\ l \neq m, l \neq n, m \neq n}} c_l c_m (a_{l,m} a_{n,l} + a_{m,l} a_{n,m}) u_m \bar{u}_l = S_3 + S_4$$

with

$$\begin{aligned} S_3 &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq m, l \neq n, m \neq n}} c_l c_m a_{l,m} a_{n,l} u_m \bar{u}_l \\ &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq m, l \neq n, m \neq n}} b_{l,m} a_{n,l} u_m \bar{u}_l = \sum_{\substack{1 \leq l \leq R \\ l \neq n}} a_{n,l} \bar{u}_l \sum_{\substack{1 \leq m \leq R \\ m \neq l, m \neq n}} b_{l,m} u_m. \end{aligned}$$

As  $\mathbf{u}$  is an eigenvector of  $B$ , it follows that

$$S_3 = \sum_{\substack{1 \leq l \leq R \\ l \neq n}} a_{n,l} \bar{u}_l (i\mu u_l - b_{l,n} u_n).$$

Likewise, noticing that  $\bar{\mathbf{u}}$  is also an eigenvector of  $B$  (with the corresponding eigenvalue  $-i\mu$ ), we obtain

$$S_4 = \sum_{\substack{1 \leq m \leq R \\ m \neq n}} a_{n,m} u_m \sum_{\substack{1 \leq l \leq R \\ l \neq n}} b_{m,l} \bar{u}_l = \sum_{\substack{1 \leq m \leq R \\ m \neq n}} a_{n,m} u_m (-i\mu \bar{u}_m - b_{m,n} \bar{u}_n).$$

From (19), we deduce that

$$S_2 = S_3 + S_4 = - \sum_{\substack{1 \leq m \leq R \\ m \neq n}} a_{n,m} b_{m,n} (\bar{u}_m u_n + u_m \bar{u}_n) = 2 \sum_{\substack{1 \leq m \leq R \\ m \neq n}} a_{m,n} b_{m,n} \Re(u_m \bar{u}_n).$$

Now, using this together with (17) and (18), we finally obtain

$$\mu^2 |u_n|^2 = \sum_{\substack{1 \leq m \leq R \\ m \neq n}} c_n^2 (c_m^2 a_{m,n}^2 |u_m|^2 + 2 c_m c_n a_{m,n}^2 \Re(u_m \bar{u}_n)),$$

which completes the proof. □

One of the many consequences of Lemma 5 is the following.

**Corollary 1.** *If  $c_1, \dots, c_R$  are all nonzero, then the eigenvalues of  $B(\mathbf{x}, \mathbf{c})$  are all distinct.*

*Proof.* If in the basis of eigenvectors of  $B$  there were two corresponding to the same eigenvalue, it would be possible to find a linear combination of them (also an eigenvector) such that one component (say,  $u_n$ ) would be equal to zero. Then by (16) we would have

$$\sum_{1 \leq m \leq R} a_{m,n}^2 c_n^2 c_m^2 |u_m|^2 = 0,$$

which is impossible, given the assumption made. □

A more precise version of Lemma 5(b) reads as follows.

**Lemma 6.** *Let  $\mathbf{u} = \mathbf{v} + i \mathbf{w}$  ( $\mathbf{v}, \mathbf{w} \in \mathbb{R}^R$ ) be an eigenvector of  $-i\mu$  corresponding to the eigenvalue  $B(\mathbf{x}, \mathbf{c})$ . Then*

$$(20) \quad \mu^2 v_n^2 = \sum_{1 \leq m \leq R} b_{n,m}^2 w_m^2 + 2c_n^2 \sum_{\substack{1 \leq m \leq R \\ m \neq n}} a_{n,m} w_m (\mu v_m - b_{m,n} w_n).$$

Moreover, if  $\mu \neq 0$ , then  $\|\mathbf{v}\| = \|\mathbf{w}\|$ , while if  $\mu = 0$ , then  $\det(B) = 0$ , so one of the eigenvectors corresponding to this eigenvalue is real.

*Proof.* Applying the proof method of Lemma 5 gives

$$\mu^2 v_n^2 = \left( \sum_{1 \leq m \leq R} b_{n,m} w_m \right)^2 = \sum_{1 \leq m \leq R} b_{n,m}^2 w_m^2 + \sum_{\substack{1 \leq l, m \leq R \\ l \neq m}} b_{n,m} b_{n,l} w_m w_l =: S_1 + S_2.$$

We can write

$$S_2 = c_n^2 \sum_{\substack{1 \leq l, m \leq R \\ l \neq m}} c_l c_m a_{n,m} a_{n,l} w_m w_l = c_n^2 (S_3 + S_4),$$

with

$$\begin{aligned} S_3 &= \sum_{\substack{1 \leq l, m \leq R \\ l \neq m, l \neq n, m \neq n}} c_l c_m a_{l,m} a_{n,l} w_m w_l = \sum_{\substack{1 \leq l \leq R \\ l \neq n}} a_{n,l} w_l \sum_{\substack{1 \leq m \leq R \\ m \neq n, m \neq l}} b_{l,m} w_m \\ &= \sum_{\substack{1 \leq l \leq R \\ l \neq n}} a_{n,l} w_l (\mu v_l - b_{l,n} w_n), \end{aligned}$$

and, likewise,

$$S_4 = \sum_{\substack{1 \leq m \leq R \\ m \neq n}} a_{n,m} w_m \sum_{\substack{1 \leq l \leq R \\ l \neq m, l \neq n}} b_{m,l} w_l = \sum_{\substack{1 \leq m \leq R \\ m \neq n}} a_{n,m} w_m (\mu v_m - b_{m,n} w_n).$$

Observing that  $S_3 = S_4$ , we obtain the formula (20).

Finally, we have by assumption that  $B(\mathbf{v} + i \mathbf{w}) = i\mu(\mathbf{v} + i \mathbf{w})$ , so

$$B \mathbf{w} = \mu \mathbf{v} \quad \text{and} \quad B \mathbf{v} = -\mu \mathbf{w}.$$

Consequently, we have

$$\mu \|\mathbf{w}\|^2 = \mu \mathbf{w}^T \mathbf{w} = (-B \mathbf{v})^T \mathbf{w} = (B^T \mathbf{v})^T \mathbf{w} = \mathbf{v}^T B \mathbf{w} = \mu \|\mathbf{v}\|^2,$$

so for  $\mu \neq 0$ , we have  $\|\mathbf{v}\| = \|\mathbf{w}\|$ . □

Finally, let us mention the following nice formula.

**Lemma 7.** *Let  $\mathbf{u}$  be an eigenvector of  $B$  corresponding to the eigenvalue  $\mu$ . Then*

$$\left| \sum_{1 \leq r \leq R} c_r u_r \right|^2 = \sum_{1 \leq r \leq R} |c_r u_r|^2.$$

*Proof.* Let  $C = \text{diag}(c_1, \dots, c_R)$  and  $X = \text{diag}(x_1, \dots, x_R)$ . Then

$$\bar{\mathbf{u}}^T (XCAC - CACX) \mathbf{u} = \bar{\mathbf{u}}^T M \mathbf{u},$$

where  $m_{r,s} = c_r c_s$  for  $r \neq s$  and 0 otherwise. Therefore,

$$\bar{\mathbf{u}}^T M \mathbf{u} = \left| \sum_{1 \leq r \leq R} c_r u_r \right|^2 - \sum_{1 \leq r \leq R} |c_r u_r|^2.$$

On the other hand,

$$\bar{\mathbf{u}}^T (XCAC - CACX) \mathbf{u} = \bar{\mathbf{u}}^T (XB - BX) \mathbf{u} = \bar{\mathbf{u}}^T X i \mu \mathbf{u} - i \mu \bar{\mathbf{u}}^T X \mathbf{u} = 0,$$

as  $\bar{\mathbf{u}}^T (-B) = \bar{\mathbf{u}}^T B^T = (B \bar{\mathbf{u}})^T = (-i \mu \bar{\mathbf{u}})^T = -i \mu \bar{\mathbf{u}}^T$ . The result follows.  $\square$

**3.4. Back to the spectral norm.** Lemma 5 also allows us to deduce the following bounds on the spectral norm of  $A(\mathbf{x})$ .

**Corollary 2.**  $\max_{1 \leq m \leq R} \sum_{1 \leq n \leq R} a_{m,n}^2 \leq \|A(\mathbf{x})\|^2 \leq 3 \max_{1 \leq m \leq R} \sum_{1 \leq n \leq R} a_{m,n}^2$ .

*Proof.* The first inequality is clear, as the  $m$ -th column of  $A$  is the image by  $A$  of the  $m$ -th canonical vector. For the second inequality, we use (16), choosing  $n$  such that  $|u_n|^2 \geq |u_m|^2$  for all  $1 \leq m \leq R$ , and  $\mu = \|A\|$ . We therefore obtain

$$\|A\|^2 |u_n|^2 = \sum_{1 \leq m \leq R} a_{m,n}^2 (|u_m|^2 + 2\Re(u_n \bar{u}_m)) \leq \sum_{1 \leq m \leq R} a_{m,n}^2 (|u_m|^2 + |u_m|^2 + |u_n|^2),$$

so

$$\|A\|^2 |u_n|^2 \leq 3 \sum_{1 \leq m \leq R} a_{m,n}^2 |u_n|^2. \quad \square$$

**3.5. The classical Hilbert matrix  $T_R$ .** The upper bound in Corollary 2 allows us to recover to the original upper bound on  $\|T_R\|$ , where  $T_R$  is the Hilbert matrix defined in the introduction:

$$\|T_R\|^2 \leq \max_{1 \leq m \leq R} 3 \sum_{\substack{1 \leq n \leq R \\ n \neq m}} \frac{1}{(m-n)^2} < 3 \cdot 2 \sum_{n \geq 1} \frac{1}{n^2} = \pi^2.$$

We now come back to the convergence speed of  $\|T_R\|$  toward  $\pi$ , already mentioned in Section 2. We prove inequality (4), namely that there exist positive constants  $a$  and  $b$  such that

$$\frac{a}{R} < \pi - \|T_R\| < \frac{b \log(R)}{R}, \quad \text{where } R \geq 2.$$

The lower bound can be deduced from Lemma 5. From (16), we indeed see that if

$R = 2S + 1$ , then

$$\|T_R\|^2 < 6 \sum_{k=1}^S \frac{1}{k^2} = \pi^2 - 6 \sum_{k>S} \frac{1}{k^2} < \pi^2 - 6 \sum_{k>S} \frac{1}{k(k+1)} = \pi^2 - \frac{6}{S+1},$$

so

$$\pi - \|T_R\| > \frac{6}{(S+1)(\pi + \|T_R\|)} > \frac{3}{\pi(S+1)},$$

which is of the type  $a/R < \pi - \|T_R\|$ . Another way to prove this lower bound is to follow the Grenander–Szegő approach of Section 2.2. Let us first recall (6):

$$\mathbf{u}^* T_R \mathbf{u} = \int_0^{2\pi} f(x) |\phi(x)|^2 dx,$$

where  $f(x) = i(x - \pi)$  for  $x \in (0, 2\pi)$  and  $\phi(x) = \frac{1}{\sqrt{2\pi}} \sum_{1 \leq n \leq R} u_n \exp(i(n-1)x)$ , and where  $\int_0^{2\pi} |\phi(x)|^2 dx = \|\mathbf{u}\|^2 = 1$ . Hence,

$$(21) \quad \pi - \mathbf{u}^* i T_R \mathbf{u} = \int_0^{2\pi} x |\phi(x)|^2 dx,$$

or, with  $E(R) = \left\{ \phi(x) = \frac{1}{\sqrt{2\pi}} \sum_{1 \leq n \leq R} u_n \exp(i(n-1)x) \mid \mathbf{u} \in \mathbb{C}^R, \sum_{1 \leq n \leq R} |u_n|^2 = 1 \right\}$ ,

$$(22) \quad \pi - \|T_R\| = \inf_{\phi \in E(R)} \int_0^{2\pi} x |\phi(x)|^2 dx.$$

It remains to show that the term on the right-hand side of (22) is bounded below by a term of order  $1/R$ . To this end, let us consider  $\phi \in E(R)$  and  $c > 0$ . Using the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \int_0^c |\phi(x)|^2 dx &= \frac{1}{2\pi} \sum_{1 \leq m, n \leq R} u_m \bar{u}_n \int_0^c \exp(i(m-n)x) dx \\ &\leq \frac{c}{2\pi} \sum_{1 \leq m, n \leq R} |u_m| |u_n| = \frac{c}{2\pi} \left( \sum_{1 \leq n \leq R} |u_n| \right)^2 \\ &\leq \frac{cR}{2\pi} \sum_{1 \leq n \leq R} |u_n|^2 = \frac{cR}{2\pi}. \end{aligned}$$

Setting  $c = \pi/R$ , we obtain  $\int_0^{\pi/R} |\phi(x)|^2 dx \leq \frac{1}{2}$ . This in turn implies that

$$\int_0^{2\pi} x |\phi(x)|^2 dx \geq \int_{\pi/R}^{2\pi} x |\phi(x)|^2 dx \geq \frac{\pi}{R} \int_{\pi/R}^{2\pi} |\phi(x)|^2 dx \geq \frac{\pi}{2R}$$

for all  $\phi \in E(R)$ , which settles the lower bound in (4).



To establish the upper bound, we need to find a function  $\phi \in E(R)$  such that

$$(23) \quad \int_0^{2\pi} x |\phi(x)|^2 dx \leq \frac{b \log R}{R}$$

for some constant  $b > 0$ . This will indeed ensure the existence of a vector  $\mathbf{u}$  — namely, the one associated to the function  $\phi \in E(R)$  — such that  $|\mathbf{u}^* T_R \mathbf{u}| \geq \pi - (b \log R)/R$ , thus implying the result.

In view of (23), our goal is to find  $\phi \in E(R)$  such that, for  $c$  and  $\varepsilon$  small,

$$(24) \quad \int_c^{2\pi} |\phi(x)|^2 dx \leq \varepsilon,$$

which does imply that

$$(25) \quad \int_0^{2\pi} x |\phi(x)|^2 dx \leq c \int_0^c |\phi(x)|^2 dx + 2\pi \int_c^{2\pi} |\phi(x)|^2 dx \leq c + 2\pi \varepsilon.$$

Let  $M$  and  $N$  be positive integers such that  $N(M - 1) + 1 \leq R$ , and let

$$g(x) = \left( \sum_{0 \leq m \leq M-1} \exp(imx) \right)^N.$$

The function defined by

$$(26) \quad \phi(x) = \frac{g(x - c/2)}{\sqrt{\int_0^{2\pi} |g(x)|^2 dx}}$$

belongs to  $E(R)$ . We claim that, for  $M$  and  $N$  appropriately chosen,  $\phi$  satisfies (24) with both  $c$  and  $\varepsilon$  small. We first estimate  $\int_0^{2\pi} |g(x)|^2 dx$ .

**Lemma 8.** 
$$\frac{M^{2N}}{N(M - 1) + 1} \leq \frac{1}{2\pi} \int_0^{2\pi} |g(x)|^2 dx \leq M^{2N-1}.$$

*Proof.* Let  $K$  be a positive integer and define the polynomial

$$P_K(t) = \left( \sum_{0 \leq m \leq M-1} t^m \right)^K = \sum_{0 \leq l \leq K(M-1)} b_{l,K} t^l.$$

Clearly,  $b_{l,K} = b_{m,K}$  if  $l + m = K(M - 1)$ . Moreover,

$$|g(x)|^2 = |P_N(\exp(ix))|^2 = \sum_{0 \leq l, m \leq N(M-1)} b_{l,N} b_{m,N} \exp(i(l - m)x),$$

so

$$\begin{aligned} \int_0^{2\pi} |g(x)|^2 dx &= 2\pi \sum_{0 \leq l \leq N(M-1)} b_{l,N}^2 = 2\pi \sum_{0 \leq l \leq N(M-1)} b_{l,N} b_{N(M-1)-l,N} \\ &= 2\pi b_{N(M-1), 2N}. \end{aligned}$$

Therefore, what remains to be proven is

$$\frac{M^{2N}}{N(M-1)+1} \leq b_{N(M-1),2N} \leq M^{2N-1}.$$

Using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} b_{N(M-1),2N} &= \sum_{0 \leq l \leq N(M-1)} b_{l,N}^2 \geq \frac{\left(\sum_{0 \leq l \leq N(M-1)} b_{l,N}\right)^2}{N(M-1)+1} = \frac{P_N(1)^2}{N(M-1)+1} \\ &= \frac{M^{2N}}{N(M-1)+1}. \end{aligned}$$

On the other hand,  $P_{2N}(t) = P_1(t) P_{2N-1}(t)$ , so

$$b_{N(M-1),2N} = \sum_{(N-1)(M-1) \leq l \leq N(M-1)} b_{l,2N-1} \leq P_{2N-1}(1) \leq M^{2N-1},$$

which completes the proof. □

We now set out to prove (24). We retain the same  $\phi$  from (26). As a result of the previous lemma, we have

$$\begin{aligned} \int_c^{2\pi} |\phi(x)|^2 dx &\leq \frac{N(M-1)+1}{M^{2N}} \frac{1}{2\pi} \int_c^{2\pi} |g(x-c/2)|^2 dx \\ &= \frac{N(M-1)+1}{M^{2N}} \frac{1}{2\pi} \int_{c/2}^{2\pi-c/2} |g(x)|^2 dx. \end{aligned}$$

Notice that

$$|g(x)|^2 = \left| \sum_{0 \leq m \leq M-1} \exp(imx) \right|^{2N} = \left( \frac{\sin(Mx/2)}{\sin(x/2)} \right)^{2N},$$

so

$$\int_{c/2}^{2\pi-c/2} |g(x)|^2 dx = 2 \int_{c/2}^{\pi} |g(x)|^2 dx \leq 2 \int_{c/2}^{\pi} \left( \frac{\pi \sin(Mx/2)}{x} \right)^{2N} dx$$

because  $\sin \frac{x}{2} \geq \frac{x}{\pi}$  for  $0 \leq x \leq \pi$ . This implies

$$\int_{c/2}^{2\pi-c/2} |g(x)|^2 dx \leq 2 \int_{c/2}^{\infty} \left( \frac{\pi}{x} \right)^{2N} dx = 2\pi \int_{c/2\pi}^{\infty} \frac{1}{y^{2N}} dy = \frac{2\pi}{2N-1} \left( \frac{2\pi}{c} \right)^{2N-1},$$

and, correspondingly,

$$\varepsilon = \int_c^{2\pi} |\phi(x)|^2 dx \leq \frac{N(M-1)+1}{M^{2N}} \frac{1}{2N-1} \left( \frac{2\pi}{c} \right)^{2N-1}.$$

Assuming  $R \geq 3$  and defining

$$M := \left\lfloor \frac{2R}{\log R} \right\rfloor, \quad N := \left\lfloor \frac{\log R}{2} \right\rfloor, \quad c := \frac{\pi e \log R}{R}$$

(where  $\lfloor x \rfloor$  denotes the integer part of  $x$ ), we verify that  $M(N - 1) + 1 \leq R$  (so  $\phi \in E(R)$ ) and prove below that (24) is satisfied with  $\varepsilon = O(1/R)$ . Indeed, as  $M \geq R/\log R$  and  $N(M - 1) + 1 \leq M(2N - 1)$ , we obtain

$$\frac{N(M-1)+1}{M^{2N}(2N-1)(c/2\pi)^{2N-1}} = \left(\frac{cM}{2\pi}\right)^{1-2N} \frac{1+N(M-1)}{M(2N-1)} \leq \left(\frac{cM}{2\pi}\right)^{1-2N} \leq e^{1-2N} \leq \frac{e^3}{R},$$

as  $1 - 2N < 3 - \log R$ . According to (25), this finally leads to

$$\int_0^{2\pi} x|\phi(x)|^2 dx \leq \frac{\pi e \log R}{R} + \frac{2\pi e^3}{R},$$

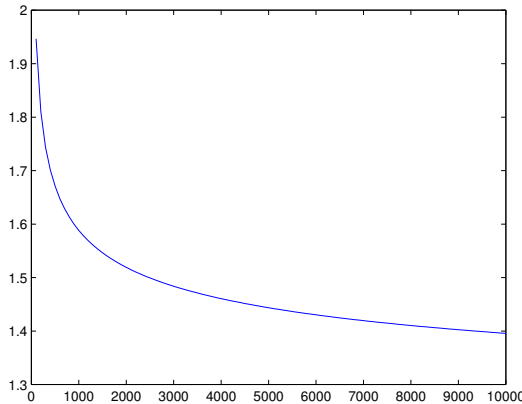
which completes the proof of the upper bound in (4). As already mentioned, it has been conjectured in [Preissmann 1985] that of the two bounds in (4), the upper bound is tight. We provide below some numerical simulation data that supports this fact. In Figure 1, the expression

$$f(R) := (\pi - \|T_R\|) \frac{R}{\log R}$$

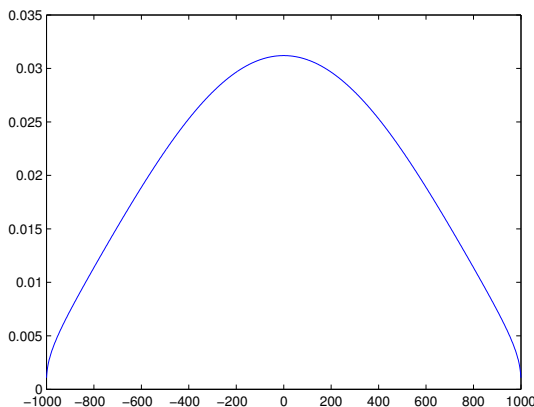
is represented as a function of  $R$ , for values of  $R$  ranging from 1 to 10,000. Detailed facts can also be established about the eigenvectors of  $T_R$ . In order to ease the notation, suppose that  $R = 2S + 1$  and that  $T_R$  is indexed from  $-S$  to  $S$ .

**Lemma 9.** *Let  $u$  be an eigenvector of  $T_R$  corresponding to the eigenvalue  $i\mu$ , and assume without loss of generality that  $u_0 = 1$ . For  $0 \leq n \leq S$ , we have*

$$u_{-n} = -\bar{u}_n.$$



**Figure 1.** Rescaled gap  $f(R)$  between the spectral norm of the infinite-dimensional operator  $T_\infty$  and that of the matrix  $T_R$ , for  $1 \leq R \leq 10,000$ .



**Figure 2.** Amplitude  $\{|u_n| : -R \leq n \leq R\}$  of the eigenvector corresponding to the largest eigenvalue of  $T_R$ , with  $R = 1000$ .

*Proof.* Define  $\mathbf{v}$  by  $v_n = -\bar{u}_{-n}$ . Then

$$(T_R \mathbf{v})_{-m} = \sum_{-S \leq n \leq S} \frac{v_n}{-m-n} = \sum_{-S \leq n \leq S} \frac{v_{-n}}{-m+n} = - \sum_{-S \leq n \leq S} \frac{v_{-n}}{m-n},$$

so

$$(T_R \mathbf{v})_{-m} = \sum_{-S \leq n \leq S} \frac{\bar{u}_n}{m-n} = (T_R \bar{\mathbf{u}})_m = (-i\mu \bar{\mathbf{u}})_m = i\mu v_{-m},$$

i.e.,  $\mathbf{v}$  is an eigenvector corresponding to the eigenvalue  $i\mu$ , with  $v_0 = 1$ . Thus,  $\mathbf{v} = \mathbf{u}$  (as the eigenspace corresponding to  $i\mu$  is of dimension 1).  $\square$

We finally make the following conjecture. Let  $\mathbf{u}$  be the eigenvector corresponding to the largest eigenvalue  $\mu$  in absolute value. Then

$$|u_m| < |u_n| \quad \text{for all } 0 \leq m < n \leq S.$$

This conjecture is confirmed numerically; in Figure 2, we represent  $|u_n|$  as a function of  $n \in \{-S, \dots, S\}$ , for  $S = 1000$ .

From the theoretical point of view, the conjecture also seems reasonable, as  $(-1)^k (T_R^{2k})_{n,n}$  (see Lemma 2) should decrease as  $n$  increases (in absolute value). If true, this fact would therefore hold in the limit  $k \rightarrow \infty$ , which would imply the conjecture on the eigenvector.

## References

- [Bombieri et al. 1986] E. Bombieri, J. B. Friedlander, and H. Iwaniec, “Primes in arithmetic progressions to large moduli”, *Acta Math.* **156**:3-4 (1986), 203–251. MR 88b:11058 Zbl 0588.10042
- [Choi 1983] M. D. Choi, “Tricks or treats with the Hilbert matrix”, *Amer. Math. Monthly* **90**:5 (1983), 301–312. MR 84h:47031 Zbl 0546.47007

- [Grenander and Szegő 1958] U. Grenander and G. Szegő, *Toeplitz forms and their applications*, University of California Press, Berkeley, 1958. Reprinted Chelsea, New York, 1984. MR 20 #1349 Zbl 0080.09501
- [Hardy et al. 1952] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2nd ed., Cambridge University Press, 1952. MR 13,727e Zbl 0047.05302
- [Laeng 2007] E. Laeng, “Remarks on the Hilbert transform and on some families of multiplier operators related to it”, *Collect. Math.* **58**:1 (2007), 25–44. MR 2008b:42012 Zbl 1135.42006
- [Laeng 2009] E. Laeng, “Sharp norm inequalities for the truncated Hilbert transform”, *J. Math. Inequal.* **3**:1 (2009), 123–127. MR 2010a:42024 Zbl 1158.42004
- [Matthews 2002] K. Matthews, “Hilbert’s inequality”, unpublished notes, 2002, available at <http://www.numbertheory.org/pdfs/hilbert.pdf>.
- [Montgomery and Vaughan 1973] H. L. Montgomery and R. C. Vaughan, “The large sieve”, *Mathematika* **20** (1973), 119–134. MR 51 #10260 Zbl 0296.10023
- [Montgomery and Vaughan 1974] H. L. Montgomery and R. C. Vaughan, “Hilbert’s inequality”, *J. London Math. Soc.* (2) **8** (1974), 73–82. MR 49 #2544 Zbl 0281.10021
- [Preissmann 1984] E. Preissmann, “Sur une inégalité de Montgomery–Vaughan”, *Enseign. Math.* (2) **30**:1-2 (1984), 95–113. MR 85e:11067 Zbl 0548.10031
- [Preissmann 1985] E. Preissmann, *Autour des inégalités de Hilbert–Montgomery–Vaughan*, Ph.D. thesis, University of Lausanne, 1985.
- [Preissmann 1987] E. Preissmann, “Équations fonctionnelles et estimations de normes de matrices”, *Aequationes Math.* **32**:2-3 (1987), 195–212. MR 88i:39014 Zbl 0625.39009
- [Slepian 1978] D. Slepian, “Prolate spheroidal wave functions, Fourier analysis and uncertainty, V: The discrete case”, *Bell System Tech. J.* **57**:5 (1978), 1371–1430. Zbl 0378.33006
- [Titchmarsh 1926] E. C. Titchmarsh, “Reciprocal formulae involving series and integrals”, *Math. Z.* **25**:1 (1926), 321–347. MR 1544814 JFM 52.0213.03
- [Varah 1993] J. M. Varah, “The prolate matrix”, *Linear Algebra Appl.* **187** (1993), 269–278. MR 94e:15058 Zbl 0782.15014
- [Wilf 1970] H. S. Wilf, *Finite sections of some classical inequalities*, *Ergebnisse der Mathematik und ihrer Grenzgebiete* **52**, Springer, New York, 1970. MR 42 #6643 Zbl 0199.38301

Received August 17, 2012.

EMMANUEL PREISSMANN  
SWISS FEDERAL INSTITUTE OF TECHNOLOGY  
CH-1015 LAUSANNE  
SWITZERLAND  
emmanuel.preissmann@gmail.com

OLIVIER LÉVÊQUE  
SWISS FEDERAL INSTITUTE OF TECHNOLOGY  
CH-1015 LAUSANNE  
SWITZERLAND  
olivier.leveque@epfl.ch



## UNIQUE PRIME DECOMPOSITION RESULTS FOR FACTORS COMING FROM WREATH PRODUCT GROUPS

J. OWEN SZEMORE AND ADAM WINCHESTER

**We use malleable deformations combined with spectral gap rigidity theory, in the framework of Popa’s deformation/rigidity theory, to prove unique tensor product decomposition results for  $\text{II}_1$  factors arising as tensor products of wreath product factors. We also obtain a similar result regarding measure equivalence decomposition of direct products of such groups.**

### Introduction

A major goal of the study of  $\text{II}_1$  factors is the classification of these algebras based on the “input data” that goes into their construction. For example, given a countable discrete group  $\Gamma$ , one can construct the associated group von Neumann algebra  $L(\Gamma)$ . It is then natural to determine the properties/isomorphism class of the algebra based on those of the group. A significant landmark was the result, due to Connes [1976], that all group von Neumann algebras,  $L(\Gamma)$ , with  $\Gamma$  amenable i.c.c., are isomorphic. However, in the nonamenable realm there is a much greater variety, and a striking classification theory has developed.

One goal of this research is to determine if some algebra, which is constructed in one manner, can be obtained in some other manner. For example, if we have a  $\text{II}_1$  factor that we know to be a free product of two  $\text{II}_1$  factors, is it also possible for it to be the tensor product of two (possibly different)  $\text{II}_1$  factors? Over the last decade many examples of such so-called “ $W^*$ -rigidity” phenomena have been discovered and, in particular, wreath products, or their ergodic theory counterparts, Bernoulli shifts, have played a prominent role. In particular, they have led to the first examples [Ioana et al. 2013; Berbec and Vaes 2012] of  $W^*$ -superrigid groups (i.e., groups  $\Gamma$  for which, for any  $\Lambda$ , isomorphism of  $L(\Gamma)$  and  $L(\Lambda)$  implies isomorphism of  $\Gamma$  and  $\Lambda$ ). For a more detailed overview of the theory we refer the reader to [Popa 2007; Vaes 2011].

Here we study whether certain factors can be written as tensor products in distinct ways. Recall that a  $\text{II}_1$  factor is prime if it is not the tensor product of two other  $\text{II}_1$  factors. The first example of a prime  $\text{II}_1$  factor was obtained by Popa [1983], who showed that the group von Neumann algebra of an uncountable free group is prime.

---

*MSC2010:* 47L65, 46L36.

*Keywords:* operator algebras, measure equivalence, group theory.

Later, Ge [1998] used techniques from Voiculescu’s free probability theory, in particular the tools of free entropy which were defined and developed in [Voiculescu 1993; 1994; 1996], to prove that all group factors coming from finitely generated free groups are prime. Note that Ge’s result, unlike Popa’s result mentioned above, gave the first example of a prime  $\text{II}_1$  factor that is separable. Using  $C^*$ -algebraic techniques, this was greatly generalized in [Ozawa 2004] to show that all i.c.c. Gromov hyperbolic groups give rise to prime  $\text{II}_1$  factors. Also, using his deformation/rigidity theory, Popa [2008] showed that all  $\text{II}_1$  factors arising from the Bernoulli actions of nonamenable groups are prime. Further, Peterson [2009] used his derivation approach to deformation/rigidity to prove that any  $\text{II}_1$  factor coming from a countable group with positive first  $l^2$ -Betti number is also prime. Finally we should also note that, using Popa’s deformation/rigidity theory, Chifan and Houdayer [2010] gave many more examples of prime  $\text{II}_1$  factors coming from amalgamated free products.

A natural question about prime factors is whether a tensor product of a finite number of such factors  $P_1, P_2, \dots, P_n$  has a “unique prime factor decomposition”; i.e., if  $P_1 \bar{\otimes} \dots \bar{\otimes} P_n = Q_1 \bar{\otimes} \dots \bar{\otimes} Q_m$ , for some  $m \geq n$  and some other prime factors  $Q_j$ , forces  $n = m$  and  $P_i$  unitary conjugate to  $Q_i$ , modulo some permutation of indices and modulo some “rescaling” by appropriate amplifications of the prime factors involved. The first such result was obtained in [Ozawa and Popa 2004], where a combination of  $C^*$ -algebraic techniques from [Ozawa 2004] and intertwining techniques from [Popa 2006c] is used to show that any  $\text{II}_1$  factor arising from a tensor product of  $\text{II}_1$  factors of the form  $L(\Gamma)$  with  $\Gamma$  hyperbolic, or more generally in Ozawa’s class  $\mathcal{S}$ , has such a unique tensor product decomposition.

In this paper we prove an analogous unique prime factor decomposition result for tensor products of group von Neumann algebras coming from wreath product groups. More precisely, let us denote by  $\mathcal{WR}_{\text{NA}}$  the class of “amenable by nonamenable” wreath product groups, by which we means groups of the form  $A \wr H$  where  $A$  is a nontrivial countable amenable group and  $H$  is a countable nonamenable group. Then we prove the following result:

**Theorem 0.1.** *Let  $\Gamma_1, \dots, \Gamma_n \in \mathcal{WR}_{\text{NA}}$  and  $Q_1, \dots, Q_m$  be diffuse von Neumann algebras such that*

$$M = L(\Gamma_1) \bar{\otimes} \dots \bar{\otimes} L(\Gamma_n) = Q_1 \bar{\otimes} \dots \bar{\otimes} Q_k.$$

*If  $m \geq n$ , then  $n = m$ , and after permutation of indices we have that  $L(\Gamma_1) \simeq Q_i^{t_1}$  for some positive numbers  $t_1, t_2, \dots, t_n$  whose product is 1.*

We can view this as a “ $W^*$ -rigidity” theorem in that it gives us large families of nonisomorphic  $\text{II}_1$  factors. In particular, picking a specific amenable group,  $\mathbb{Z}$ , and a specific nonamenable group,  $\mathbb{F}_n$ , the free group on  $n$  generators, we get the new



result that

$$L(\mathbb{Z} \wr \mathbb{F}_n) \bar{\otimes} L(\mathbb{Z} \wr \mathbb{F}_n) \bar{\otimes} L(\mathbb{Z} \wr \mathbb{F}_n) \not\cong L(\mathbb{Z} \wr \mathbb{F}_n) \bar{\otimes} L(\mathbb{Z} \wr \mathbb{F}_n).$$

Of course, the above theorem provides us with many such examples of rigidity phenomena.

Also we have a natural generalization of this theorem to unique measure-equivalence decomposition results of finite products of groups in the class  $\mathcal{WR}_{NA}$ . Such results were achieved for products of groups of the class  $\mathcal{C}_{reg}$  in [Monod and Shalom 2006, Theorem 1.16] and for products of biexact groups in [Sako 2009, Theorem 4] and, independently, in [Chifan and Sinclair 2010, Corollary C]. We refer the reader to the last section for the definition of measure equivalence for groups.

Before stating our second main result we would like to point out that Sako [2009, Theorem 7] has obtained measure equivalence rigidity results for certain classes of wreath products; however, his results were not of this type and used techniques different from the ones that we will employ.

**Theorem 0.2.** *Let  $\Gamma_1, \dots, \Gamma_n, \Lambda_1, \dots, \Lambda_m \in \mathcal{WR}_{NA}$  be such that  $\Gamma_1 \times \dots \times \Gamma_n$  is measure equivalent to  $\Lambda_1 \times \dots \times \Lambda_m$ . We denote this by*

$$\Gamma_1 \times \dots \times \Gamma_n \simeq_{ME} \Lambda_1 \times \dots \times \Lambda_m.$$

*If  $m \geq n$ , then  $n = m$ , and after permutation of indices we have that  $\Gamma_i \simeq_{ME} \Lambda_i$ ,  $\Gamma_i$  is measure equivalent to  $\Lambda_i$ .*

We prove these results by using deformation/rigidity theory. More precisely, we use the malleable deformation for wreath product group factors in [Chifan et al. 2012], combined with Popa’s spectral gap rigidity and intertwining by bimodules techniques.

### 1. Preliminaries

**Intertwining by bimodules.** Let us recall Popa’s intertwining by bimodules technique. This is a crucial tool for locating subalgebras of  $\text{II}_1$  factors, and is summed up in the following theorem:

**Theorem 1.1** [Popa 2006c]. *Let  $(M, \tau)$  be a finite von Neumann algebra with trace  $\tau$ , and let  $P, Q \subset M$  be von Neumann subalgebras. Then the following are equivalent:*

- (1) *There exist nonzero projections  $p \in P, q \in Q$ , a nonzero partial isometry  $v \in M$ , and a  $*$ -homomorphism  $\varphi : pPp \rightarrow qQq$  such that  $vx = \varphi(x)v$  for all  $x \in pPp$ .*

- (2) *There is a sub- $P$ - $Q$ -bimodule  $\mathcal{H} \subset L^2(M)$  that is finitely generated as a right  $Q$ -module.*
- (3) *There is no sequence  $u_n \in \mathcal{U}(P)$  such that*

$$\lim_{n \rightarrow \infty} \|E_Q(xu_n y)\|_2 \rightarrow 0 \quad \text{for all } x, y \in M.$$

If any of the above conditions hold, we say that *a corner of  $P$  embeds in  $Q$  inside  $M$* , denoted by  $P \prec_M Q$ .

Following [Ozawa and Popa 2010] we have the following definition:

**Definition 1.2.** Let  $(M, \tau)$  be a finite von Neumann algebra with trace  $\tau$ , and let  $P, Q \subset M$  be von Neumann subalgebras. We say that  *$P$  is amenable over  $Q$  inside  $M$* , which we denote by  $P \prec_M Q$ , if there is a  $P$ -central state,  $\varphi$ , on  $\langle M, e_Q \rangle$  such that  $\varphi|_M = \tau$ , where  $\tau$  is the trace on  $M$ .

Let us note that, by [Ozawa and Popa 2010, Theorem 2.1],  $P \prec_M Q$  is equivalent to  $L^2(P)$  being weakly contained in  $\bigoplus L^2(\langle M, e_Q \rangle)$  as  $P$ -bimodules. Further, if  $P \prec_M Q$  then  $L^2(M)$  contains a sub- $P$ - $Q$ -module,  $\mathcal{H}$ , that is finitely generated as a right  $Q$ -module. Therefore, the projection onto this module will commute with the right action of  $Q$  and will have finite trace. Therefore, it will be a vector in  $L^2(\langle M, e_Q \rangle)$ . Further, it will also commute with  $P$ , so, if we look at  $L^2(\langle M, e_Q \rangle)$  as a  $P$ -bimodule, it will contain a central vector. Since strong containment implies weak containment we get the following observation.

**Proposition 1.3.** *Let  $(M, \tau)$  be a finite von Neumann algebra with trace  $\tau$ , and let  $P, Q \subset M$  be von Neumann subalgebras. If  $P \prec_M Q$  then  $P \prec_M Q$ .*

**Deformation of wreath products.** Let  $A$  and  $H$  be countable discrete groups. Now let us consider the infinite direct sum,  $\bigoplus_H A$ , indexed by  $H$ . Now notice that  $H$  acts on  $\bigoplus_H A$  by acting on the index set on the left. The resulting semidirect product group  $\bigoplus_H A \rtimes H = A \wr H$  is known as the wreath product. Throughout this paper we will consider trace preserving actions of  $A \wr H$  on a finite von Neumann algebra  $N$  with trace  $\tau$ , and we consider the resulting crossed product algebra  $M = N \rtimes A \wr H$ .

Now let us describe the construction of a deformation for von Neumann algebras coming from wreath products as above. This is the same deformation that the first author used in [Chifan et al. 2012], and is inspired by similar free malleable deformations in [Popa 2006b; Ioana et al. 2008; Ioana 2007]. We refer to this previous work for additional discussion.

Let  $\tilde{A} = A * \mathbb{Z}$ . If we let  $u \in L(\tilde{A})$  denote the Haar unitary that generates  $L(\mathbb{Z})$  then we can find a selfadjoint element  $h \in L(\mathbb{Z})$  such that  $u = \exp(ih)$ . Thus, for every  $t \in \mathbb{R}$ , we define  $u^t \doteq \exp(iht) \in L\tilde{A}$ . This allows us to define  $\theta_t \in \text{Aut}(L(\tilde{A}))$  by  $\theta_t(x) = u^t x (u^*)^t$ . By applying this automorphism in each coordinate we can get

an automorphism of  $L(\tilde{A}^H)$ . Since the action of  $H$  is by permuting the coordinates, it commutes with  $\theta_t$  and so we can extend it to  $L(\tilde{A} \wr H)$ . If we now declare that the Haar unitaries in each coordinate do not act on the algebra  $N$ , then we can extend to an automorphism, which we still denote by  $\theta_t$ , of  $\tilde{M} = N \rtimes \tilde{A} \wr H$ .

It is easy to see that  $\lim_{t \rightarrow 0} \|u^t - 1\|_2 = 0$ , and hence  $\lim_{t \rightarrow 0} \|\theta_t(x) - x\|_2 = 0$  for all  $x \in \tilde{M}$ . Therefore, the path  $(\theta_t)_{t \in \mathbb{R}}$  is a deformation by automorphisms of  $\tilde{M}$ .

Next we show that  $\theta_t$  admits a ‘‘symmetry’’; i.e., there exists an automorphism  $\beta$  of  $\tilde{M}$  satisfying the following relations:

$$\beta^2 = \text{id}, \quad \beta|_M = \text{id}|_M, \quad \beta\theta_t\beta = \theta_{-t} \text{ for all } t \in \mathbb{R}.$$

To see this, we first define  $\beta|_{L(A^H)} = \text{id}|_{L(A^H)}$  and then for every  $h \in H$  we let  $(u)_h$  be the element in  $L\tilde{A}^H$  whose  $h$ -th entry is  $u$  and whose other entries are 1. On elements of this form we define  $\beta((u)_h) = (u^*)_h$ , and, since  $\beta$  commutes with the actions of  $H$  on  $A^H$ , it extends to an automorphism of  $L(\tilde{A} \wr H)$  by acting identically on  $L(H)$ . Finally, the automorphism  $\beta$  extends to an automorphism of  $\tilde{M}$ , still denoted by  $\beta$ , which acts trivially on  $A$ .

Let us note that, with this choice of  $\beta$ ,  $\theta_t$  is an *s-malleable deformation* of  $\tilde{M}$  in the sense of [Popa 2006c].

## 2. Intertwining techniques for wreath products

In this section we prove the necessary intertwining results for  $\text{II}_1$  factors arising from wreath product groups that we will need in order to prove our desired uniqueness of tensor product decomposition.

The following proposition is a relative version of [Chifan et al. 2012, Lemma 4.2], and will follow a similar proof.

**Proposition 2.1.** *Let  $N$  be a finite von Neumann algebra. Let  $A, H$  be groups with  $A$  nontrivial amenable and  $H$  nonamenable. Let  $Q \subset N \rtimes A \wr H = M$  be an inclusion of von Neumann algebras. Assume  $Q$  is not amenable over  $N$  inside  $M$ ; then  $Q' \cap \tilde{M}^\omega \subseteq M^\omega$ .*

*Proof.* As mentioned above this proof follows closely the proof of [Chifan et al. 2012, Lemma 4.2] as well as [Popa 2008, Lemma 5.1] and other similar results in the literature.

We will prove the contrapositive so let us assume that  $Q' \cap \tilde{M}^\omega \not\subseteq M^\omega$ . Then, proceeding as in [Popa 2008, Lemma 5.1], we see that

$$L^2(Q) \prec L^2(\tilde{M}) \ominus L^2(M)$$

as  $Q$ -bimodules. Now we decompose  $L^2(\tilde{M}) \ominus L^2(M)$  as an  $M$ -bimodule.

One can see that the above  $M$ -bimodule can be written as a direct sum of  $M$ -bimodules  $\overline{M\tilde{\eta}_s M}^{\|\cdot\|_2}$ , where the cyclic vectors  $\tilde{\eta}_s$  correspond to an enumeration of

all elements of  $\tilde{A}^H$  whose nontrivial coordinates start and end with nonzero powers of  $u$ .

Next, for every  $s$ , we denote by  $\eta_s$  the element of  $A^H$  that remains from  $\tilde{\eta}_s$  after deleting all nontrivial powers of  $u$ . Also for every  $s$  let  $\Delta_s \subset H$  be the support of  $\tilde{\eta}_s$  and observe that if  $\text{Stab}_H(\tilde{\eta}_s)$  denotes the stabilizing group of  $\tilde{\eta}_s$  inside  $H$  then we have  $\text{Stab}_H(\tilde{\eta}_s)(H \setminus \Delta_s) \subset H \setminus \Delta_s$ .

Hence we can consider the von Neumann algebra

$$K_s = N \rtimes (A \wr_{H \setminus \Delta_s} \text{Stab}_H(\tilde{\eta}_s))$$

and, using similar computations to those in [Popa 2008, Lemma 5.1], one can easily check that the map  $x\tilde{\eta}_s y \rightarrow x\eta_s e_{K_s} y$  implements an  $M$ -bimodule isomorphism between  $\overline{M\tilde{\eta}_s M}^{\|\cdot\|_2}$  and  $L^2(\langle M, e_{K_s} \rangle)$ .

Therefore, as  $M$ -bimodules, we have the isomorphism

$$L^2(\tilde{M}) \ominus L^2(M) = \bigoplus L^2(\langle M, e_{K_s} \rangle).$$

Thus we can get the weak containment of  $Q$ -bimodules

$$L^2(Q) \prec \bigoplus L^2(\langle M, e_{K_s} \rangle).$$

Notice that, since  $\Delta_s$  is finite, and the action of  $H$  on itself is free, then  $\text{Stab}_H(\tilde{\eta}_s)$  is finite for all  $s$ . Also, since  $A$  is an amenable group we have that  $K_s \prec_N N$  for all  $s$ . Thus for all  $s$  we have the weak containment of  $K_s$ -bimodules

$$L^2(K_s) \prec \bigoplus L^2(\langle K_s, e_N \rangle) \simeq \bigoplus L^2(K_s) \otimes_N L^2(K_s).$$

Now if we induce to  $M$ -bimodules and restrict to  $Q$ -bimodules and use continuity of weak containment under induction and restriction we get the inclusions of  $Q$ -bimodules

$$\begin{aligned} L^2(Q) &\prec \bigoplus L^2(\langle M, e_{K_s} \rangle) \simeq \bigoplus L^2(M) \otimes_{K_s} L^2(K_s) \otimes_{K_s} L^2(M) \\ &\prec \bigoplus L^2(M) \otimes_{K_s} L^2(K_s) \otimes_N L^2(K_s) \otimes_{K_s} L^2(M) \\ &\simeq \bigoplus L^2(M) \otimes_N L^2(M) \simeq \bigoplus L^2(\langle M, e_N \rangle). \end{aligned}$$

Thus  $Q \prec_M N$ . □

To state the next result let us recall the following standard definition.

**Definition 2.2.** Given an inclusion of von Neumann algebras  $P \subset M$  the *normalizer of  $P$  inside  $M$*  is the set

$$\mathcal{N}_M(P) = \{u \in \mathcal{U}(M) : uPu^* = P\}.$$

We say that, for such an inclusion,  $P$  is a *regular subalgebra* if  $\mathcal{N}_M(P)'' = M$ .

We finish this section with a theorem that allows us to locate regular subfactors with large commutant.

**Theorem 2.3.** *Let  $N$  be a finite von Neumann algebra. Let  $A$  and  $H$  be groups with  $A$  nontrivial amenable and  $H$  nonamenable. Let  $Q \subset N \rtimes A \wr H = M$  be a von Neumann subalgebra that is not amenable over  $N$ . Let  $P = Q' \cap M$ . If  $P$  is a regular subfactor of  $M$  then  $P \prec_M N$ .*

*Proof.* Applying Proposition 2.1 and following the proof of [Chifan et al. 2012, Theorem 4.1] we see that the deformation  $\theta_t$  converges uniformly on the unit ball of  $P$ , and thus by [Chifan et al. 2012, Theorem 3.1] we have that  $P \prec_M N \rtimes A^H$  or  $P \prec_M N \rtimes H$ .

Following the same argument as [Chifan et al. 2012, Theorem 4.1], if we assume that  $P \prec_M N \rtimes A^H$  and  $P \not\prec_M N$ , then we get  $Q \prec_M N \rtimes A \wr H_0$  for some finite subgroup  $H_0 \subset H$ . Since  $A$  is amenable and  $H_0$  is finite then  $N \rtimes A \wr H_0 \prec_M N$ . So, since  $Q \prec_M N \rtimes A \wr H_0$ , then, by Proposition 1.3, we have  $Q \prec_M N \rtimes A \wr H_0$ . Then by [Ozawa and Popa 2010, part 3 of Proposition 2.4] we have that  $Q \prec_M N$ , contradicting our assumption.

Thus  $P \prec_M N \rtimes H$ . Therefore, by Theorem 1.1, there exist nonzero projections  $p \in P$ ,  $q \in N \rtimes H$ , a nonzero partial isometry  $v \in M$ , and a \*-homomorphism  $\varphi : pPp \rightarrow q(N \rtimes H)q$  such that  $vx = \varphi(x)v$  for all  $x \in pPp$ . Furthermore we have that  $v^*v = p$  and  $vv^* = \hat{q} \in \varphi(pPp)' \cap qMq$ . Also, by [Popa 2006c, Lemma 3.5] we know that  $pPp$  is a regular subalgebra of  $pMp$ .

Then for all  $u \in \mathcal{N}_{pMp}(pPp)$  let us calculate

$$\begin{aligned} \varphi(x)vuv^* &= vxuv^* = vu(u^*xu)v^* = vuv^*v(u^*xu)v^* \\ &= vuv^*\varphi(u^*xu)v^* = vuv^*\varphi(u^*xu). \end{aligned}$$

Now assume that  $P \not\prec_M N$ ; then by [Chifan et al. 2012, part (3) of Lemma 2.4] we have that  $vuv^* \in N \rtimes H$ . Since  $pPp$  is regular in  $pMp$  we would then get that  $M \prec_M N \rtimes H$ . However, this is impossible since the fact that  $A$  is nontrivial implies that  $[M : N \rtimes H] = \infty$ .  $\square$

### 3. Proof of main theorems

In this section we prove our main theorems. Our main technical tool is the following, which is [Popa and Vaes 2011, Proposition 2.7]. Before we state the result let us recall that two von Neumann subalgebras  $M_1, M_2 \subset M$  of a finite von Neumann algebra  $M$  are said to form a commuting square if  $E_{M_1}E_{M_2} = E_{M_2}E_{M_1}$ , where  $E_{M_i}$  denotes the unique trace-preserving conditional expectation from  $M$  onto  $M_i$ .

**Theorem 3.1** [Popa and Vaes 2011]. *Let  $(M, \tau)$  be a tracial von Neumann algebra with von Neumann subalgebras  $M_1, M_2 \subset M$ . Assume that  $M_1$  and  $M_2$  form a commuting square and that  $M_1$  is regular in  $M$ . If a von Neumann subalgebra  $Q \subset pMp$  is amenable relative to both  $M_1$  and  $M_2$ , then  $Q$  is amenable relative to  $M_1 \cap M_2$ .*

Notice that this theorem allows us to eliminate the case where  $Q$  is amenable over  $M_1$ . More specifically we have the following observation.

**Proposition 3.2.** *Let  $G_1$  and  $G_2$  be groups. Let  $A$  be a finite amenable von Neumann algebra with an action of  $G_1 \times G_2$ , and let  $Q \subset A \rtimes G_1 \times G_2$  be a nonamenable subalgebra. Then there exists an  $i$  such that  $Q$  is not amenable over  $A \rtimes G_i$ .*

*Proof.* If we let  $A \rtimes G_i = M_i$  then it is easy to see that  $M_1, M_2 \subset M$  form a commuting square. So if  $Q$  is amenable over both  $M_i$  we would have that it would be amenable over the intersection, which is  $A$ . Since  $A$  is amenable this would imply that  $Q$  is amenable. □

Finally combining the above results we can prove Theorem 0.1.

*Proof.* First let us mention that, for the case  $n = 1$ , this is equivalent to the primeness of  $\text{II}_1$  factors arising from Bernoulli shifts, which was proven in [Popa 2008].

Now, since we have that  $\Gamma_i \in \mathcal{WR}_{NA}$ , there is a nontrivial amenable group  $A_i$  and a nonamenable group  $H_i$  such that  $\Gamma_i = A_i \wr H_i$ . Let us note, since the  $A_i$  are nontrivial and  $H_i$  are infinite, that  $L(A_i \wr H_i)$  and  $L(A_1 \wr H_1) \bar{\otimes} \cdots \bar{\otimes} L(A_{i-1} \wr H_{i-1})$  are  $\text{II}_1$  factors. Thus we must also have that  $Q_1 \bar{\otimes} \cdots \bar{\otimes} Q_m$  is as well and thus each  $Q_i$  is a  $\text{II}_1$  factor.

Now notice that we can write  $M$  as  $M = N_i \rtimes_{\sigma} A_i \wr H_i$ , where  $N_i = L(A_1 \wr H_1) \bar{\otimes} \cdots \bar{\otimes} L(A_{i-1} \wr H_{i-1}) \bar{\otimes} L(A_{i+1} \wr H_{i+1}) \bar{\otimes} \cdots \bar{\otimes} L(A_n \wr H_n)$  and  $\sigma$  is the trivial action. Therefore, since we can view  $M$  as a crossed product by a wreath product group, we can use the above intertwining statements to determine the location of algebras which are not amenable over  $N_i$  for some  $i$ .

In order to proceed in this manner, let us define

$$\widehat{Q}_i = (Q_i)' \cap M = Q_1 \bar{\otimes} \cdots \bar{\otimes} Q_{i-1} \bar{\otimes} Q_{i+1} \bar{\otimes} \cdots \bar{\otimes} Q_k.$$

Since each  $H_i$  is nonamenable this implies, in particular, that there is a  $j$  such that  $Q_j$  is nonamenable. Moreover, by Proposition 3.2, where we let  $A = \mathbb{C}$ , we know that there is an  $i$  such that  $Q_j$  is not amenable over  $N_i$ . With this information we can then apply our results above to finish the proof.

Specifically, since  $\widehat{Q}_j$  is a regular subalgebra of  $M$ , then by Theorem 2.3 we get that  $\widehat{Q}_j \prec_M N$ .

We complete the argument by following Proposition 12 and the induction argument of the proof of Theorem 1 in [Ozawa and Popa 2004]. □

Before we prove our final theorem let us recall the following definition:

**Definition 3.3.** We say that two group  $\Gamma$  and  $\Lambda$  are *measure equivalent*,  $\Gamma \simeq_{\text{ME}} \Lambda$ , if there is a diffuse abelian von Neumann algebra,  $A$ , and free ergodic trace preserving actions,  $\sigma, \rho$ , of  $\Gamma$  and  $\Lambda$ , respectively, such that  $A \rtimes_{\sigma} \Gamma \simeq (A \rtimes_{\rho} \Lambda)^t$ , and the isomorphism takes  $A$  onto  $A^t$ .

With this definition we can now prove our final result (Theorem 0.2).

*Proof.* Our argument here follows closely a similar argument in the proof of [Chifan and Sinclair 2010, Corollary C]. For this reason we sketch the proof here but refer the reader to the cited work for any remaining details. Let  $\Gamma_1, \dots, \Gamma_n, \Lambda_1, \dots, \Lambda_m \in \mathcal{WR}_{NA}$ . Then there are nontrivial amenable groups  $A_i$  and  $B_j$  as well as nonamenable groups  $H_i$  and  $G_j$  such that  $\Gamma_i = A_i \wr H_i$  and  $\Lambda_j = B_j \wr G_j$ . Note, for all  $i$  and  $j$ ,  $\Gamma_i$  and  $\Lambda_j$  are nonamenable.

Now we know that there are actions on  $\Gamma = A_1 \wr H_1 \times \dots \times A_n \wr H_n \curvearrowright L^\infty(X)$  and  $\Lambda = K_1 \times \dots \times K_m \curvearrowright L^\infty(Y)$  such that  $M_1 = L^\infty(X) \rtimes \Gamma$  is isomorphic to  $M_2 = (L^\infty(Y) \rtimes \Lambda)^t$  via an isomorphism  $\phi : M_1 \rightarrow M_2$  such that  $\phi(L^\infty(X)) = (L^\infty(Y))^t$ . Note that the intertwining statements which we will use remain true under amplifications; thus we may assume that  $t = 1$ .

Following [ibid.] we fix the following notation. Given a subset  $F \subset \{1, \dots, n\}$ , we denote by  $\hat{\Gamma}_F$  the subgroup of  $\Gamma = A_1 \wr H_1 \times \dots \times A_n \wr H_n$  which consists of all elements with trivial  $i$ -th coordinate, for all  $i \in F$ , and similarly for  $\Lambda$ . Also for any subset  $F \subset \{1, \dots, n\}$  and  $K \subset \{1, \dots, m\}$  we define  $\hat{M}_{1,F} = L^\infty(X) \rtimes \hat{\Gamma}_F$  and  $\hat{M}_{2,K} = L^\infty(Y) \rtimes \hat{\Lambda}_K$ .

As in [ibid.] we will show that for any subset  $F \subset \{1, \dots, n\}$  there is a subset  $K \subset \{1, \dots, m\}$  with  $|F| = |K|$  such that

$$(1) \quad \phi(L(\hat{\Gamma}_F)) \prec \hat{M}_{2,k}.$$

We will prove this via induction on  $|F|$ . For  $|F| = 1$  we are considering  $L(\hat{\Gamma}_i)$ . As in the proof of the previous theorem, since the  $\phi(L(\Gamma_i))$  are nonamenable, there is a  $j$  such that  $\phi(L(\Gamma_i))$  is nonamenable over  $\hat{M}_{2,\{j\}}$ . Now by the proof of Theorem 2.3 this implies that  $\phi(L(\Gamma_i))' \cap M_2 \prec \hat{M}_{2,\{j\}} \rtimes G_j$ , and, since we have that  $\phi(L(\hat{\Gamma}_i)) \subset \phi(L(\Gamma_i))' \cap M_2$ , we get that  $\phi(L(\hat{\Gamma}_i)) \prec \hat{M}_{2,\{j\}} \rtimes G_j$ .

Thus by [Chifan et al. 2012, Lemma 2.2] we have that  $\phi(L^\infty(X) \rtimes \hat{\Gamma}_i) \prec \hat{M}_{2,\{j\}} \rtimes G_j$ . Now since  $\phi(L^\infty(X) \rtimes \hat{\Gamma}_i)$  is a regular subalgebra we have by Theorem 2.3 that  $\phi(L^\infty(X) \rtimes \hat{\Gamma}_i) \prec \hat{M}_{2,\{j\}}$ . This proves the base case and the inductive case follows exactly as in [Chifan and Sinclair 2010].

Again following [ibid.] we can apply the same logic to  $\phi^{-1}$  to get that for each  $i \in F$  there is a  $\rho(i) \in K$  such that  $\phi(L^\infty(X) \rtimes \hat{\Gamma}_i) \prec \hat{M}_{2,\{j\}}$  and for each  $\rho(i) \in K$  there is a  $\pi(\rho(i)) \in F$  with

$$\phi(L^\infty(X) \rtimes \Gamma_i) \prec L^\infty(Y) \rtimes \Lambda_{\rho(i)}$$

and

$$\phi^{-1}(L^\infty(y) \rtimes \Lambda_{\rho(i)}) \prec L^\infty(X) \rtimes \Gamma_{\pi(\rho(i))}.$$

Thus we have

$$\phi(L^\infty(X) \rtimes \Gamma_i) \prec L^\infty(X) \rtimes \Gamma_{\pi(\rho(i))},$$

and so we have that  $\pi$  and  $\rho$  are permutations. Thus using [Ioana et al. 2008, Proposition 8.4] we get unitaries  $u_i$  such that

$$(2) \quad u_i \phi(L^\infty(X) \rtimes \Gamma_i) u_i^* = L^\infty(Y) \rtimes \Lambda_{\rho(i)}.$$

This further gives that  $\phi_{u_i} = \text{Ad}(u_i) \circ \phi$  is an isomorphism from  $L^\infty(X) \rtimes \Gamma_i$  onto  $L^\infty(Y) \rtimes \Lambda_{\rho(i)}$  which satisfies

$$\phi_{u_i}(a) u_i = u_i \phi(a)$$

for all  $a \in L^\infty(X)$ .

Now we would like to finish the proof by showing that we can map the Cartan subalgebras onto each other. Toward this goal let us consider  $L^\infty(Y) \rtimes (\Lambda_{\rho(i)} \times \hat{\Lambda}_{\rho(i)}) = (L^\infty(Y) \rtimes \Lambda_{\rho(i)}) \rtimes \hat{\Lambda}_{\rho(i)}$ . Then we can consider the Fourier decomposition  $u = \sum_{\lambda \in \hat{\Lambda}_{\rho(i)}} x_\lambda v_\lambda$  with  $x_\lambda \in L^\infty(Y) \rtimes \Lambda_{\rho(i)}$  and, using the above equation, there exists a nonzero element  $x_\lambda \in L^\infty(Y) \rtimes \Lambda_{\rho(i)}$  such that for all  $a \in L^\infty(X)$  we have

$$\phi_{u_i}(a) x_\lambda = x_\lambda \sigma_\lambda(\phi(a)),$$

where  $\sigma_\lambda$  represents the actions of  $v_\lambda$  on  $L^\infty(Y) \rtimes \Lambda_{\rho(i)}$ .

Now we can take the polar decomposition of  $x_\lambda$  to get a partial isometry  $w_\lambda$  such that

$$(3) \quad \phi_{u_i}(a) w_\lambda = w_\lambda \sigma_\lambda(\phi(a)).$$

Notice that the left side of the above equation is  $\phi_{u_i}(L^\infty(X))$  while the right side is  $\phi(L^\infty(X)) = L^\infty(Y)$ . Thus (3) implies that we know  $\phi_{u_i}(A) \prec_{L^\infty(Y) \rtimes \Lambda_{\rho(i)}} L^\infty(Y)$ . Since they are both Cartan subalgebras then by [Popa 2006a, Theorem A2] we can extend this to unitary conjugacy and thus get our result.  $\square$

### Acknowledgement

We would like to thank Sorin Popa for suggesting this problem and for his encouragement throughout.

### References

- [Berbec and Vaes 2012] M. Berbec and S. Vaes, “ $W^*$ -superrigidity for group von Neumann algebras of left-right wreath products”, preprint, 2012. arXiv 1210.0336
- [Chifan and Houdayer 2010] I. Chifan and C. Houdayer, “Bass–Serre rigidity results in von Neumann algebras”, *Duke Math. J.* **153**:1 (2010), 23–54. MR 2012a:46126 Zbl 1201.46057
- [Chifan and Sinclair 2010] I. Chifan and T. Sinclair, “On the structural theory of  $\text{II}_1$  factors of negatively curved groups”, preprint, 2010. arXiv 1103.4299v2
- [Chifan et al. 2012] I. Chifan, S. Popa, and J. O. Sizemore, “Some OE- and  $W^*$ -rigidity results for actions by wreath product groups”, *J. Funct. Anal.* **263**:11 (2012), 3422–3448. MR 2984072 Zbl 06115544



- [Connes 1976] A. Connes, “Classification of injective factors: cases  $\text{II}_1$ ,  $\text{II}_\infty$ ,  $\text{III}_\lambda$ ,  $\lambda \neq 1$ ”, *Ann. of Math.* (2) **104**:1 (1976), 73–115. MR 56 #12908 Zbl 0343.46042
- [Ge 1998] L. Ge, “Applications of free entropy to finite von Neumann algebras, II”, *Ann. of Math.* (2) **147**:1 (1998), 143–157. MR 99c:46068 Zbl 0924.46050
- [Ioana 2007] A. Ioana, “Rigidity results for wreath product  $\text{II}_1$  factors”, *J. Funct. Anal.* **252**:2 (2007), 763–791. MR 2008j:46046 Zbl 1134.46041
- [Ioana et al. 2008] A. Ioana, J. Peterson, and S. Popa, “Amalgamated free products of weakly rigid factors and calculation of their symmetry groups”, *Acta Math.* **200**:1 (2008), 85–153. MR 2009a:46119 Zbl 1149.46047
- [Ioana et al. 2013] A. Ioana, S. Popa, and S. Vaes, “A class of superrigid group von Neumann algebras”, *Ann. of Math.* (2) **178**:1 (2013), 231–286. arXiv 1007.1412
- [Monod and Shalom 2006] N. Monod and Y. Shalom, “Orbit equivalence rigidity and bounded cohomology”, *Ann. of Math.* (2) **164**:3 (2006), 825–878. MR 2007k:37007 Zbl 1129.37003
- [Ozawa 2004] N. Ozawa, “Solid von Neumann algebras”, *Acta Math.* **192**:1 (2004), 111–117. MR 2005e:46115 Zbl 1072.46040
- [Ozawa and Popa 2004] N. Ozawa and S. Popa, “Some prime factorization results for type  $\text{II}_1$  factors”, *Invent. Math.* **156**:2 (2004), 223–234. MR 2005g:46117 Zbl 1060.46044
- [Ozawa and Popa 2010] N. Ozawa and S. Popa, “On a class of  $\text{II}_1$  factors with at most one Cartan subalgebra”, *Ann. of Math.* (2) **172**:1 (2010), 713–749. MR 2011j:46101 Zbl 1201.46054
- [Peterson 2009] J. Peterson, “ $L^2$ -rigidity in von Neumann algebras”, *Invent. Math.* **175**:2 (2009), 417–433. MR 2010b:46128 Zbl 1170.46053
- [Popa 1983] S. Popa, “Orthogonal pairs of  $*$ -subalgebras in finite von Neumann algebras”, *J. Operator Theory* **9**:2 (1983), 253–268. MR 84h:46077 Zbl 0521.46048
- [Popa 2006a] S. Popa, “On a class of type  $\text{II}_1$  factors with Betti numbers invariants”, *Ann. of Math.* (2) **163**:3 (2006), 809–899. MR 2006k:46097 Zbl 1120.46045
- [Popa 2006b] S. Popa, “Some rigidity results for non-commutative Bernoulli shifts”, *J. Funct. Anal.* **230**:2 (2006), 273–328. MR 2007b:46106 Zbl 1097.46045
- [Popa 2006c] S. Popa, “Strong rigidity of  $\text{II}_1$  factors arising from malleable actions of  $w$ -rigid groups, I”, *Invent. Math.* **165**:2 (2006), 369–408. MR 2007f:46058 Zbl 1120.46043
- [Popa 2007] S. Popa, “Deformation and rigidity for group actions and von Neumann algebras”, pp. 445–477 in *International Congress of Mathematicians* (Madrid, 2006), vol. I, edited by M. Sanz-Solé et al., Eur. Math. Soc., Zürich, 2007. MR 2008k:46186 Zbl 1132.46038
- [Popa 2008] S. Popa, “On the superrigidity of malleable actions with spectral gap”, *J. Amer. Math. Soc.* **21**:4 (2008), 981–1000. MR 2009e:46056 Zbl 1222.46048
- [Popa and Vaes 2011] S. Popa and S. Vaes, “Unique Cartan decomposition for  $\text{II}_1$  factors arising from arbitrary actions of free groups”, preprint, 2011. arXiv 1111.6951v1
- [Sako 2009] H. Sako, “Measure equivalence rigidity and bi-exactness of groups”, *J. Funct. Anal.* **257**:10 (2009), 3167–3202. MR 2010k:37008 Zbl 1256.37002
- [Vaes 2011] S. Vaes, “Rigidity for von Neumann algebras and their invariants”, pp. 1624–1650 in *International Congress of Mathematicians* (Hyderabad, 2010), vol. III, edited by R. Bhatia et al., Hindustan Book Agency, New Delhi, 2011. MR 2012g:46006 Zbl 1235.46058
- [Voiculescu 1993] D. Voiculescu, “The analogues of entropy and of Fisher’s information measure in free probability theory, I”, *Comm. Math. Phys.* **155**:1 (1993), 71–92. MR 94k:46137 Zbl 0781.60006
- [Voiculescu 1994] D. Voiculescu, “The analogues of entropy and of Fisher’s information measure in free probability theory, II”, *Invent. Math.* **118**:3 (1994), 411–440. MR 96a:46117 Zbl 0820.60001

[Voiculescu 1996] D. Voiculescu, “The analogues of entropy and of Fisher’s information measure in free probability theory, III: The absence of Cartan subalgebras”, *Geom. Funct. Anal.* **6**:1 (1996), 172–199. MR 96m:46119 Zbl 0856.60012

Received October 19, 2011. Revised January 26, 2013.

J. OWEN SIZEMORE  
DEPARTMENT OF MATHEMATICS  
UNIVERSITY OF WISCONSIN-MADISON  
480 LINCOLN DR.  
UNIVERSITY OF WISCONSIN  
MADISON, WI 53706  
UNITED STATES  
josizemore@math.wisc.edu

ADAM WINCHESTER  
UCLA  
MATH SCIENCES BUILDING  
LOS ANGELES, CA 90095  
UNITED STATES  
lagwadam@math.ucla.edu

## ON VOLUME GROWTH OF GRADIENT STEADY RICCI SOLITONS

GUOFANG WEI AND PENG WU

**In this paper we study volume growth of gradient steady Ricci solitons. We show that if the potential function satisfies a uniform condition, then the soliton has at most euclidean volume growth.**

### 1. Introduction

$(M^n, g)$  is a gradient Ricci soliton if there is a smooth function  $f : M \rightarrow \mathbb{R}$  and constant  $\lambda \in \mathbb{R}$  such that

$$(1-1) \quad \text{Ric} + \text{Hess } f = \lambda g.$$

We refer to  $f$  as the potential function. The soliton is called shrinking, steady, and expanding when  $\lambda > 0$ ,  $\lambda = 0$ , and  $\lambda < 0$  respectively.

Ricci solitons are self-similar solutions of the Ricci flow, and play an important role in the study of singularity formation. They are also natural extensions of Einstein manifolds, and special cases of smooth metric measure spaces.

Volume growth of gradient Ricci solitons is of particular interest to mathematicians. Estimates of the potential functions plays an important role in the study of volume growth. Hamilton [1995] proved the following identity for gradient Ricci solitons:

$$R + |\nabla f|^2 - 2\lambda f = \Lambda,$$

where  $\Lambda$  is a constant, and  $R$  is the scalar curvature.

For gradient shrinking Ricci solitons, the answer is complete. Perelman [2003] and Cao and Zhou [2010] proved that  $f$  always grows quadratically. Cao and Zhou [2010] further proved that any gradient Ricci shrinking soliton has at most euclidean volume growth. Recently, Munteanu and Wang [2012] proved that any noncompact gradient Ricci shrinking soliton has at least linear volume growth.

For gradient steady Ricci solitons, B. L. Chen [2009] proved that  $R \geq 0$ . Hence,  $\Lambda \geq 0$ , and equal to zero if and only if  $f$  is constant and  $(M, g)$  is Ricci flat. When

---

*MSC2010:* 53CXX.

*Keywords:* Ricci solitons, volume growth.

$\Lambda > 0$  we can assume  $\Lambda = 1$  after scaling; that is,

$$(1-2) \quad R + |\nabla f|^2 = 1.$$

Combined with the trace of the steady Ricci soliton equation  $R + \Delta f = 0$ , we have

$$(1-3) \quad \Delta f - |\nabla f|^2 = -1.$$

Therefore,  $f$  has no local minimum. Equation (1-2) and  $R \geq 0$  give  $|\nabla f| \leq 1$ . Namely,  $f$  decays at most linearly.

Cao and Chen [2012] proved that  $f$  decays linearly when Ricci curvature is positive, and  $R$  attains its maximum at some point. However, the simple example of  $\mathbb{R}^2$  with the canonical metric  $g_0$  and  $f(x) = x_1$  shows that this is not the case;  $f$  is constant along the  $x_2$  direction. Note that the Riemannian product of any two steady gradient Ricci solitons is still a steady gradient Ricci soliton. Hence, a steady gradient Ricci soliton multiplied with a trivial one ( $f$  is a constant) will have constant direction. Though one can take the product of two shrinking ones, all trivial shrinking ones are compact, so they will not give a constant direction by taking a product. Munteanu and Sesum [2013] and Wu [2013] independently showed that the infimum of  $f$  does decay linearly. In fact,

$$(1-4) \quad -r \leq \inf_{y \in \partial B(x,r)} f(y) - f(x) \leq -r + \sqrt{2n}(\sqrt{r} + 1), \quad r \gg 1.$$

In particular,  $\liminf_{y \rightarrow \infty} R(y) = 0$ . See also [Fernández-López and García-Río 2011; Chow et al. 2011].

We note that among all known examples of steady gradient Ricci solitons, the infimum of  $f$  is like  $-r + O(\ln r)$ . See the survey article [Cao 2010] for a list of examples. One naturally asks if one can improve the second order term in (1-4) from  $\sqrt{r}$  to  $\ln r$ . We show this is indeed the case for a large class of steady gradient Ricci solitons. To study the second order term, write the potential function in polar coordinates:

$$f(r, \theta) = -r + \phi(r, \theta),$$

where  $r(\cdot) = d(x, \cdot)$  for some  $x \in M^n$ ,  $\theta \in S^{n-1}$ . Without loss of generality, we assume  $\phi(0, \theta) = 0$  by adding a constant to  $f$ . Since  $|\nabla f| \leq 1$ , we have  $|\partial f / \partial r| \leq 1$ , so  $f(r) \geq -r$ ,  $\phi(r) \geq 0$  and  $\phi(r, \theta)$  are nondecreasing in  $r$  for any fixed  $\theta$ . We show that the estimate (1-4) can be improved to  $\ln r$  if  $\phi$  in one direction is comparable to the minimum of  $\phi$  among all spherical directions for all large  $r$ .

**Theorem 1.1.** *Let  $(M^n, g, f)$  be a complete gradient steady Ricci soliton satisfying (1-2). Assume that there exist  $\theta_0 \in S^{n-1}$ , and constants  $C_1 \geq 0, C_2 \geq 0$  such that*

$$(1-5) \quad \int_0^r (\phi(r, \theta_0) - \phi(t, \theta_0)) dt \leq C_1 \min_{\theta \in S^{n-1}} \int_0^r (\phi(r, \theta) - \phi(t, \theta)) dt + C_2 r$$

for sufficiently large  $r$ . Then for any  $x \in M^n$ , there exist constants  $C \geq 0$  and  $r_0 > 0$  such that for  $r \geq r_0$ ,

$$(1-6) \quad -r \leq \inf_{y \in \partial B(x,r)} f(y) - f(x) \leq -r + \left(\frac{n}{2}C_1 + C_2\right) \ln r + C.$$

**Remark 1.2.** All known examples of gradient steady Ricci solitons satisfy the condition (1-5). We suspect that the estimate (1-6) holds for all gradient steady Ricci solitons.

In [Munteanu and Sesum 2013], it was proven that any gradient steady Ricci soliton has at least linear volume growth, and at most a growth rate of  $e^{\sqrt{r}}$ . We show that if the potential function satisfies a uniform condition in the spherical directions, then the gradient steady Ricci soliton has at most euclidean volume growth.

**Theorem 1.3.** Let  $(M^n, g, f)$  be a complete gradient steady Ricci soliton satisfying (1-2). Assume that there exist constants  $C_1, C_2 \geq 0$  such that

$$(1-7) \quad \max_{\theta \in S^{n-1}} \int_0^r (\phi(r, \theta) - \phi(t, \theta)) dt \leq C_1 \min_{\theta \in S^{n-1}} \int_0^r (\phi(r, \theta) - \phi(t, \theta)) dt + C_2 r$$

for sufficiently large  $r$ . Then for any  $x \in M^n$ , there exist constants  $C \geq 0$  and  $r_0 > 0$  such that for any  $r \geq r_0$ ,

$$(1-8) \quad -r \leq f(y) - f(x) \leq -r + C \ln r$$

for any  $y \in \partial B(x, r)$ . Moreover, the soliton has at most euclidean volume growth; that is, for any  $x \in M^n$  there exists  $r_0 > 0$ , and for any  $r \geq r_0$ ,

$$\text{Vol } B(x, r) \leq Cr^n.$$

If, in addition,  $\phi(r) \geq \delta \ln r$  for large  $r$ , then

$$\text{Vol } B(x, r) \leq Cr^{n-\delta}.$$

**Remark 1.4.** (1) If  $\phi$  increases uniformly along all spherical directions; that is,  $\max_{\theta} \partial\phi/\partial r \leq C \min_{\theta} \partial\phi/\partial r$ , where  $\theta \in S^{n-1}$ , then  $\phi$  satisfies (1-7) with  $C_1 = C$  and  $C_2 = 0$ .

(2) Theorem 1.3 can be considered an analogue of the volume growth theorem of [Cao and Zhou 2010], valid for gradient shrinking Ricci solitons. As the potential function for such solitons automatically satisfies a uniform condition, here too we need to impose a uniform condition for gradient steady Ricci solitons.

(3) If the soliton is rectifiable (see [Petersen and Wylie 2009])—i.e.,  $f$  is the distance function from a set—then  $\phi$  satisfies (1-7) with  $C_1 = 1$  if the set is bounded (this is the case with all nonproduct examples).

To prove the results, the following estimate for  $\phi$ , which holds for all gradient steady Ricci solitons, is the key:

**Proposition 1.5.** *Let  $(M^n, g, f)$  be a complete gradient steady Ricci soliton satisfying (1-2). Then*

$$(1-9) \quad \min_{y \in \partial B(x,r)} \int_0^r (\phi(y) - \phi(t)) dt \leq \frac{n}{2} (r + \sqrt{r}) + o\left(\frac{1}{r}\right).$$

This estimate improves the one in [Wu 2013]. In the next section we derive a volume comparison for the solitons by adapting the volume comparison for smooth metric measure spaces in [Wei and Wylie 2009]. Then we prove Proposition 1.5 by combining this with (1-3). In Section 3 we prove the main theorems using this estimate and an ODE.

### 2. The preliminary estimate

In this section we prove Proposition 1.5 by applying a weighted volume comparison argument for smooth metric measure spaces as in [Wei and Wylie 2009; Wu 2013].

Recall that a smooth metric measure space is a triple  $(M^n, g, e^{-f} d\text{vol}_g)$ , where  $(M^n, g)$  is a smooth Riemannian manifold, and  $f : M^n \rightarrow \mathbb{R}$  is a smooth function. Write the volume element in polar coordinates  $d\text{vol} = J(r, \theta) dr d\theta$ . Define the weighted volume element as  $J_f(r, \theta) = e^{-f} J(r, \theta)$  and the weighted volume as  $\text{Vol}_f B(x, r) = \int_{B(x,r)} e^{-f} d\text{vol}$ .

Wei and Wylie [2009] obtained the following  $f$ -volume comparison theorem for smooth metric measure spaces:

**Theorem 2.1** ( $f$ -volume comparison). *Suppose  $(M^n, g, e^{-f} d\text{vol})$  is a smooth metric measure space with  $\text{Ric}_f \geq (n - 1)H$ . Fix  $x \in M$ . If  $|f| \leq \Lambda$ , then for  $R \geq r > 0$  (and  $R \leq \pi/4\sqrt{H}$  if  $H > 0$ ),*

$$\frac{\text{Vol}_f B(x, R)}{\text{Vol}_f B(x, r)} \leq \frac{V_H^{n+4\Lambda}(B_R)}{V_H^{n+4\Lambda}(B_r)}$$

where  $V_H^n(B_r)$  is the volume of the ball of radius  $r$  in  $M_H^n$  (the simply connected model space of dimension  $n$  with constant sectional curvature  $H$ ).

One observes that the dimension of the model space in the volume comparison depends on the potential function  $f$ . Further investigation of the dimension will lead to Proposition 1.5.

Denote the  $f$ -mean curvature by  $m_f = (\ln J_f)'$ . For  $0 < r_1 \leq r_2$ , let  $A(x, r_1, r_2) = \{y | r_1 \leq d(x, y) \leq r_2\}$  be the annulus, and

$$a = \min_{y \in A(x,r_1,r_2)} \frac{2}{r(y)^2} \int_0^{r(y)} (\phi(y) - \phi(t)) dt.$$

Clearly  $a \geq 0$ . By (1-4), we have  $a \leq C/\sqrt{r_1}$  for  $r_1 \gg 1$ . For the rest we assume  $r_1 \gg 1$  and therefore we can assume  $a < 1$ .

**Proposition 2.2.** *For a gradient steady Ricci soliton, we have*

$$(2-1) \quad m_f(r, \theta) \leq \frac{n-1}{r} + 1 - \frac{2}{r^2} \int_0^r (\phi(r, \theta) - \phi(t, \theta)) dt \leq \frac{n-1}{r} + 1,$$

$$(2-2) \quad \frac{\text{Vol}_f \partial B(x, r_2)}{\text{Vol}_f A(x, r_1, r_2)} \leq \frac{n/r_2 + 1 - a}{1 - (r_1/r_2)^{n+(1-a)r_2}}.$$

*Proof.* For a smooth metric space  $(M^n, g, f)$  with  $\text{Ric}_f \geq 0$ , recall the following estimate for  $m_f$  from [Wei and Wylie 2009, (3.19)]:

$$m_f(r, \theta) \leq \frac{n-1}{r} + \frac{2}{r^2} \int_0^r (f(t) - f(r)) dt.$$

Plugging in  $f = -r + \phi$  gives (2-1).

Now let

$$\bar{m}(r) = \begin{cases} \frac{n-1}{r} + 1 & \text{if } r \leq r_1, \\ \frac{n-1+(1-a)r_2}{r} & \text{if } r_1 < r \leq r_2. \end{cases}$$

Then

$$(2-3) \quad m_f(r) \leq \bar{m}(r) \text{ for } 0 < r \leq r_2.$$

Let  $\bar{A}(r) = e^{\int_0^r \bar{m}(t) dt}$  and  $\bar{V}(r_0, r) = \int_{r_0}^r \bar{A}(t) dt$ . From the mean curvature relation (2-3), we have  $(A_f/\bar{A})' \leq 0$ ; therefore

$$\frac{\text{Vol}_f \partial B(x, r_2)}{\text{Vol}_f A(x, r_1, r_2)} \leq \frac{\bar{A}(r_2)}{\bar{V}(r_1, r_2)}.$$

We compute

$$\begin{aligned} \frac{\bar{A}(r_2)}{\bar{V}(r_1, r_2)} &= \frac{e^{\int_0^{r_2} \bar{m}(t) dt}}{\int_{r_1}^{r_2} e^{\int_0^s \bar{m}(t) dt} ds} = \frac{e^{\int_{r_1}^{r_2} \bar{m}(t) dt}}{\int_{r_1}^{r_2} e^{\int_{r_1}^s \bar{m}(t) dt} ds} \\ &= \frac{(r_2/r_1)^{n-1+(1-a)r_2}}{\int_{r_1}^{r_2} (s/r_1)^{n-1+(1-a)r_2} ds} = \frac{n/r_2 + 1 - a}{1 - (r_1/r_2)^{n+(1-a)r_2}}. \end{aligned}$$

This gives (2-2). □

*Proof of Proposition 1.5.* Integrating (1-3) and using  $|\nabla f| \leq 1$  we have, for any  $x \in M$ ,

$$\begin{aligned} \int_{B(x,r)} 1 \cdot e^{-f} d\text{vol} &= - \int_{B(x,r)} (\Delta f - |\nabla f|^2) \cdot e^{-f} d\text{vol} \\ &= - \int_{\partial B(x,r)} \frac{\partial f}{\partial n} e^{-f} d\text{vol} \leq \int_{\partial B(x,r)} e^{-f} d\text{vol}. \end{aligned}$$

Therefore,

$$(2-4) \quad \frac{\text{Vol}_f \partial B(x, r)}{\text{Vol}_f B(x, r)} \geq 1.$$

Combining (2-2) and (2-4) we have

$$a \leq \frac{n}{r_2} + \left(\frac{r_1}{r_2}\right)^{n+(1-a)r_2}.$$

Let  $r_1 = r$  and  $r_2 = r + \sqrt{r}$ . Then  $r_1/r_2 = (1 + 1/\sqrt{r})^{-1}$ . When  $r$  is large,

$$\left(1 + \frac{1}{\sqrt{r}}\right)^{-(n+(1-a)(r+\sqrt{r}))} = O(e^{-(1-a)\sqrt{r}}).$$

Therefore, for all  $r$  large enough,

$$a = \min_{y \in A(x, r, r+\sqrt{r})} \frac{2}{r(y)^2} \int_0^{r(y)} (\phi(y) - \phi(t)) dt \leq \frac{n}{r + \sqrt{r}} + O(e^{-(1-a)\sqrt{r}}).$$

Suppose the minimum above is attained at  $y_0 = (r_0, \theta_1)$  with  $r \leq r_0 \leq r + \sqrt{r}$ . Then

$$\begin{aligned} \min_{y \in \partial B(x, r)} \int_0^r (\phi(y) - \phi(t)) dt &\leq \int_0^{r_0} (\phi(y_0) - \phi(t)) dt \\ &\leq \frac{r_0^2}{2} \left( \frac{n}{r + \sqrt{r}} + O(e^{-(1-a)\sqrt{r}}) \right) \\ &\leq \frac{n}{2} (r + \sqrt{r}) + o\left(\frac{1}{r}\right). \quad \square \end{aligned}$$

### 3. Proof of main results

*Proof of Theorem 1.1.* From (1-9) and (1-5), we have

$$(3-1) \quad \int_0^r (\phi(r, \theta_0) - \phi(t, \theta_0)) dt \leq \frac{nC_1}{2}(r + \sqrt{r}) + C_2r + o\left(\frac{1}{r}\right).$$

For simplicity, when there is no confusion we omit  $\theta_0$  in the function. Let  $\Phi(r) = \int_0^r \phi(t) dt$ ; then (3-1) can be written as

$$(3-2) \quad \Phi'(r) - \frac{1}{r}\Phi(r) \leq \frac{nC_1}{2} + C_2 + O\left(\frac{1}{\sqrt{r}}\right).$$

Multiplying by the integrating factor  $1/r$  and integrating from some fixed  $t_0 \gg 1$  to  $r$ , we get

$$\frac{\Phi(r)}{r} \leq \left(\frac{nC_1}{2} + C_2\right) \ln r + C_3.$$



So, we have

$$\begin{aligned} \phi(r, \theta_0) &= \Phi'(r, \theta_0) \leq \frac{\Phi(r, \theta_0)}{r} + \frac{nC_1}{2} + C_2 + O\left(\frac{1}{\sqrt{r}}\right) \\ &\leq \left(\frac{nC_1}{2} + C_2\right) \ln r + C_4 \\ f(r, \theta_0) &= -r + \phi(r, \theta_0) \leq -r + \left(\frac{nC_1}{2} + C_2\right) \ln r + C_4. \end{aligned}$$

This gives (1-6). □

*Proof of Theorem 1.3.* From (1-9) and (1-7), we have

$$\int_0^r (\phi(r, \theta) - \phi(t, \theta)) dt \leq \frac{nC_1}{2} (r + \sqrt{r}) + C_2r + o\left(\frac{1}{r}\right)$$

for all  $\theta \in S^{n-1}$ . Therefore, (1-6) holds for all  $y$ . Namely, for all  $y \in \partial B(x, r)$ ,

$$-r \leq f(y) - f(x) \leq -r + \left(\frac{nC_1}{2} + C_2\right) \ln r + C_4.$$

By (2-1), for all  $r > 0$ ,

$$\begin{aligned} \frac{\partial}{\partial r} \ln J &= m_f(r) + \langle \nabla f, \nabla r \rangle \\ &\leq \frac{n-1}{r} + 1 - \frac{2}{r} \phi(r) + \frac{2}{r^2} \int_0^r \phi(t) dt + \langle \nabla f, \nabla r \rangle. \end{aligned}$$

Integrating from 1 to  $r$  and performing integration by parts for the double integral, we get

$$\begin{aligned} (3-3) \quad &\ln J(r) - \ln J(1) \\ &\leq (n-1) \ln r + (r-1) - \int_1^r \frac{2}{s} \phi(s) ds + \left(-\frac{2}{s} \int_0^s \phi(t) dt\right) \Big|_1^r \\ &\quad + \int_1^r \frac{2}{s} \phi(s) ds + f(r) - f(1) \\ &= (n-1) \ln r + \phi(r) - \frac{2}{r} \int_0^r \phi(t) dt + 2 \int_0^1 \phi(t) dt - f(1) - 1 \\ &= (n-1) \ln r - \phi(r) + 2 \left(\phi(r) - \frac{1}{r} \int_0^r \phi(t) dt\right) + 2 \int_0^1 \phi(t) dt - f(1) - 1. \end{aligned}$$

Using (3-2) we have, for large  $r$ ,

$$\ln J(r) \leq (n-1) \ln r - \phi(r) + C \leq (n-1) \ln r + C.$$

Hence,

$$J(r) \leq e^C e^{(n-1) \ln r} = e^C r^{n-1},$$

and the volume of a geodesic ball centered at  $x$  satisfies

$$\text{Vol}B(x, r) \leq C'r^n.$$

If further  $\phi(s) \geq \delta \ln s$ , then we have

$$J(r) \leq Cr^{n-1} \exp(-\phi(r)) \leq Cr^{n-\delta-1},$$

therefore the volume growth is strictly less than euclidean volume growth:

$$\text{Vol}B(x, r) \leq Cr^{n-\delta}. \quad \square$$

For general gradient steady Ricci solitons, the estimate of a potential function can be reduced to the following:

**Question 3.1.** Suppose  $\phi$  is nondecreasing along any minimal geodesic starting from  $x$ . Assume that for sufficiently large  $r$ ,  $\inf_{y \in \partial B(x, r)} \phi(y) \leq C\sqrt{r}$ , and

$$\inf_{y \in \partial B(x, r)} \int_1^r (\phi(y) - \phi(\gamma_y(t))) dt \leq \frac{nr}{2}.$$

Does the following hold?

$$\inf_{y \in \partial B(x, r)} \phi(y) \leq C \ln r$$

**Remark 3.2.** From (3-3), we see that if

$$-r \leq f(y) - f(x) \leq -r + C \ln r$$

for  $y \in \partial B(x, r)$ , then for any  $x \in M^n$ , there exists  $r_0 > 0$  such that for any  $r \geq r_0$ ,

$$\text{Vol}B(x, r) \leq C'r^{n+C}.$$

### Acknowledgements

The second author would like to thank Professor Thomas Sideris for helpful discussions.

### References

- [Cao 2010] H.-D. Cao, "Recent progress on Ricci solitons", pp. 1–38 in *Recent advances in geometric analysis* (Taipei, 2007), edited by Y.-I. Lee et al., Adv. Lect. Math. **11**, International Press, Somerville, MA, 2010. MR 2011d:53061 Zbl 1201.53046 arXiv 0908.2006
- [Cao and Chen 2012] H.-D. Cao and Q. Chen, "On locally conformally flat gradient steady Ricci solitons", *Trans. Amer. Math. Soc.* **364**:5 (2012), 2377–2391. MR 2888210 Zbl 1245.53038
- [Cao and Zhou 2010] H.-D. Cao and D. Zhou, "On complete gradient shrinking Ricci solitons", *J. Differential Geom.* **85**:2 (2010), 175–185. MR 2011k:53040 Zbl 1246.53051
- [Chen 2009] B.-L. Chen, "Strong uniqueness of the Ricci flow", *J. Differential Geom.* **82**:2 (2009), 363–382. MR 2010h:53095 Zbl 1177.53036

- [Chow et al. 2011] B. Chow, P. Lu, and B. Yang, “Lower bounds for the scalar curvatures of noncompact gradient Ricci solitons”, *C. R. Math. Acad. Sci. Paris* **349**:23–24 (2011), 1265–1267. MR 2861997 Zbl 1230.53036
- [Fernández-López and García-Río 2011] M. Fernández-López and E. García-Río, “Maximum principles and gradient Ricci solitons”, *J. Differential Equations* **251**:1 (2011), 73–81. MR 2012d:53136 Zbl 1217.53042
- [Hamilton 1995] R. S. Hamilton, “The formation of singularities in the Ricci flow”, pp. 7–136 in *Proceedings of the conference on geometry and topology* (Cambridge, MA, 1993), edited by C. C. Hsiung and S.-T. Yau, *Surveys in Differential Geometry* **2**, International Press, Somerville, MA, 1995. MR 97e:53075 Zbl 0867.53030
- [Munteanu and Sesum 2013] O. Munteanu and N. Sesum, “On gradient Ricci solitons”, *J. Geom. Anal.* **23**:2 (2013), 539–561. MR 3023848
- [Munteanu and Wang 2012] O. Munteanu and J. Wang, “Analysis of weighted Laplacian and applications to Ricci solitons”, *Comm. Anal. Geom.* **20**:1 (2012), 55–94. MR 2903101 Zbl 1245.53039 arXiv 1112.3027
- [Perelman 2003] G. Y. Perelman, “Ricci flow with surgery on three manifolds”, preprint, 2003. Zbl 1130.53002 arXiv math.DG/0303109
- [Petersen and Wylie 2009] P. Petersen and W. Wylie, “On gradient Ricci solitons with symmetry”, *Proc. Amer. Math. Soc.* **137**:6 (2009), 2085–2092. MR 2010a:53073 Zbl 1168.53021
- [Wei and Wylie 2009] G. Wei and W. Wylie, “Comparison geometry for the Bakry–Emery Ricci tensor”, *J. Differential Geom.* **83**:2 (2009), 377–405. MR 2011a:53064 Zbl 1189.53036
- [Wu 2013] P. Wu, “On the potential function of gradient steady Ricci solitons”, *J. Geom. Anal.* **23**:1 (2013), 221–228. MR 3010278 Zbl 06136841

Received July 12, 2012. Revised November 9, 2012.

GUOFANG WEI  
MATH DEPARTMENT  
UC SANTA BARBARA  
SANTA BARBARA, CA 93106  
UNITED STATES  
wei@math.ucsb.edu

PENG WU  
MATH DEPARTMENT  
CORNELL UNIVERSITY  
ITHACA, NY 14853  
UNITED STATES  
wupenguin@math.cornell.edu



## CLASSIFICATION OF MODULI SPACES OF ARRANGEMENTS OF NINE PROJECTIVE LINES

FEI YE

**In the study of line arrangements, searching for minimal examples of line arrangements whose fundamental groups are not combinatorially invariant is a very interesting and hard problem. It is known that such a minimal arrangement must have at least 9 lines. In this paper, we extend the number to 10 by a new method. We classify arrangements of 9 projective lines according to the irreducibility of their moduli spaces and show that fundamental groups of complements of arrangements of 9 projective lines are combinatorially invariant. The idea and results have been used to classify arrangements of 10 projective lines.**

### 1. Introduction

A hyperplane arrangement  $\mathcal{A} = \{L_1, L_2, \dots, L_n\}$  in  $\mathbb{C}\mathbb{P}^r$  is a finite collection of hyperplanes. We call  $M(\mathcal{A}) = \mathbb{C}\mathbb{P}^r \setminus (\bigcup_{L \in \mathcal{A}} L)$  the complement of  $\mathcal{A}$ . The set  $L(\mathcal{A}) = \{\bigcap_{i \in S} L_i \mid S \subseteq \{1, 2, \dots, n\}\}$  partially ordered by reverse inclusion is called the *intersection lattice* of  $\mathcal{A}$ . Let  $\mathcal{A}_1$  and  $\mathcal{A}_2$  be two arrangements of  $n$  hyperplanes. We say that intersection lattices  $L(\mathcal{A}_1)$  and  $L(\mathcal{A}_2)$  are isomorphic, denoted by  $L(\mathcal{A}_1) \sim L(\mathcal{A}_2)$ , if there is a permutation  $\phi$  of the set  $\{1, 2, \dots, n\}$  such that

$$\dim \bigcap_{\substack{i \in S \\ G_i \in \mathcal{A}_1}} G_i = \dim \bigcap_{\substack{j \in \phi(S) \\ H_j \in \mathcal{A}_2}} H_j$$

for any nonempty subset  $S \subseteq \{1, 2, \dots, n\}$ . Two arrangements are *lattice isomorphic* if their lattices are isomorphic. In this paper, we only consider line arrangements in  $\mathbb{C}\mathbb{P}^2$ .

An essential topic in hyperplane arrangements theory is to study the interaction between topology of complements and combinatorics of intersection lattices. Naturally enough, one may ask how close topology and combinatorics of a given arrangement

---

This work was partially supported by the Oswald Veblen Fund and by the Minerva Foundation of Germany.

*MSC2010:* 14N20, 32S22, 52C35.

*Keywords:* line arrangements, moduli spaces.

are related. Two arrangements,  $\mathcal{A}_1$  and  $\mathcal{A}_2$ , are *homeomorphic equivalent* if there is a homeomorphism between their complements. A more concrete question is: how close are lattice isomorphism and homeomorphic equivalence to being in one-to-one correspondence?

The deepest theorem in the theory of line arrangement in projective 2-dimensional space is that of Jiang and Yau [1998], which asserts that the intersection lattice of the line arrangement is a topological invariant. It is natural to ask to what extent the converse of the Jiang–Yau theorem is true. Jiang and Yau [1994], and subsequently Wang and Yau [2005], have shown that the converse statement is also true for a large class of line arrangements. Therefore, the Jiang–Yau theorem initiates a new research direction: Can one find a Zariski pair of line arrangements; that is, a pair of arrangements which are lattice isomorphic but not homeomorphic equivalent.

A pair of arrangements which are lattice isomorphisms but not homeomorphic equivalent is called a *Zariski pair*. Our definition is stronger than the definition introduced by Artal [1994], which we shall call *weak Zariski pairs* (see [Artal et al. 2008] for a survey on Zariski pairs). The first Zariski pair of arrangements was constructed by Rybnikov [2011]. Each arrangement in Rybnikov’s example consists of 13 lines and 15 triple points. Artal et al. [2005] provided another (weak) Zariski pair of two arrangements  $\mathcal{H}^+ := \mathcal{C}^+ \cup \{N^+\}$  and  $\mathcal{H}^- := \mathcal{C}^- \cup \{N^-\}$ , where  $\mathcal{C}^+$ ,  $\mathcal{C}^-$  are arrangements (Figure 7) extending Falk–Sturmfels arrangements (Figure 2), and  $N^+$ ,  $N^-$  are lines passing through a triple point and a double point of  $\mathcal{C}^\pm$ . The proof is based on the observation that there is no order-preserving homeomorphism between  $(\mathbb{P}^2, \mathcal{C}^+)$  and  $(\mathbb{P}^2, \mathcal{C}^-)$ . In the contrary direction, Garber, Teicher, and Vishne [Garber et al. 2003] proved that there is no Zariski pair of arrangements of up to 8 real lines which covered the result of Fan [1997] on arrangements of 6 lines. This result was recently generalized to arrangements of 8 complex lines by Nazir and Yoshinaga [2012].

A natural question is: what is the minimal number of lines of a Zariski pair of line arrangements?

On the other hand, it was Jiang and Yau [1994] who first observed that the statement “two lattice isotopy line arrangements (that is, they are connected by a one-parameter family with constant intersection lattice) have diffeomorphic complements” follows trivially from Teissier’s numerical characterization of the Whitney condition. In [Jiang and Yau 1994] and [Wang and Yau 2005], the authors found large classes of line arrangements, called *nice arrangements* and *simple arrangements*, whose intersection lattices determine the topology of the complements. Nazir and Yoshinaga [2012] found new classes of line arrangements whose intersection lattices determine the topology of the complements. Unlike nice and simple arrangements whose intersection lattices have special properties, Nazir and Yoshinaga’s new classes require that all intersection points with multiplicity at

least 3 be in special positions. This makes their results more useful for studying arrangements of a few lines. Indeed, in their paper they classify arrangements of 8 lines and give a list of classes of arrangements of 9 lines.

In this paper, we introduce new ideas to classify arrangements of lines. We prove that Nazir and Yoshinaga’s list on the classification of arrangements of 9 lines is complete. As a corollary, we conclude that there is no Zariski pair of arrangements of 9 lines. The idea and results of this paper have been used to classify moduli spaces of arrangements of 10 projective lines (see [Amram et al. 2012]).

The paper is organized as follows: In Section 2, we recall results in Nazir and Yoshinaga. In Section 3, we prove that their list of classes of arrangements of 9 lines is complete. In Section 4, we consider the example of arrangements of 10 lines  $\mathcal{C}^\pm$  and give an explicit diffeomorphism between the complements  $M(\mathcal{C}^\pm)$ .

### 2. Simple $C_{\leq 3}$ line arrangements

Consider the dual space  $(\mathbb{C}\mathbb{P}^2)^*$  of the projective space  $\mathbb{C}\mathbb{P}^2$ . A line arrangement  $\mathcal{A} = \{L_1, L_2, \dots, L_n\}$  can be viewed as an  $n$ -tuple of points  $(L_1^*, L_2^*, \dots, L_n^*)$  in the product of the dual spaces  $((\mathbb{C}\mathbb{P}^2)^*)^n$ . We define the moduli space of arrangements with the fixed lattice  $L(\mathcal{A})$  as

$$\mathcal{M}_{\mathcal{A}} = \frac{\{\mathcal{B} \in ((\mathbb{C}\mathbb{P}^2)^*)^n \mid L(\mathcal{B}) \sim L(\mathcal{A})\}}{\text{PGL}_3(\mathbb{C})} \subseteq \frac{((\mathbb{C}\mathbb{P}^2)^*)^n}{\text{PGL}_3(\mathbb{C})}.$$

We say that a singular point  $P$  of  $L_1 \cup L_2 \cup \dots \cup L_n$  is a *multiple point* of  $\mathcal{A}$  if the multiplicity of  $P$  is at least 3.

The following definition is a combination of Nazir and Yoshinaga’s original definitions of  $C_1$ ,  $C_2$ , and simple  $C_3$  arrangements.

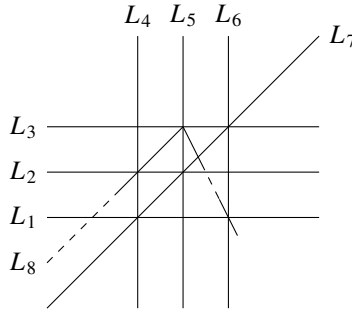
**Definition 2.1.** A line arrangement is called  $C_{\leq 3}$  if all the multiple points are on at most three lines; say,  $L_1, L_2$ , and  $L_3$ . A line arrangement is called simple  $C_{\leq 3}$  if it is  $C_{\leq 3}$ , and one of the following condition holds:

- (i)  $L_1 \cap L_2 \cap L_3 \neq \emptyset$ , or
- (ii) one of  $L_1, L_2$  and  $L_3$  contains at most one more multiple point apart from the possible multiple points  $L_1 \cap L_2, L_2 \cap L_3$ , and  $L_1 \cap L_3$ .

Here are some examples of arrangements which are not simple  $C_{\leq 3}$ :

**Example 2.2.** A *Mac Lane arrangement* (see Figure 1) consists of eight lines and eight triple points such that each line passes through exactly three triple points. It is not hard to check that the moduli space of Mac Lane arrangements consists of two points. Representatives of the two points can be defined by the equation

$$xy(x - z)(y - z)(x - y)(x - \varepsilon^\pm z)(y - \varepsilon^\pm z)(-\varepsilon^\mp x - y + z) = 0,$$



**Figure 1.** A Mac Lane arrangement.

where  $\varepsilon^\pm = \frac{1}{2}(1 \pm \sqrt{-3})$  are the roots of  $x^2 - x + 1 = 0$ .

Since each line passes through three triple points, there are at most seven triple points on three lines. Thus, Mac Lane arrangements cannot be simple  $C_{\leq 3}$ .

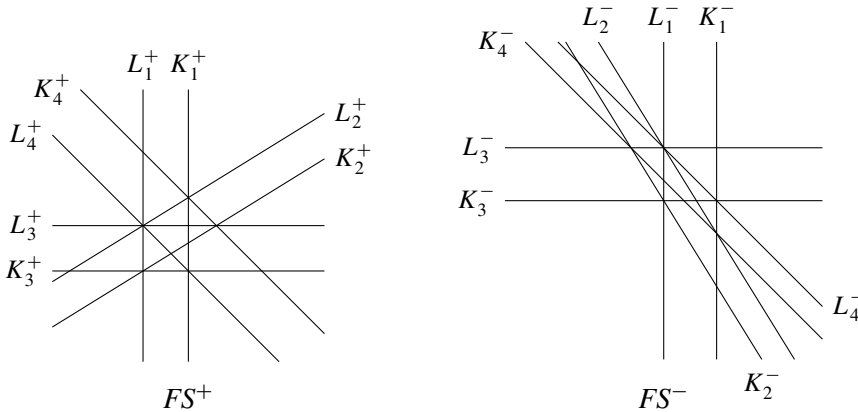
**Example 2.3.** *Falk–Sturmfels arrangements* are the arrangements of nine lines with one quadruple point, eight triple points, and one line passing through four triple points (Figure 2). We denote them by

$$FS^\pm = \{L_i^\pm, K_i^\pm, H_9^\pm, i = 1, 2, 3, 4\},$$

where the lines are defined by

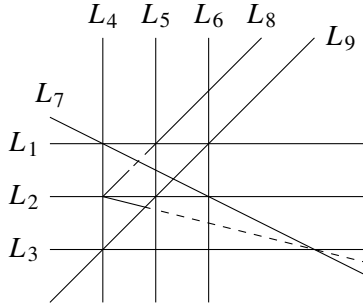
$$\begin{aligned} L_1^\pm : x = 0, \quad L_2^\pm : x = \gamma_\pm(y - z), \quad L_3^\pm : y = z, \quad L_4^\pm : x + y = z, \\ K_1^\pm : x = z, \quad K_2^\pm : x = \gamma_\pm y, \quad K_3^\pm : y = 0, \quad K_4^\pm : x + y = (\gamma_\pm + 1)z, \\ H_9^\pm : z = 0, \end{aligned}$$

with  $\gamma_\pm = \frac{1}{2}(1 \pm \sqrt{5})$  the roots of  $x^2 - x - 1 = 0$ . It is known (see [Nazir and



**Figure 2.** Falk–Sturmfels arrangements.





**Figure 3.** The arrangements  $\mathcal{A}^{\pm\sqrt{-1}}$ .

Yoshinaga 2012, Example 5.2], for instance) that the moduli space  $\mathcal{M}_{L(FS^\pm)}$  consists of 2 points,  $\{FS^+, FS^-\}$ .

**Example 2.4** [Nazir and Yoshinaga 2012, Example 5.3]. The arrangements  $\mathcal{A}^{\pm\sqrt{-1}}$  consist of nine lines and ten triple points such that there are three lines which do not intersect at a point and have four triple points on each. Moreover, each of the other six lines contains exactly three triple points. Those arrangements (see Figure 3) can be defined by the equation

$$xy(x - z)(y - z)(x \mp iz)(y \mp iz)(x - y)((\pm i - 1)x \pm iy + z)((1 \mp i)x + y - z) = 0.$$

**Lemma 2.5** [Nazir and Yoshinaga 2012, Lemma 4.4]. *If a line arrangement is not simple  $C_{\leq 3}$ , then it has 6 lines  $L_1, L_2, \dots, L_6$  such that  $L_1 \cap L_2 \cap L_3 \neq \emptyset$ ,  $L_4 \cap L_5 \cap L_6 \neq \emptyset$ , and  $(L_1 \cup L_2 \cup L_3) \cap (L_4 \cup L_5 \cup L_6)$  consists of 9 distinct double points.*

Let  $\mathcal{A}_s = \{L_1, L_2, \dots, L_6\}$  be the arrangement which has two triple points  $L_1 \cap L_2 \cap L_3$  and  $L_4 \cap L_5 \cap L_6$ , and nine double points  $Q_{ij} = L_i \cap L_{j+3}$ , where  $i, j \in \{1, 2, 3\}$ .

Using Lemma 2.5, one can easily prove that an arrangement of 7 lines is simple  $C_{\leq 3}$ . It is also not hard to prove the following result:

**Proposition 2.6** [Nazir and Yoshinaga 2012, Proposition 4.6]. *An arrangement of eight lines is either a simple  $C_{\leq 3}$  line arrangement or a Mac Lane arrangement.*

More generally:

**Theorem 2.7** [Nazir and Yoshinaga 2012, Theorem 3.5]. *The moduli space  $\mathcal{M}_{\mathcal{A}}$  of simple  $C_{\leq 3}$  line arrangements with the fixed intersection lattice  $L(\mathcal{A})$  is irreducible.*

Let  $\mathcal{A} = \{L_1, L_2, \dots, L_n\}$  be a line arrangement, and  $\mathcal{A}' = \{L_1, L_2, \dots, L_{n-1}\}$  be a subarrangement. The following lemma shows when the irreducibility of the moduli space  $\mathcal{M}_{\mathcal{A}'}$  will be inherited:

**Lemma 2.8** [Nazir and Yoshinaga 2012, Lemma 2.4]. *Assume that the line  $L_n$  passes through at most two multiple points of the arrangement  $\mathcal{A}$ . Then the moduli space  $\mathcal{M}_{\mathcal{A}}$  is a fiber bundle over the moduli space of  $\mathcal{M}_{\mathcal{A}'}$ . In particular, the moduli space  $\mathcal{M}_{\mathcal{A}}$  is irreducible if  $\mathcal{M}_{\mathcal{A}'}$  is irreducible.*

Applying this lemma to arrangements of 9 lines, we have the following corollary:

**Corollary 2.9.** *Let  $\mathcal{A}$  be an arrangement of 9 lines. If there is a line in  $\mathcal{A}$  which passes through at most two multiple points of  $\mathcal{A}$ , then either  $\mathcal{A}$  contains a Mac Lane arrangement as a subarrangement, or the moduli space  $\mathcal{M}_{\mathcal{A}}$  is irreducible.*

*Proof.* The conclusion follows directly from Proposition 2.6 and Lemma 2.8.  $\square$

### 3. Classification of arrangements of 9 lines

For a line arrangement  $\mathcal{A}$ , we denote by  $m_{\mathcal{A}}$  the highest multiplicity of a multiple point of  $\mathcal{A}$ . We will divide the classification of arrangements of 9 lines into three cases according to the value of  $m_{\mathcal{A}}$ .

Let  $n_r$  be the number of multiple points of multiplicity  $r$ . We first recall two well-known results on the number of multiple points.

**Theorem 3.1** [Hirzebruch 1986]. *Let  $\mathcal{A}$  be an arrangement of  $t$  lines in  $\mathbb{C}\mathbb{P}^2$ . Assume that  $n_t = n_{t-1} = n_{t-2} = 0$ . Then,*

$$n_2 + \frac{3}{4}n_3 \geq t + \sum_{r \geq 5} (2r - 9)n_r.$$

**Lemma 3.2** (see, for instance, [Hirzebruch 1986]). *Let  $\mathcal{A}$  be a line arrangement of  $n$  lines in  $\mathbb{C}\mathbb{P}^2$ . We have the intersection formula*

$$\frac{n(n-1)}{2} = \sum_{r \geq 2} \left( n_r \cdot \frac{r(r-1)}{2} \right).$$

#### 3A. The case $m_{\mathcal{A}} \geq 5$ .

**Proposition 3.3.** *Let  $\mathcal{A}$  be an arrangement of 9 lines. If  $\mathcal{A}$  has multiple points of multiplicity (at least 5), then the moduli space  $\mathcal{M}_{\mathcal{A}}$  is irreducible.*

*Proof.* Assume that  $L_1 \cap L_2 \cap \cdots \cap L_5 \neq \emptyset$ . There are at most 6 double points in  $L_6 \cup L_7 \cup L_8 \cup L_9$ . Then, there are at most 7 multiple points in  $L_1 \cup L_2 \cup \cdots \cup L_5$ . So, at least one of the five lines  $L_1, L_2, \dots, L_5$  contains only two multiple points. By Corollary 2.9, the moduli space  $\mathcal{M}_{\mathcal{A}}$  is irreducible.  $\square$

**3B. The case  $m_{\mathcal{A}} = 4$ .** Let  $\mathcal{A}$  be an arrangement of 9 lines. In this subsection, we assume that multiple points of  $\mathcal{A}$  are at most quadruple points.

**Proposition 3.4.** *Assume that each line of  $\mathcal{A}$  passes through at least three multiple points, and  $n_4 \geq 1$ . Then either  $\mathcal{M}_{\mathcal{A}}$  is irreducible, or  $\mathcal{A}$  is lattice isomorphic to a Falk–Sturmfels arrangement.*

*Proof.* We will first show that  $n_4 = 1$ .

Let  $L_1 \cap L_2 \cap L_3 \cap L_4$  be a quadruple point of  $\mathcal{A}$ . Since each line passes through at least three multiple points,  $L_1, L_2, L_3$  and  $L_4$  should pass through two more multiple points besides the quadruple point  $L_1 \cap L_2 \cap L_3 \cap L_4$ . Then, there will be at least 9 multiple points on those four lines. Since multiple points of  $\mathcal{A}$  are at most quadruple points, there are  $n_4$  quadruple points. Therefore, there should be at least  $9 - n_4$  triple points on those four lines such that each line passes through at least 3 multiple points. By Theorem 3.1 and Lemma 3.2, we have

$$36 = 6n_4 + 3n_3 + n_2 \geq 6n_4 + \frac{9}{4}n_3 + 9 \geq 6n_4 + \frac{9}{4}(9 - n_4) + 9.$$

Solving the inequality, we obtain that  $n_4 \leq \frac{9}{5} < 2$ . Therefore, by the assumption, we have  $n_4 = 1$ .

Now we claim that all triple points should be on the lines passing through the quadruple point.

Let  $L_1 \cap L_2 \cap L_3 \cap L_4$  be the quadruple. Suppose, contrary to our claim, that  $L_5 \cap L_6 \cap L_7$  is a triple point which is not on  $L_1 \cup L_2 \cup L_3 \cup L_4$ . Note that there are at most 7 double points on  $L_5 \cup L_6 \cup L_7 \cup L_8 \cup L_9$ . Then the intersection set  $(L_1 \cup L_2 \cup L_3 \cup L_4) \cap (L_5 \cup L_6 \cup L_7 \cup L_8 \cup L_9)$  will contain at most 7 triple points which are on  $L_1 \cup L_2 \cup L_3 \cup L_4$ . However, there should be at least 8 triple points so that each of the four lines  $L_1, L_2, L_3$  and  $L_4$  passes through at least three multiple points. Therefore, by the assumption, all triple points must be on the lines passing through the quadruple point.

If  $\mathcal{A}$  is simple  $C_{\leq 3}$ , then the moduli space  $\mathcal{M}_{\mathcal{A}}$  is irreducible. We only need to consider the case that  $\mathcal{A}$  is not simple  $C_{\leq 3}$ . By Lemma 2.5, we know that the arrangement  $\mathcal{A}$  has a subarrangement  $\mathcal{A}_s$ . It is not hard to see that the quadruple point should be  $Q_{ij}$ , where  $i, j \in \{1, 2, 3\}$ .

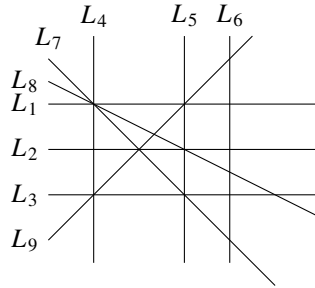
Up to a lattice isomorphism, we may assume that the only quadruple point is  $L_1 \cap L_4 \cap L_7 \cap L_8 = Q_{11}$ .

Since all triple points should be on  $L_1 \cup L_4 \cup L_7 \cup L_8$ , then all possible triple points on  $L_7$  and  $L_8$  should be in the following set of points:

$$\{Q_{22}, Q_{23}, Q_{32}, Q_{33}, L_7 \cap L_9, L_8 \cap L_9\}.$$

The following figure is an example, but an excluding one, for  $L_6$  passes through only one triple point.

Hence, each of the lines  $L_7$  and  $L_8$  will have at least one  $Q_{ij}$ , where  $i, j \in \{2, 3\}$ .



**Figure 4.** An excluding arrangement.

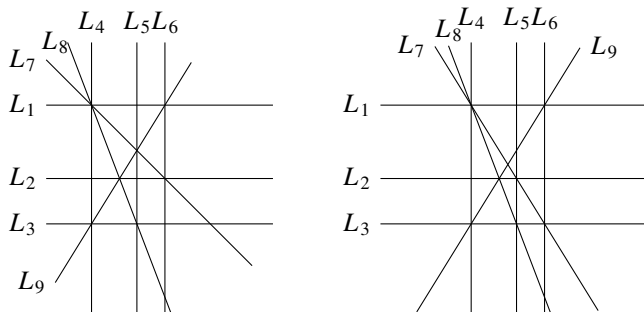
(1) Assume that each of the lines  $L_7$  and  $L_8$  passes through exactly one of the points  $\{Q_{22}, Q_{23}, Q_{32}, Q_{33}\}$ .

If those two  $Q_{ij}$  are on the same line, then one of the four lines  $L_2, L_3, L_5$  and  $L_6$  will have at most two multiple points. For example, in Figure 4, the line  $L_6$  passes through only one multiple point,  $L_4 \cap L_5 \cap L_6$ .

Assume that they are not on the same line. Up to switching labels between  $L_2$  and  $L_3$ , correspondingly  $L_5$  and  $L_6$ , we may assume that  $Q_{32} \in L_7$  and  $Q_{23} \in L_8$ . Then, either  $\{Q_{31}, Q_{13}\} \subset L_9$  or  $\{Q_{21}, Q_{12}\} \subset L_9$ . Correspondingly,  $\{L_2 \cap L_7, L_5 \cap L_8\} \subset L_9$  or  $\{L_3 \cap L_7, L_6 \cap L_8\} \subset L_9$ . By switching the labels between  $L_2$  and  $L_3, L_5$  and  $L_6$ , and  $L_7$  and  $L_8$ , we see that those two arrangements are lattice isomorphic. Moreover, one can check that both arrangements (see Figure 5, left) are lattice isomorphic to Falk–Sturmfels arrangements.

(2) Assume that either the line  $L_7$  or  $L_8$  passes through two points out of the four points  $Q_{22}, Q_{23}, Q_{32}$  and  $Q_{33}$ , but the other one passes through only one point out of the four points  $Q_{22}, Q_{23}, Q_{32}$  and  $Q_{33}$ .

Up to a lattice isomorphism, we may assume that  $\{Q_{11}, Q_{22}, Q_{33}\} \subset L_7$ , and  $\{Q_{11}, Q_{32}\} \subset L_8$ . Then either  $L_2 \cap L_8 \in L_9$ , or  $L_6 \cap L_8 \in L_9$ . Otherwise,  $L_8$  will have only two multiple points. Correspondingly,  $\{Q_{31}, Q_{13}\} \subset L_9$ , or  $\{Q_{21}, Q_{12}\} \subset L_9$ .



**Figure 5.** Falk–Sturmfels arrangements 1 and 2.

By first switching the labels between  $L_1$  and  $L_4$ ,  $L_2$  and  $L_5$ , and  $L_3$  and  $L_6$ , then switching the labels between  $L_2$  and  $L_3$ , and  $L_5$  and  $L_6$ , we see that those two arrangements are lattice isomorphic. Moreover, we check that  $\mathcal{A}$  (see Figure 5, right) is also lattice isomorphic to Falk–Sturmfels arrangements.

(3) Assume that  $L_7$  and  $L_8$  each contain two of  $\{Q_{22}, Q_{23}, Q_{32}, Q_{33}\}$ , then  $L_9$  will contain at most two multiple points.

Therefore, we conclude that either  $\mathcal{M}_{\mathcal{A}}$  is irreducible or  $\mathcal{A}$  is lattice isomorphic to a Falk–Sturmfels arrangement.  $\square$

**3C. The case  $m_{\mathcal{A}} = 3$ .** Now we consider the last case in which all multiple points are triple points. We will first investigate possible values of  $n_3$  such that each line has at least three triple points. Notice that  $n_3$  should be no less than 9. On the other hand, we observe the following result:

**Lemma 3.5.** *Let  $\mathcal{A}$  be an arrangement of 9 lines, all of whose multiple points are triple points. Assume that  $\mathcal{A}$  does not contain a Mac Lane arrangement as a subarrangement and is not simple  $C_{\leq 3}$ . Then  $\mathcal{A}$  has at most 10 triple points.*

*Proof.* By Lemma 3.2, to show that  $n_3 \leq 10$ , it is enough to show that  $n_2 \geq 4$ .

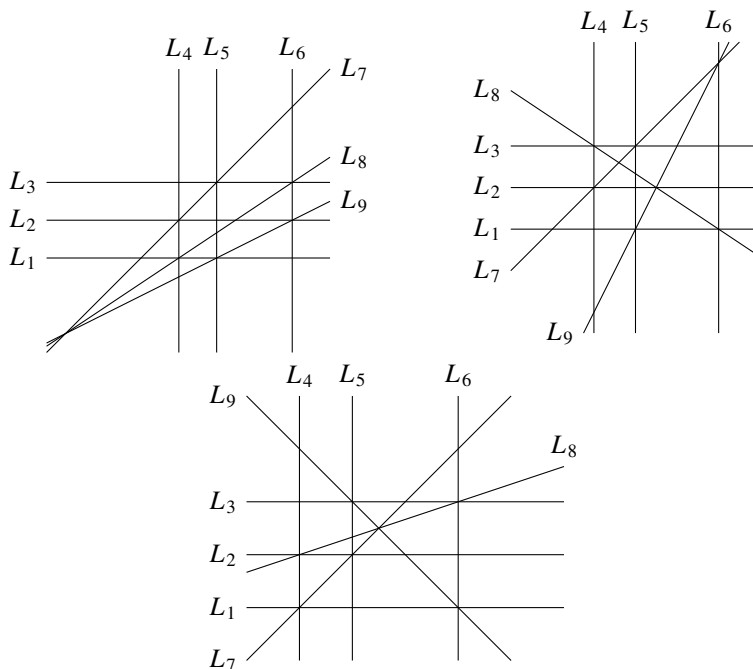
Since  $\mathcal{A}$  does not contain a Mac Lane arrangement, at most one of the lines  $L_7$ ,  $L_8$ , and  $L_9$  passes through three  $Q_{ij}$ , where  $i, j \in \{1, 2, 3\}$  (defined as above). We may assume that each of the lines  $L_7$  and  $L_8$  passes through at most two  $Q_{ij}$ . By our assumption and Lemma 2.5, the arrangement  $\mathcal{A}$  has a subarrangement  $\mathcal{A}_s$ .

Let  $x$  be the number of  $Q_{ij}$  which are not in  $L_7 \cup L_8 \cup L_9$ . It is clear that  $x \geq 2$ . Let  $y$  and  $z$  be the number of double points of  $\mathcal{A}$  which are in  $L_7 \cap (L_1 \cup L_2 \cup \dots \cup L_6)$  and  $L_8 \cap (L_1 \cup L_2 \cup \dots \cup L_6)$  respectively. If  $y + z \geq 2$ , then we have  $n_2 \geq x + (y + z) \geq 4$ .

Assume that  $y + z \leq 1$ . Then each of the lines  $L_7$  and  $L_8$  should pass through exactly two  $Q_{ij}$ . Moreover,  $L_7 \cap L_8$  must be a triple point in  $L_1 \cup L_2 \cup \dots \cup L_6$ . We see now the subarrangement  $\mathcal{A}' = \{L_1, L_2, \dots, L_8\}$  has 7 double points. Without a loss of generality, we assume that  $L_7 \cap L_8$  is on  $L_2$ . It is not hard to see that the 7 double points of  $\mathcal{A}'$  are all on  $L_4 \cup L_5 \cup L_6$ . The line  $L_9$  can only pass through at most three double points of  $\mathcal{A}'$ . Therefore, the arrangement  $\mathcal{A}$  still has at least 4 double points.  $\square$

**Remark 3.6.** By Theorem 2.15 in [Csima and Sawyer 1993], if our arrangements are *real arrangements*, that is, if the coefficients of the defining equations of the lines are real numbers, then there are at least  $\frac{60}{13} > 4$  double points. Hence, there should be at most 10 triple points. However, there seems to be no similar result for complex line arrangements.

**Proposition 3.7.** *Let  $\mathcal{A}$  be an arrangement of 9 lines with 9 triple points. Assume that all multiple points of  $\mathcal{A}$  are triple points, and each line passes through exactly three triple points. Then, the moduli space  $\mathcal{M}_{\mathcal{A}}$  is irreducible.*



**Figure 6.**  $9_3$  arrangements.

*Proof.* By Theorem 2.2.1 in [Grünbaum 2009] that  $\mathcal{A}$  is lattice isomorphic to one of the three arrangements appearing in Figure 6.

One can check that the moduli space  $\mathcal{M}_{\mathcal{A}}$  is irreducible in each case. □

**Proposition 3.8.** *Let  $\mathcal{A}$  be an arrangement of 9 lines with 10 triple points. Assume that all multiple points of  $\mathcal{A}$  are triple points and each line passes through at least three triple points. If  $\mathcal{A}$  is not simple  $C_{\leq 3}$ , then it is isomorphic to  $\mathcal{A}^{\pm\sqrt{-1}}$  (Figure 3).*

*Proof.* Let  $a$  be the number of lines that pass through 4 triple points and  $b$  the number of lines that pass through 3 triple points. Then  $a + b = 9$  and  $4a + 3b = 30$ . We have  $a = 3$  and  $b = 6$ .

If the three lines with 4 triple points on each of them intersect at a triple point, then all 10 triples should be on them. Consequently, the arrangement is simple  $C_{\leq 3}$ .

Assume that  $L_1, L_2$  and  $L_4$  are the three lines with 4 triple points on each of them and  $L_1 \cap L_2 \cap L_4 = \emptyset$ . Then, at least two of  $\{L_1 \cap L_2, L_1 \cap L_4, L_2 \cap L_4\}$  are triple points. Otherwise there should be at least 11 triple points so that  $L_1, L_2,$  and  $L_4$  will have 4 triple points. So, we may assume that  $L_1 \cap L_2 \cap L_3$  and  $L_1 \cap L_4 \cap L_7$  are triple points. Let  $L_4 \cap L_5 \cap L_6$  be a triple point which is not on  $L_1 \cup L_2 \cup L_3$ . Then,  $L_7$  must pass through  $L_2 \cap L_5$  or  $L_2 \cap L_6$ . Otherwise,  $L_2$  will have at most 3 triples. By switching labels of  $L_5$  and  $L_6$ , we may assume that

$L_2 \cap L_6 \cap L_7 \neq \emptyset$ . Then the two points  $Q_{21}$  and  $Q_{22}$  must be on  $L_8 \cup L_9$  so that  $L_2$  will pass through 4 triple points. We may assume that  $Q_{21} \in L_8$  and  $Q_{22} \in L_9$ . Since the line  $L_4$  also passes through 4 triple points, then  $Q_{31}$  should be on  $L_9$ . Similarly, since the line  $L_1$  passes through 4 triple points, then  $Q_{13}$  should be on  $L_9$  and  $Q_{12}$  should be on  $L_8$ . Now we have 9 triple points. The last triple point must be  $L_3 \cap L_7 \cap L_8$  so that  $L_7$  will pass through three triple points. The arrangements with such intersection lattices are just  $\mathcal{A}^{\pm\sqrt{-1}}$  (see Figure 3).  $\square$

**3D. Classification and applications.** We summarize Section 3 so far as follows:

**Theorem 3.9.** *Any arrangement of nine lines in  $\mathbb{C}\mathbb{P}^2$  belongs to one of the following classes:*

- (i) *arrangements whose moduli spaces are irreducible;*
- (ii) *arrangements containing Mac Lane arrangements (Example 2.2);*
- (iii) *Falk–Sturmfels arrangements (Example 2.3);*
- (iv)  *$\mathcal{A}^{\pm\sqrt{-1}}$  arrangements (Example 2.4).*

*Proof.* The classification simply follows from Corollary 2.9 and Propositions 3.3, 3.4, 3.7, and 3.8.  $\square$

As an application, we obtain the following result which generalizes a result of Theorem 8.3 in [Garber et al. 2003].

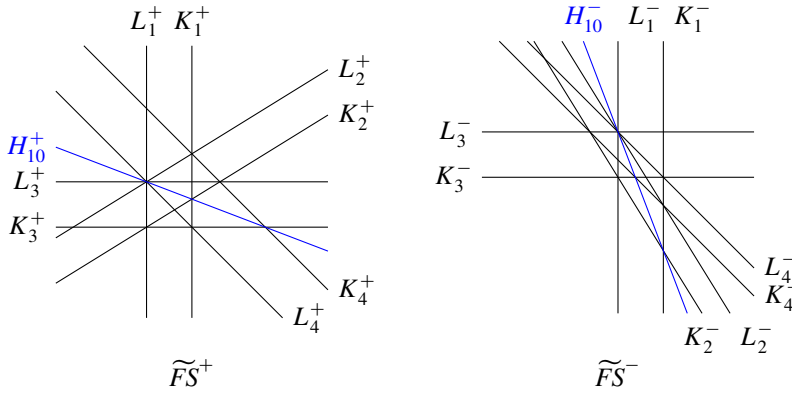
**Theorem 3.10.** *The fundamental group of the complement of an arrangement of 9 lines is determined by the intersection lattice.*

*Proof.* If the moduli space is irreducible, then the fundamental group is determined by the lattice according to the lattice-isotopy theorem.

It follows from Example 5.2 in [Nazir and Yoshinaga 2012] (see also Section 7.5 in [Cohen and Suciu 1997]) that the fundamental groups  $\pi_1(M(FS^+))$  and  $\pi_1(M(FS^-))$  are isomorphic. Let  $\mathcal{A}_1$  and  $\mathcal{A}_2$  be two arrangements containing Mac Lane arrangements. Then, either they are in the same connected component of the moduli spaces, or  $\mathcal{A}_1$  and the conjugate of  $\mathcal{A}_2$  are in the same connected component. By Theorem 3.9 in [Cohen and Suciu 1997], the fundamental groups of  $\mathcal{A}_1$  and  $\mathcal{A}_2$  are isomorphisms. According to the same theorem, the fundamental groups of  $\mathcal{A}^{+\sqrt{-1}}$  and  $\mathcal{A}^{-\sqrt{-1}}$  are isomorphic too.  $\square$

#### 4. Arrangements of 10 lines: an example

We have seen that there is no Zariski pair of arrangements of 9 lines, but we do not know if there is a Zariski pair of arrangements of 10 lines. To get a Zariski pair, a naive idea is to add lines to those arrangements whose moduli spaces are disconnected. In general, it is very hard to determine if the resulting pair of arrangements is a Zariski pair. The following example is a trial:



**Figure 7.** Extended Falk–Sturmfels arrangement.

**Example 4.1.** Starting from the Falk–Sturmfels arrangements (see Example 2.3), we will construct new arrangements of 10 lines such that the moduli space is disconnected.

We define two line arrangements of 10 lines, called extended Falk–Sturmfels arrangements (see Figure 7):

$$\widetilde{FS}^\pm = \{L_i^\pm, K_i^\pm, H_9^\pm, H_{10}^\pm, i = 1, 2, 3, 4\}$$

by adding lines:

$$H_{10}^\pm : y = \left(\frac{1}{\gamma_\pm} - 1\right)x + z$$

to  $FS^\pm$  respectively.

Notice that  $\widetilde{FS}^\pm$  are both fiber-type line arrangements according to Theorem 3.12 in [Jiang et al. 2001] .

It is not hard to see that  $\mathcal{M}_{\widetilde{FS}^\pm} \cong \mathcal{M}_{FS^\pm}$ . In fact, the line  $H_{10}^+$  (respectively,  $H_{10}^-$ ) is always passing through three points of  $L(FS^\pm)$ :  $L_1^+ \cap L_2^+, K_1^+ \cap K_2^+$  and  $K_3^+ \cap K_4^+$  (respectively,  $K_2^- \cap K_4^-, K_3^- \cap K_4^-$  and  $K_1^- \cap K_2^-$ ).

This pair of arrangements has been studied by Artal, Carmona, Cogolludo, and Marco. They show (Theorem 4.19 in [Artal et al. 2005]) that there is no order-preserving homeomorphism between the pairs  $(\mathbb{P}^2, \widetilde{FS}^+)$  and  $(\mathbb{P}^2, \widetilde{FS}^-)$ . Here, we present an explicit diffeomorphism between the complements  $M(\widetilde{FS}^+)$  and  $M(\widetilde{FS}^-)$ . In fact, by Example 5.2 in [Nazir and Yoshinaga 2012], we know that there is an automorphism  $A \in \text{PGL}(\mathbb{C}^3)$  of  $\mathbb{C}\mathbb{P}^2$ ,

$$A := \begin{pmatrix} -\gamma_- & -1 & 0 \\ -\gamma_- & 0 & 0 \\ \gamma_- & 1 & 1 \end{pmatrix},$$



acting from the right (via matrix multiplication) on points  $[x, y, z]$  in the projective space  $\mathbb{P}^2$ , which sends

$$\begin{aligned} L_1^+ &\mapsto L_3^-, & L_2^+ &\mapsto L_4^-, & L_3^+ &\mapsto L_2^-, & L_4^+ &\mapsto L_1^-, \\ K_1^+ &\mapsto K_3^-, & K_2^+ &\mapsto K_4^-, & K_3^+ &\mapsto K_2^-, & K_4^+ &\mapsto K_1^-, \\ H_9^+ &\mapsto H_9^-. \end{aligned}$$

To see that  $A$  induces a diffeomorphism between  $M(\widetilde{FS}^+)$  and  $M(\widetilde{FS}^-)$ , it suffices to show that the automorphism  $A$  sends  $H_{10}^+$  to  $H_{10}^-$ .

Recall that  $\gamma_{\pm} = \frac{1}{2}(1 \pm \sqrt{5})$ . One can check that for any point

$$P := [x, (1/\gamma_+ - 1)x + z, z]$$

on  $H_{10}^+$ , the image  $P \cdot A$  is a point on  $H_{10}^-$ . In fact,

$$\begin{pmatrix} x & (1/\gamma_+ - 1)x + z & z \end{pmatrix} \cdot A \cdot \begin{pmatrix} 1/\gamma_- - 1 \\ -1 \\ 1 \end{pmatrix} \equiv 0.$$

Therefore, the pair  $(\widetilde{FS}^+, \widetilde{FS}^-)$  is not a Zariski pair.

From this example, we see that moduli spaces of fiber-type projective line arrangements do not have to be connected. In fact, we can produce infinitely many fiber-type projective line arrangements whose moduli spaces are disconnected. On the other hand, we do not know if fundamental groups of complements of fiber-type projective line arrangements are determined by intersection lattices.

### Acknowledgements

The author is grateful to M. Amram, M. Cohen, M. Eliyahu, D. Garber, M. Teicher, E. Artal Bartolo, and J.I. Cogolludo Agustín for their helpful conversations and comments, and especially to D. Garber for comments on a draft of this manuscript.

### References

[Amram et al. 2012] M. Amram, M. Teicher, and F. Ye, “Moduli spaces of arrangements of 10 projective lines with quadruple points”, preprint, 2012. To appear in *Adv. Appl. Math.* arXiv 1206.2486

[Artal 1994] E. Artal Bartolo, “Sur les couples de Zariski”, *J. Algebraic Geom.* **3:2** (1994), 223–247. MR 94m:14033 Zbl 0823.14013

[Artal et al. 2005] E. Artal Bartolo, J. Carmona Ruber, J. I. Cogolludo-Agustín, and M. Marco Buzunáriz, “Topology and combinatorics of real line arrangements”, *Compos. Math.* **141:6** (2005), 1578–1588. MR 2006k:32055 Zbl 1085.32012

[Artal et al. 2008] E. Artal Bartolo, J. I. Cogolludo, and H.-o. Tokunaga, “A survey on Zariski pairs”, pp. 1–100 in *Algebraic geometry in East Asia* (Hanoi, 2005), edited by K. Konno and V. Nguyen-Khac, Adv. Stud. Pure Math. **50**, Math. Soc. Japan, Tokyo, 2008. MR 2009g:14030 Zbl 1141.14015

- [Cohen and Suciu 1997] D. C. Cohen and A. I. Suciu, “The braid monodromy of plane algebraic curves and hyperplane arrangements”, *Comment. Math. Helv.* **72**:2 (1997), 285–315. MR 98f:52012 Zbl 0959.52018
- [Csima and Sawyer 1993] J. Csima and E. T. Sawyer, “There exist  $6n/13$  ordinary points”, *Discrete Comput. Geom.* **9**:2 (1993), 187–202. MR 94a:52015 Zbl 0771.52003
- [Fan 1997] K.-M. Fan, “Direct product of free groups as the fundamental group of the complement of a union of lines”, *Michigan Math. J.* **44**:2 (1997), 283–291. MR 98j:14039 Zbl 0911.14007
- [Garber et al. 2003] D. Garber, M. Teicher, and U. Vishne, “ $\pi_1$ -classification of real arrangements with up to eight lines”, *Topology* **42**:1 (2003), 265–289. MR 2004b:32048 Zbl 1074.14050
- [Grünbaum 2009] B. Grünbaum, *Configurations of points and lines*, Graduate Studies in Mathematics **103**, American Mathematical Society, Providence, RI, 2009. MR 2011j:52001 Zbl 1205.51003
- [Hirzebruch 1986] F. Hirzebruch, “Singularities of algebraic surfaces and characteristic numbers”, pp. 141–155 in *The Lefschetz centennial conference, I* (Mexico City, 1984), edited by D. Sundararaman, Contemp. Math. **58**, Amer. Math. Soc., Providence, RI, 1986. MR 87j:14057 Zbl 0601.14030
- [Jiang and Yau 1994] T. Jiang and S. S.-T. Yau, “Diffeomorphic types of the complements of arrangements of hyperplanes”, *Compositio Math.* **92**:2 (1994), 133–155. MR 95e:32042 Zbl 0828.57018
- [Jiang and Yau 1998] T. Jiang and S. S.-T. Yau, “Intersection lattices and topological structures of complements of arrangements in  $\mathbb{C}P^2$ ”, *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **26**:2 (1998), 357–381. MR 99g:52020 Zbl 0973.32015
- [Jiang et al. 2001] T. Jiang, S. S.-T. Yau, and L.-Y. Yeh, “Simple geometric characterization of supersolvable arrangements”, *Rocky Mountain J. Math.* **31**:1 (2001), 303–312. MR 2001m:55047 Zbl 1008.32015
- [Nazir and Yoshinaga 2012] S. Nazir and M. Yoshinaga, “On the connectivity of the realization spaces of line arrangements”, *Ann. Sc. Norm. Super. Pisa, Cl. Sci.* **11**:4 (2012), 921–937. Zbl 06142478
- [Rybnikov 2011] G. L. Rybnikov, “On the fundamental group of the complement of a complex hyperplane arrangement”, *Funktsional. Anal. i Prilozhen.* **45**:2 (2011), 71–85. MR 2012i:14067
- [Wang and Yau 2005] S. Wang and S. S.-T. Yau, “Rigidity of differentiable structure for new class of line arrangements”, *Comm. Anal. Geom.* **13**:5 (2005), 1057–1075. MR 2007d:32021 Zbl 1115.52010

Received July 3, 2012. Revised November 13, 2012.

FEI YE  
 DEPARTMENT OF MATHEMATICS  
 THE UNIVERSITY OF HONG KONG  
 POKFULAM  
 HONG KONG  
 fye@maths.hku.hk  
 http://hkumath.hku.hk/~fye

## Guidelines for Authors

Authors may submit manuscripts at [msp.berkeley.edu/pjm/about/journal/submissions.html](http://msp.berkeley.edu/pjm/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](mailto: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  $\text{Bib}\TeX$  is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to [pacific@math.berkeley.edu](mailto: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 with 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 265 No. 1 September 2013

---

Genus-two Goeritz groups of lens spaces	1
SANGBUM CHO	
A compact embedding theorem for generalized Sobolev spaces	17
SENG-KEE CHUA, SCOTT RODNEY and RICHARD L. WHEEDEN	
Partial integrability of almost complex structures and the existence of solutions for quasilinear Cauchy–Riemann equations	59
CHONG-KYU HAN and JONG-DO PARK	
An overdetermined problem in potential theory	85
DMITRY KHAVINSON, ERIK LUNDBERG and RAZVAN TEODORESCU	
Quasisymmetric homeomorphisms on reducible Carnot groups	113
XIANGDONG XIE	
Capillarity and Archimedes’ principle	123
JOHN MCCUAN and RAY TREINEN	
Generalized eigenvalue problems of nonhomogeneous elliptic operators and their application	151
DUMITRU MOTREANU and MIEKO TANAKA	
Weighted Ricci curvature estimates for Hilbert and Funk geometries	185
SHIN-ICHI OHTA	
On generalized weighted Hilbert matrices	199
EMMANUEL PREISSMANN and OLIVIER LÉVÊQUE	
Unique prime decomposition results for factors coming from wreath product groups	221
J. OWEN SIZEMORE and ADAM WINCHESTER	
On volume growth of gradient steady Ricci solitons	233
GUOFANG WEI and PENG WU	
Classification of moduli spaces of arrangements of nine projective lines	243
FEI YE	



0030-8730(201309)265:1;1-1