Pacific Journal of Mathematics

Volume 275 No. 1 May 2015

PACIFIC JOURNAL OF MATHEMATICS

msp.org/pjm

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

See inside back cover or msp.org/pjm for submission instructions.

The subscription price for 2015 is US $420/year for the electronic version, and $570/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 Web of Knowledge (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/

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.1

CONSTANT-SPEED RAMPS

OSCAR M. PERDOMO

It is easy to show that if the kinetic coefficient of friction between a block and a ramp is μ_k and this ramp is a straight line with slope $-\mu_k$, then this block will move along the ramp with constant speed. A natural question to ask is the following: besides straight lines, are there other shapes of ramps such that a block will go down the ramp with constant speed? Here we classify all possible shapes of these ramps, and, surprisingly, we show that the planar ramps can be parametrized in terms of elementary functions: trigonometric functions, exponential functions and their inverses. They provide basic examples of curves explicitly parametrized by arclength. A video explaining the main results in this paper can be found at http://youtu.be/iBrvbb0efVk.

1. Introduction

It is known experimentally that when a block slides on a surface under the presence of gravity, besides the force exerted on the block by the surface — called the normal force — there is a second force on the block, with direction opposite to the direction of the velocity vector of the block and with length proportional to the length of the normal force. This second force is called the *friction force*. If we use symbols in a bold font to denote vectors and we use the same symbols (not in bold) to represent their magnitudes, then the friction force $\boldsymbol{F}$ satisfies

$$F = \mu N, \tag{1-1}$$

where $\boldsymbol{N}$ is the normal force and μ is a constant called the *kinetic coefficient of friction*.

In this paper we will assume that (1-1) holds true, even though in real life it is just a good approximation and its formal deduction from basic physical laws is not known. Let us quote [Feynman et al. 1963] in this regard: "It is quite difficult to do accurate quantitative experiments in friction, and the laws of friction are still not analyzed very well, in spite of the enormous engineering value of an accurate analysis. Although the law $F = \mu N$ is fairly accurate once the surfaces are standardized, the reason for this form of the law is not really understood."

MSC2010: 53Z05.
Keywords: friction, constant speed, ramps, sliding.

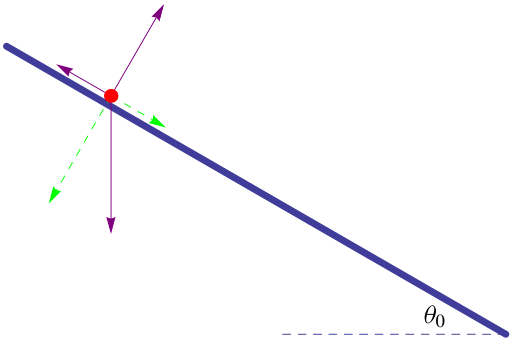


Figure 1. A block sliding down under the effect of gravity and the friction force. This figure shows the three forces acting on the block. The dashed forces are just a decomposition of the gravity force.

We will take (1-1) and the discussion of the direction of the forces in the previous paragraphs as a definition of the friction force and the kinetic coefficient of friction. One of the difficulties dealing with friction forces is that they are not continuous, due to the fact that their direction must be opposite to the velocity vector of the block. For example, friction forces are responsible for making a car stop; they produce a force opposite to the motion of the car, but once the car has stopped, this force disappears.

A basic computation shows that given a kinetic coefficient of friction μ_k, the angle of a linear ramp can be adjusted so that a block on this ramp will move down with constant speed. In order to cancel the gravity, friction, and normal forces, it is enough to take a ramp given by a line with slope $-\mu_k$, that is, we must take the inclination of the ramp to be θ_0 with $\tan\theta_0 = \mu_k$. See Figure 1.

In this paper we study other ramps on which a block can slide down with constant speed under only the effect of gravity and the friction force. Let us mention some of the differences and similarities between the motion on a linear ramp and on these other ramps: (i) As happens for linear ramps, if a block is moving with constant speed on *any* ramp only under the effect of gravity and the friction force, then the velocity vector of the block must have a negative vertical component, that is, the motion must be downward (see Proposition 3.9). (ii) When the ramp is linear, the angle of the ramp (i.e., the shape of the ramp) is independent of the velocity of the block; it depends only on the coefficient of friction. For other ramps, the shape of the ramp depends on the desired constant speed of the block. Corollary 4.2 provides an exact procedure for changing the shape of the ramp according to the desired constant speed of the block.

Let us assume that we want a block to move down with speed v_0 and we know that the kinetic coefficient of friction is μ_k. In this note, we classify the shapes of

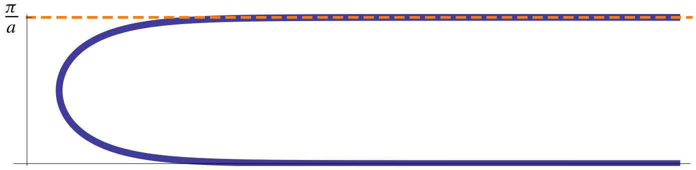


Figure 2. The graph of the curve $\boldsymbol{\alpha}$, which has two horizontal asymptotes, separated by a distance of π/a.

all possible ramps on which this block will move down with speed v_0. We now construct the 2-dimensional *constant-speed ramp*. Given v_0 and $\mu_k = \tan\delta$, define $a = g/(v_0^2 \sin\delta)$, where g is the acceleration due to gravity, and

$$\boldsymbol{\alpha}(s) = \left(s + \frac{1}{a}\ln(1+\mathrm{e}^{-2as}), \frac{2}{a}\operatorname{arccot}(\mathrm{e}^{-as})\right). \tag{1-2}$$

This curve $\boldsymbol{\alpha}$ has a “U” shape with two horizontal asymptotes: one at $y = 0$ and the second one at $y = \pi/a$. See Figure 2.

If we rotate the curve $\boldsymbol{\alpha}$ clockwise through an angle δ (see Figure 3), then the highest point divides this rotated curve into two ramps, on which a block can move down with constant speed v_0 under the assumption that the kinetic coefficient of friction is $\mu_k = \tan\delta$.

The sequence of pictures in the Appendix shows the motion on the ramps when the desired speed is 5 m/s and $\mu_k = 0.5$. The pictures also display the three forces acting on the block under the assumption that the mass is 1 kg and $g = 9.81$ m/s^2 (recall that g changes from place to place on earth). On the highest part of the ramp on the left, the normal force is zero. Also, although the normal force in the upper part of this ramp is pointing down, the block does not fall down due to its speed.

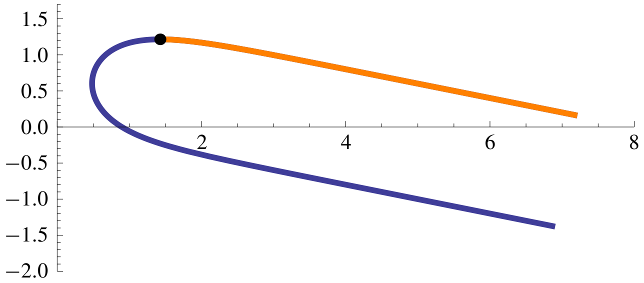


Figure 3. Graph of the two *constant-speed ramps*. At the highest point, a block with speed v_0 must be placed on top if we want to use the ramp to the right of this highest point, and it must be placed underneath if we want to use the ramp to the left of this highest point.

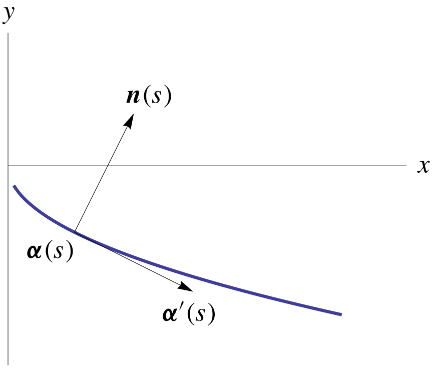


Figure 4. Given that s is the arc-length parameter, the vectors $\boldsymbol{\alpha}'(s) = (\cos\theta(s), \sin\theta(s))$ and $\boldsymbol{n}(s) = (-\sin\theta(s), \cos\theta(s))$ are perpendicular.

In Section 2 we provide a first proof of the classification of 2-dimensional ramps. In Section 3 we provide a formal definition of ramp in order to state the result as a mathematical theorem. This section may be skipped. Section 4 deals with 3-dimensional ramps.

2. A first approach

Let us find every possible 2-dimensional curve with the property that a block, sliding down on it, will move with constant speed $v > 0$ under the assumption that the kinetic coefficient of friction is $\mu = \tan\delta$ for some constant δ between 0 and $\pi/4$ radians. Start by assuming that this curve is parametrized by arclength; that is, if $\boldsymbol{\alpha}(s) = (x(s), y(s))$ denotes such a curve, then $|\boldsymbol{\alpha}'(s)| = 1$. Under this assumption we can assume that for a smooth function $\theta(s)$ we have

$$x'(s) = \cos\theta(s) \quad \text{and} \quad y'(s) = \sin\theta(s).$$

The function $\theta(s)$ will help us describe the curve $\boldsymbol{\alpha}$. It is clear that the vector $\boldsymbol{n}(s) = (-\sin\theta(s), \cos\theta(s))$ is a unit vector perpendicular to $\boldsymbol{\alpha}'(s)$, and the chain rule gives that $\boldsymbol{\alpha}''(s) = \theta'(s)\boldsymbol{n}(s)$. Figure 4 shows these two vectors.

Let us assume that $\boldsymbol{\beta}(t) = \boldsymbol{\alpha}(vt)$ describes the motion of the block. Since $s = vt$ and s is the arc-length parameter, the speed of the block is the constant v. Since we have that $\boldsymbol{\beta}''(t) = v^2\boldsymbol{\alpha}''(vt)$, we can write the free-body diagram for the constant-speed problem on $\boldsymbol{\alpha}$ with speed v as in Figure 5.

As the free-body diagram shows, the following equation must hold true (a free-body diagram is essentially a picture that shows the forces acting on a body. For further discussion on free-body diagrams, see [Beer et al. 2010]):

$$mv^2\theta'(s)\boldsymbol{n}(s) = \lambda(s)\boldsymbol{n}(s) - \tan(\delta)\lambda(s)\boldsymbol{\alpha}'(s) - (0, mg). \tag{2-1}$$

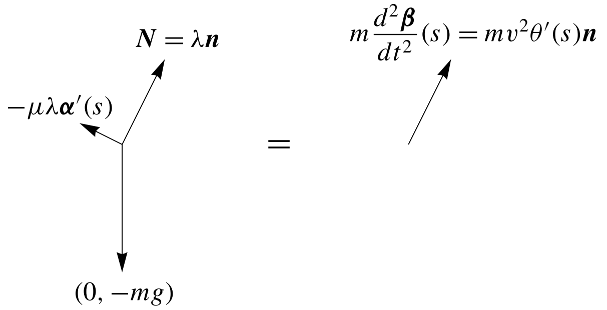


Figure 5. By Newton's second law, the sum of the normal force $\boldsymbol{N}$, the weight $(0, -mg)$ and the friction force $-\mu\lambda\boldsymbol{\alpha}'(s)$ must be $m\boldsymbol{\beta}''(s)$.

By applying the inner product with the vector $\boldsymbol{\alpha}'(s)$ to both sides of (2-1) we obtain

$$\lambda(s) = -mg\cot\delta\sin\theta(s).$$

Likewise, applying the inner product with the vector $\boldsymbol{n}(s)$ to both sides of (2-1), we obtain that $mv^2\theta' = \lambda - mg\cos\theta$. Therefore, with $a = g/(v^2\sin\delta)$,

$$\theta'(s) = \frac{-g}{v^2\sin\delta}\sin(\theta(s)+\delta) = -a\sin(\theta(s)+\delta). \tag{2-2}$$

Since the differential equation (2-2) does not have the variable s, all the solutions differ by a horizontal translation. That is, if $\theta(s)$ is a solution, then $\theta(s+c)$ is also a solution for every real number c. Due to the geometry of our problem we do not need to consider an integrating constant, since just one solution will give us all the solutions $\boldsymbol{\alpha}(s)$. Recall that the equilibrium solution of (2-2) is $\theta(s) = -\delta$ for all s. This solution corresponds to the case $\boldsymbol{\alpha}(s) = (\cos(\delta)s, -\sin(\delta)s)$, which is the straight line ramp shown in Figure 1. When we solve this differential equation by separation of variables we notice that we need to integrate the function $\csc(\theta+\delta)$. Instead of using the classical formula $\int\csc u\,du = -\ln(\csc u + \cot u)$ we will use the formula $\int\csc u\,du = \ln(\tan(u/2))$, which leads to the formula

$$\theta(s) = -\delta + 2\arctan(\mathrm{e}^{-as}).$$

It is clear that if $\boldsymbol{\gamma}(s) = (z(s), w(s))$ denotes a counterclockwise rotation of δ radians of the curve $\boldsymbol{\alpha}(s)$, then

$$z'(s) = \cos(2\arctan(\mathrm{e}^{-as})) \quad\text{and}\quad w'(s) = \sin(2\arctan(\mathrm{e}^{-as})).$$

Integrating the equations above we obtain that

$$z(s) = s + \frac{\ln(1+\mathrm{e}^{-2as})}{a} \quad\text{and}\quad w(s) = \frac{2}{a}\operatorname{arccot}(\mathrm{e}^{-as}).$$

The curve $(z(s), w(s))$ is shown in Figure 2. As mentioned in the introduction, the solution that we are looking for is a clockwise rotation of the curve $\boldsymbol{\gamma}$ by an angle δ. In the next section we give a more detailed explanation of this solution.

3. Understanding the solution of the ODE: a mathematical definition of ramps

With the intention of setting up notation, let us start with the following well-known definition (see, for example, [do Carmo 1976]).

Definition 3.1. We will say that a curve $\boldsymbol{\gamma} : [a_1, a_2] \to \mathbb{R}^2$ is *regular* if $\boldsymbol{\gamma}'(t)$ never vanishes. We say $\boldsymbol{n} : [a_1, a_2] \to \mathbb{R}^2$ is a *normal* of the curve $\boldsymbol{\gamma}$ if $\boldsymbol{n}(t)$ has length 1 and the inner product of $\boldsymbol{n}(t)$ and $\boldsymbol{\gamma}'(t)$ is zero, that is, $\boldsymbol{n}(t) \cdot \boldsymbol{\gamma}(t) = 0$ for all t.

Intuitively, we can think of a planar ramp as a 2-dimensional region whose boundary is a curve. It is clear that the interesting part of the ramp is a portion of the boundary of the region. The following definition provides an alternative way to describe a ramp without mentioning the 2-dimensional region. Figure 6 provides the interpretation of our definition.

Definition 3.2. A *ramp* in the plane $\mathbb{R}^2$ is an ordered pair $(\boldsymbol{\gamma}, \boldsymbol{n})$, where $\boldsymbol{\gamma} : [a_1, a_2] \to \mathbb{R}^2$ is a regular curve and $\boldsymbol{n} : [a_1, a_2] \to \mathbb{R}^2$ is a normal to $\boldsymbol{\gamma}$. We will interpret the ramp $(\boldsymbol{\gamma}, \boldsymbol{n})$ as a portion of the plane whose boundary contains $\boldsymbol{\gamma}$ and whose outer normal vector is $\boldsymbol{n}$.

Example 3.3. The pairs $(\boldsymbol{\gamma}_1, \boldsymbol{n}_1)$ and $(\boldsymbol{\gamma}_2, \boldsymbol{n}_2)$, mapping $[0, \pi] \to \mathbb{R}^2$ and given by

$$\begin{aligned} \boldsymbol{\gamma}_1(t) &= 5(\cos t, \sin t) \quad \text{and} \quad \boldsymbol{n}_1(t) = (\cos t, \sin t), \\ \boldsymbol{\gamma}_2(t) &= 5(\cos t, \sin t) \quad \text{and} \quad \boldsymbol{n}_2(t) = -(\cos t, \sin t), \end{aligned}$$

are examples of ramps, shown in Figure 6.

The following definition is based on Newton's second law:

Definition 3.4. A *ramp with external force* is a triple $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$, where $\boldsymbol{F} : [a_1, a_2] \to \mathbb{R}^2$ is a smooth function and $(\boldsymbol{\gamma}, \boldsymbol{n})$ is a ramp. Given a positive number m, a *solution of the ramp with external force* $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$ is given by a curve $\boldsymbol{\beta}(t) = \boldsymbol{\gamma}(h(t))$ and a nonnegative function $\lambda : [a_1, a_2] \to \mathbb{R}$ such that

$$\boldsymbol{F}(h(t)) + \lambda(t)\boldsymbol{n}(h(t)) = m\boldsymbol{\beta}''(t).$$

(Notice that $\boldsymbol{\beta}$ is just a reparametrization of the curve $\boldsymbol{\gamma}$ and is completely determined by the function $h : [a_1, a_2] \to \mathbb{R}$.)

Remark 3.5. In Definition 3.4, the term $\lambda(t)\boldsymbol{n}(t)$ represents the force that the surface of the ramp is exerting on the object under the action of the external force $\boldsymbol{F}$. By Newton's third law, $-\lambda(t)\boldsymbol{n}(t)$ is the force that the object is exerting on the

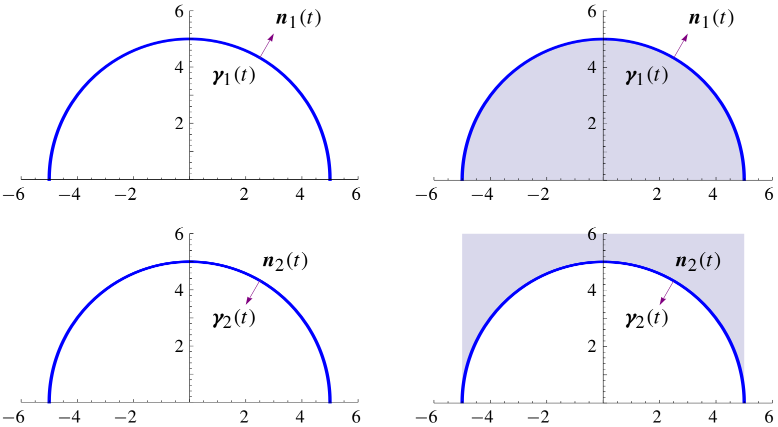


Figure 6. A normal vector of a curve helps us to define the part of a curve where we want a object to slide on a ramp

ramp. The condition $\lambda \geq 0$ is needed so that the object stays on the ramp. Also notice that the curve $\boldsymbol{\beta}(t)$ describes the position of the object at time t.

Example 3.6. If $(\boldsymbol{\gamma}, \boldsymbol{n})$ is a ramp, $m > 0$ and $\boldsymbol{F}(t) = (0, -mg)$, where g is the gravitational acceleration, then the triple $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$ represents the action of gravity on a particle with mass m that moves on the ramp without friction.

The following easy example illustrates that not every ramp with external force has a solution.

Example 3.7. Let us consider the ramp $(\boldsymbol{\gamma}, \boldsymbol{n})$ where $\boldsymbol{\gamma}$, $\boldsymbol{n} : [-1.5, 1.5] \to \mathbb{R}^2$ are given by

$$\boldsymbol{\gamma}(t) = (t, t^2) \quad \text{and} \quad \boldsymbol{n}(t) = \left(\frac{2t}{\sqrt{4t^2+1}}, \frac{-1}{\sqrt{4t^2+1}} \right).$$

If we assume that the mass of the block is 1 kg, and $\boldsymbol{F}(t) = (0, -9.81)$, then the ramp with external force $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$ has no solution. Figure 7 shows this ramp.

Proof. Let us argue by contradiction. If a solution exists, there will exist a positive function $\lambda(t)$ and a function $h(t)$ such that

$$\begin{aligned}(0, -9.81) + \lambda(t) \left(\frac{2h(t)}{\sqrt{4h^2(t)+1}}, \frac{-1}{\sqrt{4h^2(t)+1}} \right) &= \boldsymbol{\beta}''(t) \\ &= \big(h''(t), 2(h'(t))^2 + 2h(t)h''(t)\big).\end{aligned}$$

If we take the dot product with the vector $(2h(t), -1)$ on both sides of the equation above we obtain the equation

$$9.81 + \lambda(t)\sqrt{4h^2(t)+1} = -2(h'(t))^2.$$

This is impossible because we assumed that $\lambda(t)$ is a positive function. □

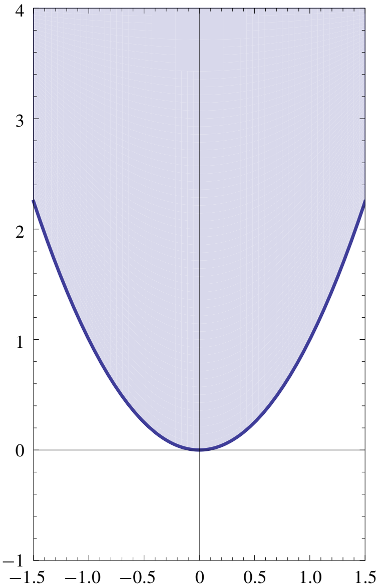


Figure 7. This ramp with the external force $\boldsymbol{F}(t) = (0, -9.81)$ does not have a solution.

The definition of the solution of a particle moving on the ramp under the action of the force $\boldsymbol{F}$ is provided by Definition 3.4. In the case that $\boldsymbol{F}$ denotes all the forces acting on the particle that moves on the ramp *except* the friction force, the definition of the solution needs to be changed to:

Definition 3.8. Given a ramp with external force $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$ defined on the interval $[a_1, a_2]$ and two positive numbers $\mu < 1$ and m, a *solution of the ramp with external force* $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$ *and kinetic coefficient of friction* μ is given by a curve $\boldsymbol{\beta}(t) = \boldsymbol{\gamma}(h(t))$ and a nonnegative function $\lambda : [a_1, a_2] \to \mathbb{R}$ such that

$$\boldsymbol{F}(h(t)) + \lambda(t)\boldsymbol{n}(h(t)) - \mu\lambda(t)\frac{\boldsymbol{\beta}'(t)}{|\boldsymbol{\beta}'(t)|} = m\boldsymbol{\beta}''(t).$$

Intuitively, when we think of a block moving with constant speed on a ramp, we think of the block moving down. The following proposition shows, using Definition 3.8, that indeed the block must move down.

Proposition 3.9. *Let us assume that* $\boldsymbol{F} = (0, -mg)$ *is the gravitational force, where m is a mass, and* μ *is a kinetic coefficient of friction. Let us further assume that* $\boldsymbol{\beta}(t) = \boldsymbol{\gamma}(h(t))$ *is a solution of the ramp with external force* $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$ *and kinetic coefficient of friction* μ. *If the speed of the solution is constant, then the velocity vector of the solution* $\boldsymbol{\beta}$ *cannot point up; that is, for any t, the dot product* $\boldsymbol{\beta}'(t) \cdot (0, 1)$ *cannot be positive.*

Proof. Since we are assuming that the curve $\boldsymbol{\gamma}$ is regular, we can find a reparametrization $\tilde{\boldsymbol{\gamma}}$ that is parametrized by arclength (see for example [do Carmo 1976]). That is, we can write $\boldsymbol{\gamma}(t) = \tilde{\boldsymbol{\gamma}}(u(t))$. We therefore have that

$$\boldsymbol{\beta}(t) = \boldsymbol{\gamma}(h(t)) = \tilde{\boldsymbol{\gamma}}(u(h(t)) = \tilde{\boldsymbol{\gamma}}(\tilde{h}(t)), \quad \text{where } \tilde{h} = u \circ h.$$

Since $\tilde{\boldsymbol{\gamma}}$ is parametrized by arclength and we are assuming that the speed of $\boldsymbol{\beta}$ is constant, we have that the function $\tilde{h}'$ is constant; that is $\tilde{h}'(t) = v$ for all t and for some nonzero real number v. Therefore $\boldsymbol{\beta}''(t) = v^2 \tilde{\boldsymbol{\gamma}}''(t)$. Since we are assuming that $\boldsymbol{\beta}$ is a solution of the ramp with external force $(\boldsymbol{\gamma}, \boldsymbol{n}, \boldsymbol{F})$ and kinetic coefficient of friction μ,

$$(0, -mg) + \lambda(t)\boldsymbol{n}(h(t)) - \mu\lambda(t)\frac{\boldsymbol{\beta}'(t)}{|\boldsymbol{\beta}'(t)|} = m\boldsymbol{\beta}''(t).$$

Multiplying the equation above by the velocity vector $\boldsymbol{\beta}'(t) = v\tilde{\boldsymbol{\gamma}}'(t)$ and keeping in mind that (i) this velocity vector is perpendicular to the normal vector $\boldsymbol{n}$ and (ii) the acceleration vector $\boldsymbol{\beta}''$ is parallel to the normal vector $\boldsymbol{n}$, we obtain

$$-mg\boldsymbol{\beta}'(t) \cdot (0, 1) - \mu\lambda(t)|v| = 0.$$

From the equation above we conclude that $\boldsymbol{\beta}'(t) \cdot (0, 1)$ cannot be positive. □

Remark 3.10. Let us define an *elementary function* to be a function of one variable with real values built from a finite number of trigonometric, exponential, constant, n-th power functions and their inverses, through composition and the four basic elementary operations $(+, -, \times, \div)$. When we want to define basic examples of curves parametrized by arclength (that is, if we want to define $\boldsymbol{\gamma}(s) = (x(s), y(s))$ such that $x'(s)^2 + y'(s)^2 = 1$ for all t), the first thing that comes to mind is to find easy possibilities for $x'(t)$ and $y'(t)$. The easiest one would be a pair of numbers c_1 and c_2 such that $c_1^2 + c_2^2 = 1$. If we make $x'(s) = c_1$ and $y'(s) = c_2$, then, after an easy integration, we obtain that the curve $\boldsymbol{\gamma}$ is a straight line. If we want to use the fact that $\cos^2(as) + \sin^2(as) = 1$ and we decide to make $x'(s) = \cos(as)$ and $y'(s) = \sin(as)$, then, after an easy integration, we obtain that $\boldsymbol{\gamma}$ should be a circle. The most important curves in this paper are those that are obtained by using the identity

$$\tanh^2 t + \operatorname{sech}^2 t = 1 \quad \text{for all } t.$$

That is, we will be using curves that satisfy $x'(s) = \tanh(as)$ and $y'(s) = \operatorname{sech}(as)$.

The following lemma is a direct computation and provides a definition of the curve $\boldsymbol{\alpha}$ whose graph is shown in Figure 2.

Lemma 3.11. *For any nonzero real number a, the curve*

$$\boldsymbol{\alpha}(s) = (x(s), y(s)) = \Big(s + \frac{1}{a}\ln(1 + \mathrm{e}^{-2as}), \frac{2}{a}\operatorname{arccot}(\mathrm{e}^{-as})\Big) \tag{3-1}$$

is an arc-length parametrized curve. Moreover,

$$x'(s) = \tanh(as) \quad \textit{and} \quad y'(s) = \operatorname{sech}(as).$$

Lemma 3.12. *For any positive number δ smaller than $\pi/2$, suppose $\boldsymbol{\alpha}_\delta$ represents the curve obtained by rotating the curve $\boldsymbol{\alpha} = (x(s), y(s))$ defined in* Lemma 3.11 *through an angle of δ radians; that is, let*

$$\begin{aligned}\boldsymbol{\alpha}_\delta(s) &= (x_\delta(s), y_\delta(s)) \\ &= \big(\cos(\delta)x(s) + \sin(\delta)y(s), -\sin(\delta)x(s) + \cos(\delta)y(s)\big).\end{aligned}$$

Then the maximum value for y_δ is achieved when $s = \operatorname{arcsinh}(\cot\delta)/a$. Also, we have that

$$\boldsymbol{\alpha}''_\delta(s) \cdot (y'_\delta(s), -x'_\delta(s)) = a\operatorname{sech}(as). \tag{3-2}$$

The graph of the curve $\boldsymbol{\alpha}_\delta$ is shown in Figure 3.

Proof. A direct computation using Lemma 3.11 shows that

$$y'_\delta(s) = -\sin\delta\tanh(as) + \cos\delta\operatorname{sech}(as),$$

and so the only solution of the equation $y'_\delta(s) = 0$ is $s = \operatorname{arcsinh}(\cot\delta)/a$. In order to prove the identity (3-2) we point out that since the inner product is invariant under rotations, this identity is equivalent to showing that $\boldsymbol{\alpha}''(s) \cdot (y'(s), -x'(s)) = a\operatorname{sech}(as)$, which follows because

$$\begin{aligned}x''(s)y'(s) - y''(s)x'(s) &= a\operatorname{sech}^3 as + a\tanh^2 as\operatorname{sech} as \\ &= a\operatorname{sech} as\big(\operatorname{sech}^2 as + \tanh^2 as\big) = a\operatorname{sech} as. \qquad \square\end{aligned}$$

Remark 3.13. In this proof, we obtained $y'_\delta(s) = -\sin\delta\operatorname{sech}(as)(\sinh(as) - \cot\delta)$. It follows that $y'_\delta(s) > 0$ if $s < s_0$ and $y'_\delta(s) < 0$ if $s > s_0$.

Definition 3.14. For any positive number δ smaller than $\pi/2$ and any $a > 0$, we define the ramps $(\boldsymbol{\gamma}_\delta, \boldsymbol{n}_\delta)$ and $(\tilde{\boldsymbol{\gamma}}_\delta, \tilde{\boldsymbol{n}}_\delta)$ on the interval $[0, \infty)$ by

$$\begin{aligned}\boldsymbol{\gamma}_\delta(s) &= \boldsymbol{\alpha}_\delta(s_0 - s), & \boldsymbol{n}_\delta(s) &= (y'_\delta(s_0 - s), -x'_\delta(s_0 - s)), \\ \tilde{\boldsymbol{\gamma}}_\delta(s) &= \boldsymbol{\alpha}_\delta(s + s_0), & \tilde{\boldsymbol{n}}_\delta(s) &= (-y'_\delta(s + s_0), x'_\delta(s + s_0)),\end{aligned}$$

where $s_0 = \operatorname{arcsinh}(\cot\delta)/a$ and the map $\boldsymbol{\alpha}_\delta$ and the functions x_δ and y_δ are defined in Lemma 3.12. These two ramps are shown in Figure 8.

Let us state and proof the main theorem in this section:

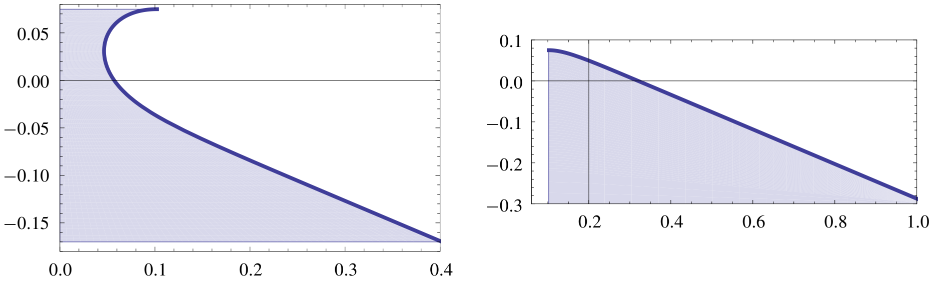


Figure 8. The two ramps given in Definition 3.14.

Theorem 3.15. *Given an angle δ between* 0 *and* $\pi/4$ *radians and a speed* $v > 0$, *set* $\mu = \tan\delta$, $a = g/(v^2 \sin\delta)$ *and* $\boldsymbol{F}(t) = (0, -mg)$, *where g is the acceleration due to gravity. We have*:

(a) *For the ramp* $(\boldsymbol{\gamma}_\delta, \boldsymbol{n}_\delta)$ *given in* Definition 3.14,

$$\boldsymbol{\beta}(t) = \boldsymbol{\gamma}_\delta(vt) \quad \textit{and} \quad \lambda(t) = mg\cot\delta\, y_\delta'(s_0 - vt)$$

is a solution for the ramp with external force $(\boldsymbol{\gamma}_\delta, \boldsymbol{n}_\delta, \boldsymbol{F})$ *and kinetic coefficient of friction* $\tan\delta$.

(b) *For the ramp* $(\tilde{\boldsymbol{\gamma}}_\delta, \tilde{\boldsymbol{n}}_\delta)$ *given in* Definition 3.14,

$$\boldsymbol{\beta}(t) = \tilde{\boldsymbol{\gamma}}_\delta(vt) \quad \textit{and} \quad \lambda(t) = -mg\cot\delta\, y_\delta'(s_0 + vt)$$

is a solution for the ramp with external force $(\tilde{\boldsymbol{\gamma}}_\delta, \tilde{\boldsymbol{n}}_\delta, \boldsymbol{F})$ *and kinetic coefficient of friction* $\tan\delta$.

(c) *Since in both cases* $|\boldsymbol{\beta}'(t)| = v$, *these motions have constant speed.*

Proof. Let us prove part (a). First of all, notice that, as is required by Definition 3.8, $\lambda > 0$ by Remark 3.13. From the definition of $\boldsymbol{\beta}$ we obtain that

$$\boldsymbol{\beta}'(t) = -v\boldsymbol{\alpha}_\delta'(s_0 - vt) \quad \text{and} \quad \boldsymbol{\beta}''(t) = v^2\,\boldsymbol{\alpha}_\delta''(s_0 - vt).$$

Therefore, the equation

$$\boldsymbol{F}(h(t)) + \lambda(t)\boldsymbol{n}(h(t)) - \mu\lambda(t)\frac{\boldsymbol{\beta}'(t)}{|\boldsymbol{\beta}'(t)|} = m\boldsymbol{\beta}''(t)$$

is equivalent to

$$(0, -mg) + \lambda(t)\boldsymbol{n}_\delta(vt) + \tan\delta\lambda(t)\boldsymbol{\alpha}_\delta'(s_0 - vt) = mv^2\,\boldsymbol{\alpha}_\delta''(t). \tag{3-3}$$

Notice that the vectors $\boldsymbol{u}_1 = \boldsymbol{n}_\delta(vt)$ and $\boldsymbol{u}_2 = \boldsymbol{\alpha}_\delta'(s_0 - tv)$ form an orthonormal basis. Therefore, in order to establish (3-3) it is enough to prove that the dot products with $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of the left-hand side and right-hand side of the equation are the same. The dot product of the LHS of (3-3) with $\boldsymbol{u}_1$ is equal to

$$\begin{aligned} mg\,x_\delta'(s_0-vt)+\lambda(t) &= mg\,x_\delta'(s_0-vt)+mg\cot\delta y_\delta'(s_0-vt)\\ &= mg\big(\cos\delta\tanh(a(s_0-vt))+\sin\delta\,\mathrm{sech}(a(s_0-vt))\big)\\ &\quad +mg\cot\delta\big(-\sin\delta\tanh(a(s_0-vt))+\cos\delta\,\mathrm{sech}(a(s_0-vt))\big)\\ &= \frac{mg}{\sin\delta}\,\mathrm{sech}(a(s_0-vt)). \end{aligned}$$

Using (3-2) we get that the dot product of the RHS of (3-3) with $\boldsymbol{u}_1$ is equal to

$$mv^2a\,\mathrm{sech}(a(s_0-vt))=\frac{mg}{\sin\delta}\,\mathrm{sech}(a(s_0-vt)).$$

Therefore, $\mathrm{LHS}\cdot\boldsymbol{u}_1=\mathrm{RHS}\cdot\boldsymbol{u}_1$. It is easy to check that the dot product of the RHS of (3-3) with $\boldsymbol{u}_2$ vanishes. On the other hand, the dot product of the LHS of (3-3) with $\boldsymbol{u}_2$ equals

$$-mg\,y_\delta'(s_0-vt)+v\tan\delta\,\lambda=-mg\,y_\delta'(s_0-vt)+\tan\delta\,mg\cot\delta y_\delta'(s_0-vt)=0.$$

Therefore, part (a) follows. Part (b) is similar. Part (c) follows because $|\boldsymbol{\alpha}'|=1$, and, since $\boldsymbol{\alpha}_\delta$ is a rotation of $\boldsymbol{\alpha}$, we have $|\boldsymbol{\alpha}_\delta'|=1$ and $\boldsymbol{\beta}'(t)=-v\boldsymbol{\alpha}'(s_0-vt)$, so $|\boldsymbol{\beta}'(t)|=v$. □

4. Three-dimensional ramps

In this section we describe ramps in the space on which an object can move with constant speed v_0 under the assumption that the kinetic coefficient of friction is μ. We will prove that, if v_0 and μ are fixed, then for every continuously differentiable unit tangent vector field in the southern hemisphere there will correspond a constant-speed family of ramps. In the correspondence that we will establish, the ramp that we defined in Section 2 corresponds to a particular choice of a tangent vector field.

For curves that represent planar ramps, we obtained a description by studying their velocity vector. Recall that the function $\theta(s)$ was used to describe a point in the lower part of the semicircle that represented a possible velocity vector of the curve that described the ramp. When this curve is free to move in the whole 3-dimensional space, the tangent unit vector to the curve lies in the unit sphere, and since the block must go down the ramp we can assume that the tangent unit vector of the ramp lies in the southern hemisphere. See Figure 9.

In order to completely determine the ramp, we need to specify the desired direction of the normal to the ramp; see Figure 10. We can do this by choosing a unit normal vector field $\boldsymbol{H}(y)$ defined in the lower hemisphere and requesting the normal vector of the ramp to be equal to $\boldsymbol{N}(y)$ any time the tangent unit vector to the ramp is the vector y. See Figure 11.

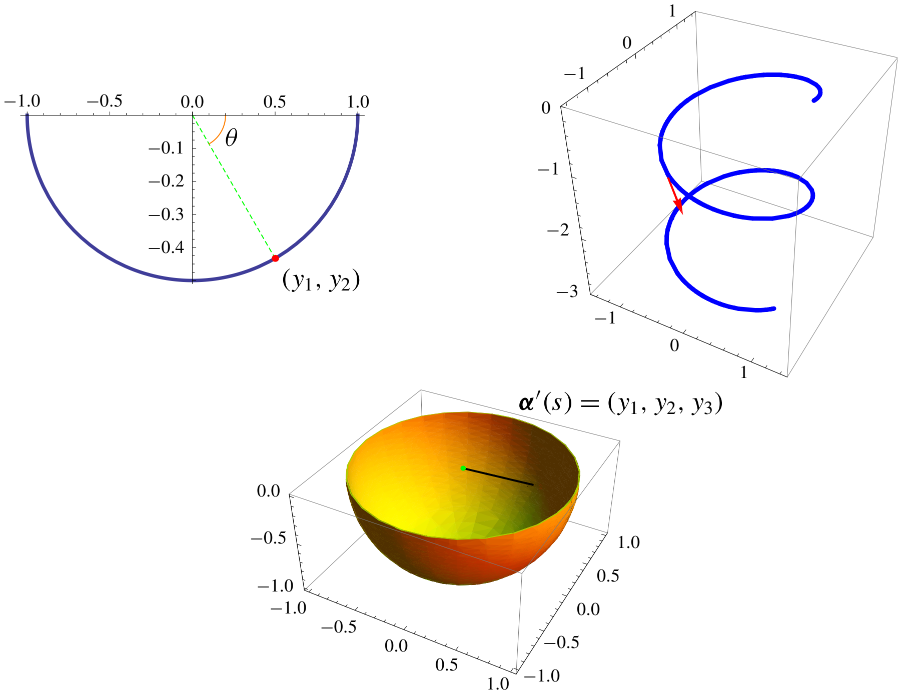


Figure 9. The unit tangent vector of a possible ramp in the plane lies in the lower part of the unit semicircle. If the ramp does not lie in a plane, but instead in the whole space, then the unit tangent vector lies in the lower hemisphere of the unit sphere.

We will prove that for any continuously differentiable choice of a normal field $\boldsymbol{n}(s)$ in the lower hemisphere, there exists a family of ramps on which a motion with constant speed is possible under the effect of the forces of gravity and friction.

More precisely, we have the following theorem:

Theorem 4.1. *Let* $\Sigma = \{(y_1, y_2, y_3) \in \mathbb{R}^3 : y_1^2 + y_2^2 + y_3^2 = 1 \text{ and } y_3 \leq 0\}$ *be the southern hemisphere and let* $\boldsymbol{N} : \Sigma \to \mathbb{R}^3$ *be any continuously differentiable unit tangent vector field. That is, for any* $y \in \Sigma$, $\boldsymbol{N}(y)$ *has norm* 1 *and is perpendicular to* y. *For any positive numbers* m, v *and* $\mu < 1$, *and any* $y_0 \in \Sigma$, *there exists a unique curve* $\boldsymbol{\alpha}(s)$ *parametrized by arclength, such that* $\boldsymbol{\alpha}(0) = (0, 0, 0)$ *and* $\boldsymbol{\alpha}'(0) = y_0$. *Furthermore, the motion* $\boldsymbol{\beta}(t) = \boldsymbol{\alpha}(vt)$ *in the ramp given by*

$$\boldsymbol{R}(s, r) = \boldsymbol{\alpha}(s) + r\boldsymbol{\alpha}'(s) \times \boldsymbol{N}(\boldsymbol{\alpha}'(s))$$

represents a solution of Newton's second law for a block with mass m *moving on the ramp under the assumption that the only forces acting on the block are gravity and friction with kinetic constant coefficient* μ. *Recall that since the parameter* s *is the arclength,* $|\boldsymbol{\beta}'(t)| = v$ *for all* t; *that is, the motion has constant speed* v.

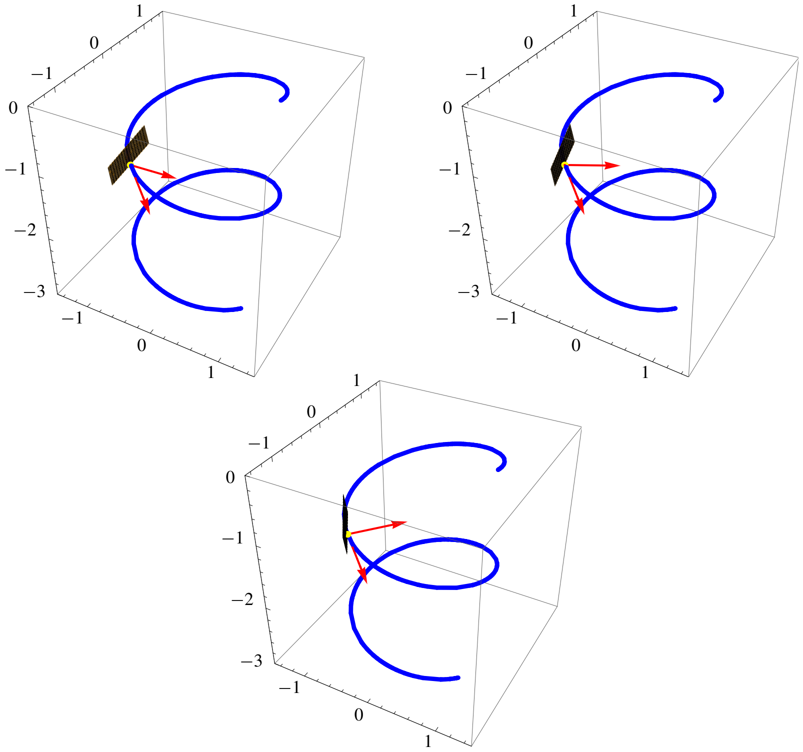


Figure 10. To build a ramp we need to choose a normal direction at every point in the ramp.

Proof. Recall that we denote by $\boldsymbol{u} \cdot \boldsymbol{w}$ the inner product of the vectors $\boldsymbol{u}$ and $\boldsymbol{w}$. A direct computation shows that if $\boldsymbol{e}_3 = (0,0,1)$, then $\boldsymbol{e}_3^T : \Sigma \to \mathbb{R}^3$ given by $\boldsymbol{e}_3^T(y) = \boldsymbol{e}_3 - (\boldsymbol{e}_3 \cdot y)\, y$ is a tangent vector field. Notice that $\boldsymbol{e}_3^T(y)$ must be in the tangent space $T_y\Sigma$ because $\boldsymbol{e}_3^T(y) \cdot y = 0$.

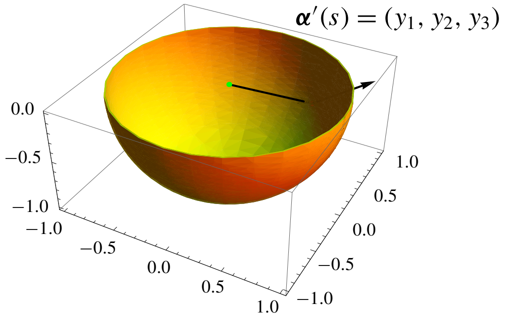


Figure 11. We will make the choice of the normal direction on the ramp to depend on the unit tangent vector.

Let $\lambda : \Sigma \to \mathbb{R}$ be the function given by

$$\lambda(y) = -\frac{g}{\mu}\boldsymbol{e}_3 \cdot y = -\frac{g}{\mu} y_3.$$

Since $y_3 \leq 0$, $\lambda \geq 0$ and λ only vanishes on the boundary of Σ. (Recall that the boundary of Σ is given by the equation $y_3 = 0$.) Let us consider the tangent vector field

$$\text{(4-1)} \qquad \boldsymbol{X} = -\frac{1}{v^2}(g\boldsymbol{e}_3^T - \lambda \boldsymbol{N}(y)) = -\frac{g}{v^2}\Big(\boldsymbol{e}_3^T + \frac{y_3}{\mu}\boldsymbol{N}(y)\Big).$$

Observe that along points in the boundary of Σ, $\boldsymbol{X}(y) = -(g/v^2)\,\boldsymbol{e}_3$ points toward Σ. It follows that any integral curve of the tangent vector field $\boldsymbol{X}$ remains in Σ. Let $\boldsymbol{\gamma}(s)$ be the integral curve of the vector field $\boldsymbol{X}$ that satisfies $\boldsymbol{\gamma}(0) = y_0$. We will prove the theorem by showing the curve $\boldsymbol{\alpha}(s)$ given by

$$\boldsymbol{\alpha}(s) = \int_0^s \boldsymbol{\gamma}(u)\,du$$

satisfies all the conditions in the theorem.

Clearly, $\boldsymbol{\alpha}(0) = (0,0,0)$ and $\boldsymbol{\alpha}'(0) = \boldsymbol{\gamma}(0) = y_0$. We also have that s is an arc-length parameter because $|\boldsymbol{\alpha}'(s)| = |\boldsymbol{\gamma}(s)| = 1$. A direct computation shows that the normal vector of the surface (or ramp) $\boldsymbol{R}$ along the curve $\boldsymbol{\alpha}$ (along points in the surface with $r = 0$) is given by $\boldsymbol{N}(\boldsymbol{\gamma}(s))$.

We will now prove that Newton's second law holds true for $\boldsymbol{\beta}(t) = \boldsymbol{\alpha}(vt)$. Notice that since v is constant, $\boldsymbol{\beta}'(t) = v\boldsymbol{\alpha}'(vt) = v\boldsymbol{\gamma}(s)$ and $\boldsymbol{\beta}''(vt) = v^2\boldsymbol{\gamma}'(vt) = v^2\boldsymbol{\gamma}'(s)$. If m denotes the mass of the particle and λ is defined as in the beginning of the proof, we have

$$\begin{aligned} m\boldsymbol{\beta}''(t) = mv^2\boldsymbol{\gamma}'(s) &= -mv^2\frac{1}{v^2}(g\,\boldsymbol{e}_3^T(\boldsymbol{\gamma}) - \lambda \boldsymbol{N}(\boldsymbol{\gamma})) \\ &= m\Big(-g\,\boldsymbol{e}_3 + g(\boldsymbol{e}_3 \cdot \boldsymbol{\gamma})\,\boldsymbol{\gamma} - \frac{g}{\mu}(\boldsymbol{e}_3 \cdot \boldsymbol{\gamma})\,\boldsymbol{N}(\boldsymbol{\gamma}(t))\Big) \\ &= -mg\boldsymbol{e}_3 + m\lambda \boldsymbol{N}(\boldsymbol{\gamma}(s)) - m\lambda\mu\boldsymbol{\gamma}(s). \end{aligned}$$

From the equation above, we conclude that the normal force imparted by the ramp onto the block has magnitude $m\lambda$. Notice that the last equation above is Newton's second law. □

It is clear that the form of the ramp depends on the desired constant speed of the block. For example, if the block is to travel upside down on some portion of the ramp and the speed v is not much, then the curvature of that portion of the ramp must be big so that the block does not fall off the ramp. The following corollary explains how to change the ramp if we want to change the constant speed or if we want to change the gravity force:

Corollary 4.2. *Suppose that a ramp $\boldsymbol{R} \subset \mathbb{R}^3$ allows a block to move down with constant speed v under a gravity g. Then, for any positive κ, the ramp $\kappa\boldsymbol{R} = \{\kappa(x, y, z) : (x, y, z) \in \boldsymbol{R}\}$ allows a block to move down*

(a) *with constant speed $\sqrt{\kappa}\, v$ and gravity g, or*

(b) *with constant speed v and gravity g/κ.*

Proof. This corollary is a consequence of the definition of the vector field $\boldsymbol{X}$ in (4-1). We can check that if $\boldsymbol{\gamma}(t)$ is an integral curve of the vector field $\boldsymbol{X}$ and $\boldsymbol{\alpha}(s) = \int_0^s \boldsymbol{\gamma}(u)\, du$, then $\tilde{\boldsymbol{\gamma}}(\tau) = \boldsymbol{\gamma}(\tau/\kappa)$ is a solution of the vector field $(1/\kappa)\boldsymbol{X}$, which can be interpreted as either the vector field coming from the tangent unit vector field $\boldsymbol{N}(y)$ with velocity $\sqrt{\kappa}\, v$ and gravity g or can be interpreted as the vector field coming from the tangent unit vector field $\boldsymbol{N}(y)$ with velocity v and gravity g/κ. The result follows by noticing that

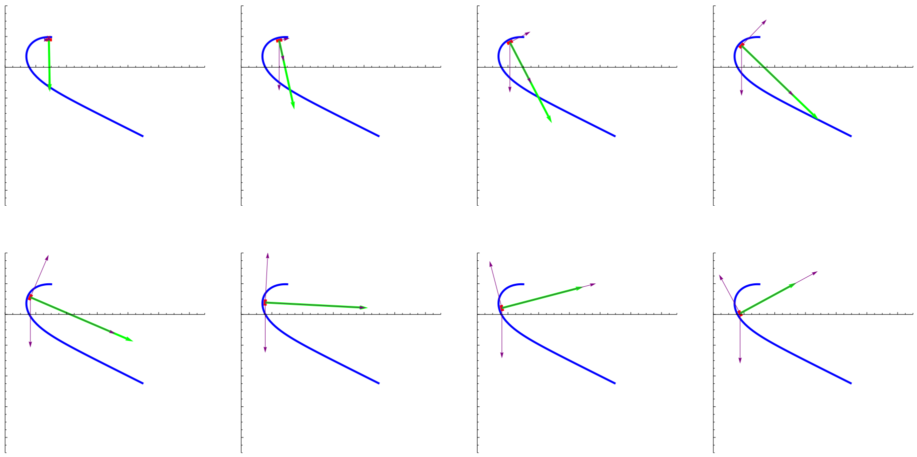
$$\tilde{\boldsymbol{\alpha}}(s) = \int_0^s \tilde{\boldsymbol{\gamma}}(u)\, du = \int_0^s \boldsymbol{\gamma}\left(\frac{u}{\kappa}\right) du = \kappa \int_0^{s/\kappa} \boldsymbol{\gamma}(v)\, dv = \kappa\boldsymbol{\alpha}\left(\frac{s}{\kappa}\right). \qquad \square$$

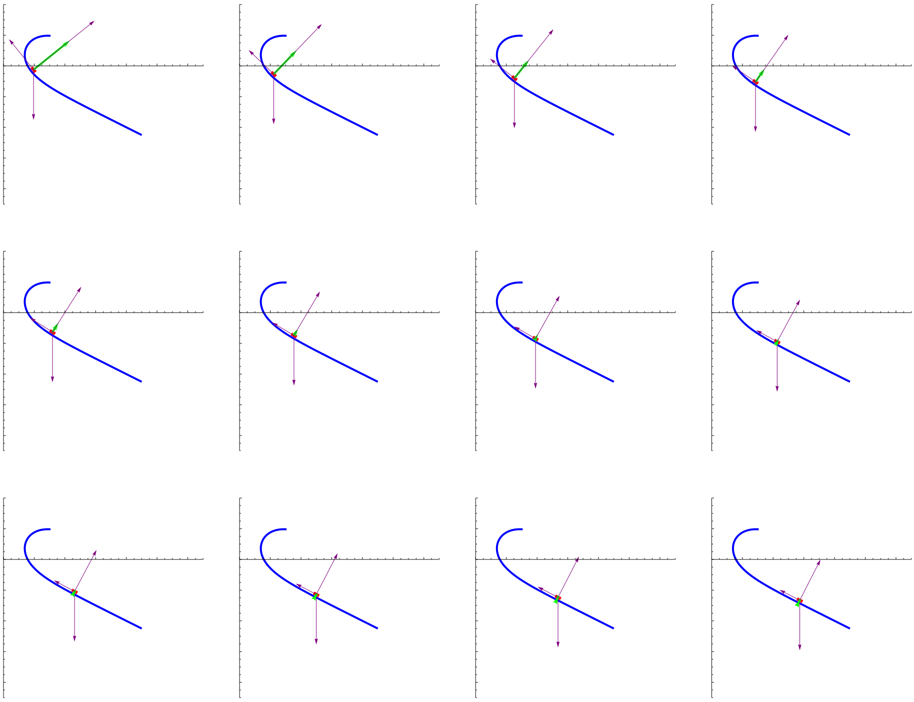
Remark 4.3. If a ramp $\boldsymbol{R}$ on earth has the property that an object will fall down with constant speed v, then, if we dilate this ramp by a factor of 6, the same block will move down with the same constant speed v on this dilated ramp when it is placed on the moon.

Appendix: Graphs

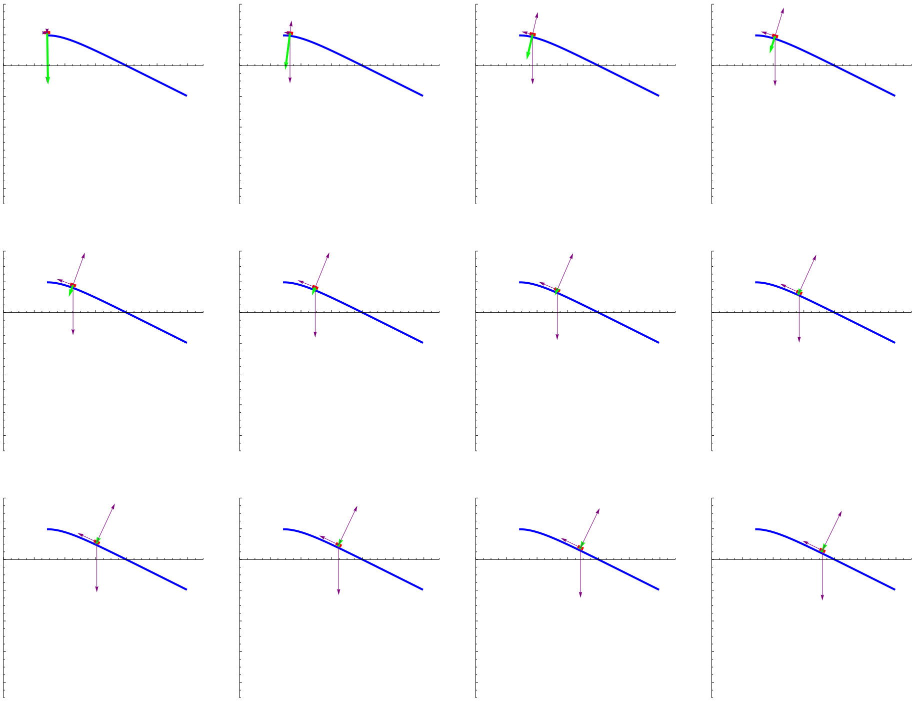
The graphs show snapshots of the motion on the longer part of the ramps, followed by snapshots of the motion on the shorter part. For a proper animation and explanatory video, see http://youtu.be/iBrvbb0efVk.

Longer Part





Shorter Part



Acknowledgements

The author would like to express his gratitude to Frank Gould, Roger Vogeler, Nidal Al-Masoud and Clifford Anderson for their valuable comments on this work. He would also like to thank the referee for several comment that helped improve this paper.

References

[Beer et al. 2010] F. P. Beer, E. R. Johnston, D. F. Mazurek, and P. J. Cornwell, *Vector mechanics for engineers: Statics and dynamics*, 9th ed., McGraw-Hill, New York, 2010.

[do Carmo 1976] M. do Carmo, *Differential geometry of curves and surfaces*, Prentice-Hall, Englewood Cliffs, NJ, 1976. MR 52 #15253 Zbl 0326.53001

[Feynman et al. 1963] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics, I: Mainly mechanics, radiation, and heat*, Addison-Wesley, Reading, MA, 1963. MR 35 #3942 Zbl 0131.38703

Received May 10, 2013. Revised October 5, 2013.

OSCAR M. PERDOMO
DEPARTMENT OF MATHEMATICS
CENTRAL CONNECTICUT STATE UNIVERSITY
NEW BRITAIN, CT 06050
UNITED STATES

perdomoosm@ccsu.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.19

SURFACES IN $\mathbb{R}^3_+$ WITH THE SAME GAUSSIAN CURVATURE INDUCED BY THE EUCLIDEAN AND HYPERBOLIC METRICS

NILTON BARROSO AND PEDRO ROITMAN

We show how to construct infinitely many immersions into the upper half-space such that the Gaussian curvatures induced from the ambient Euclidean and hyperbolic metrics coincide. We show how these immersions are related geometrically to classical minimal surfaces in Euclidean space and timelike minimal surfaces in Minkowski space.

1. Introduction

The typical scenario in problems about the geometry of submanifolds is usually given by a Riemannian manifold M and the search for a submanifold $S \subset M$ with some special geometric property with respect to the Riemannian ambient metric on M.

In the present work we will treat a problem that generalizes the above setting in the following sense. Instead of just one Riemannian metric on M, we will consider a pair of such metrics, say g_1 and g_2, and look for a submanifold $S \subset M$ with a special property that depends on both metrics g_1 and g_2. Of course, if the metrics g_1 and g_2 are arbitrary, it is hard to imagine that an interesting question will show up from this setup. However, assuming that g_1 and g_2 are in the same conformal class of metrics, we believe that there is a fertile and still unexplored field waiting for geometers. In this spirit, we will consider here an example that shows how it is possible to have a nice interplay between the geometric aspects induced by the two metrics g_1 and g_2. Specifically, we will treat the problem below.

Let S be an immersed surface in $\mathbb{R}^3_+ := \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_3 > 0\}$, and let ds_e^2 and ds_h^2 be the Euclidean and hyperbolic metrics on $\mathbb{R}^3_+$, respectively, given by

$$ds_e^2 = dx_1^2 + dx_2^2 + dx_3^2 \quad \text{and} \quad ds_h^2 = \frac{ds_e^2}{x_3^2}.$$

MSC2010: primary 53A10, 53A35; secondary 35L70, 35J96.

Keywords: minimal surfaces, Euclidean geometry, hyperbolic geometry, Gaussian curvature, Monge–Ampère equations.

We will denote by K_e and K_h the Gaussian curvatures of the metrics on S induced by ds_e^2 and ds_h^2, respectively.

Problem. Find surfaces immersed in $\mathbb{R}^3_+$ such that $K_h = K_e$.

To simplify our exposition, these surfaces will be called *isocurved* surfaces. To get a naive feeling of the problem, locally one can represent an isocurved surface as a graph $(u, v, \varphi(u, v))$, with $\varphi(u, v) > 0$ defined in a domain contained in $\partial\mathbb{R}^3_+ = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_3 = 0\}$. The function φ must then be a solution of a Monge–Ampère equation; see (2-2).

This is a mixed type equation, and — according to the type of Monge–Ampère solution associated to it — we divide isocurved surfaces into three classes: elliptic, parabolic, and hyperbolic.

As often happens in the field of differential geometry of surfaces, it is very hard to find explicit solutions of PDEs describing geometric objects and clearly some kind of geometric method is necessary to find nontrivial examples of isocurved surfaces.

Our strategy to attack the problem is based on a surprising geometric relation between isocurved surfaces and minimal surfaces in $\mathbb{R}^3$ (for the elliptic case) and timelike minimal surfaces in Minkowski space $\mathbb{L}^3$ (for the hyperbolic case). This geometric relation, together with the well-known machinery to generate minimal surfaces, provides a relatively simple and efficient method to generate isocurved surfaces.

Our geometric construction stems from the curious fact that isocurved surfaces appear in a one parameter family of parallel surfaces in the sense of hyperbolic geometry. It is therefore natural to seek a method that enables one to consider this one parameter family of isocurved surfaces as a whole. This is achieved by looking at the congruence of geodesics of hyperbolic geometry that have isocurved surfaces as orthogonal surfaces.

We will present a technique to construct such congruences of geodesics by starting with an appropriate simply connected minimal surface. This technique yields an explicit constructive method to generate isocurved surfaces that are either elliptic or hyperbolic from minimal surfaces in $\mathbb{R}^3$ or $\mathbb{L}^3$.

We have organized this paper as follows. In Section 2 we present the simplest examples of isocurved surfaces and the PDE for isocurved graphs. We also show that isocurved surfaces appear in a one parameter family of parallel surfaces with respect to ds_h^2. In Section 3 we show how elliptic and hyperbolic isocurved surfaces can be constructed from minimal surfaces. We offer a detailed discussion about the elliptic case in Section 3A, and apply the theory to construct many examples in Section 3B. In Sections 3C and 3D we deal with the case of hyperbolic isocurved surfaces and give some examples. Section 4 is devoted to some final remarks about isocurved surfaces.

2. Basics about isocurved surfaces

2A. ***Simple examples.*** Throughout this work we will use (x_1, x_2, x_3) as coordinates in $\mathbb{R}^3$. We will also adopt the following terminology. A geodesic of hyperbolic geometry will be called an h-geodesic. We recall that in the upper half-space model $(\mathbb{R}^3_+)$ these geodesics are represented either by circles orthogonal to $\partial\mathbb{R}^3_+$ or by vertical lines. Accordingly, the parallel surfaces in hyperbolic geometry will be called h-parallel surfaces.

We now start our study of isocurved surfaces with a description of the simplest examples. The trivial example is a horizontal plane $x_3 = \text{const.} > 0$ (horosphere in hyperbolic geometry), since $K_e = K_h = 0$ for any such plane. In fact, any surface that is flat with respect to the Euclidean and hyperbolic metrics is isocurved. For instance, if we consider the part of vertical right circular cone with vertex at $\partial\mathbb{R}^3_+$ that lies in the region $x_3 > 0$, it is well-known that $K_e = K_h = 0$ for this surface. A not so obvious example of this type is provided by the surface with horizontal rulings orthogonal to a tractrix contained in a vertical plane and asymptotic to $\partial\mathbb{R}^3_+$; see Figure 1.

An example that is not flat is simply a (round) sphere properly placed in $\mathbb{R}^3_+$. To see that this is true, consider any sphere in $\mathbb{R}^3_+$. Of course, K_e is invariant with respect to vertical Euclidean translations, but such motion clearly changes K_h. Note that for a given sphere with fixed Euclidean radius, if we vertical translate upwards, then K_h increases without limit as we go up and if we translate downwards then K_h tends to zero as we approach $\partial\mathbb{R}^3_+$. So, by continuity, there is a specific placement of the sphere such that $K_e = K_h$.

A particular feature of all these examples is the following: if S is one of the examples cited above, then the h-parallel surfaces are also examples.

This is clear for the horizontal planes and the circular cones, and a simple computation shows that it is also true for spheres and the ruled example. In fact, as we shall see in Theorem 2.2, this property is valid for any isocurved surface.

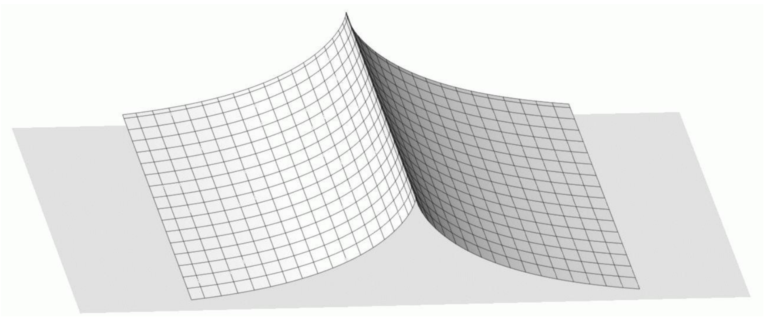


Figure 1. Flat ruled surface generated by a tractrix.

2B. ***The PDE for isocurved graphs.*** We will now present a PDE that is related to isocurved surfaces. The straightforward approach we adopt here is to consider an isocurved surface that is the graph of some function defined in a domain in $\partial\mathbb{R}^3_+$ and study the corresponding PDE that has φ as a solution.

To derive this PDE, we start by recalling that K_h and K_e are related by

$$K_h = x_3^2 K_e + 2H_e x_3 n_3 + n_3^2 - 1, \tag{2-1}$$

where x_3 is the third coordinate in $\mathbb{R}^3_+$, n_3 is the third coordinate of a unit normal vector in the Euclidean sense, and H_e is the mean curvature with respect to the Euclidean metric. Equation (2-1) follows easily from the well-known expression relating the principal curvatures of an immersed surface in $\mathbb{R}^3_+$ with respect to ds_e^2 and ds_h^2 [López 2001] and the Gauss equation

$$K_{\text{ext}} = K_h + 1,$$

relating K_h to the extrinsic curvature K_{ext} of a surface induced by the ambient metric ds_h^2.

Using the well-known expressions for K_e, H_e, n_3 for a graph $(u, v, \varphi(u, v))$ over a domain in the boundary of $\mathbb{R}^3_+$, we conclude that φ is a solution of the PDE:

$$(1-\varphi^2)\frac{\det\nabla^2\varphi}{1+|\nabla\varphi|^2} - \varphi\frac{(1+\varphi_v^2)\varphi_{uu} - 2\varphi_u\varphi_v\varphi_{uv} + (1+\varphi_u^2)\varphi_{vv}}{1+|\nabla\varphi|^2} + |\nabla\varphi|^2 = 0, \tag{2-2}$$

where $\det\nabla^2\varphi = \varphi_{uu}\varphi_{vv} - \varphi_{uv}^2$ and $|\nabla\varphi|^2 = \varphi_u^2 + \varphi_v^2$.

Equation (2-2) is a Monge–Ampère mixed type PDE. It turns out that the analytical classification of such equations can be interpreted geometrically, and the method to construct isocurved surfaces depends on the type of the equation. It is then appropriate to recall the classification of the solutions of this class of PDEs. A nice discussion about Monge–Ampère equations from a geometric point of view can be found in [Ivey and Landsberg 2003].

Consider a PDE of the form

$$A(\varphi_{uu}\varphi_{vv} - \varphi_{uv}) + B\varphi_{uu} + 2C\varphi_{uv} + D\varphi_{vv} + E = 0, \tag{2-3}$$

where A, B, C, D, and E are functions of u, v, φ, φ_u, and φ_v, and define the quantity $\Delta = AE - BD + C^2$. Equation (2-3) (or its solutions for some authors) is called *elliptic* if $\Delta < 0$, *hyperbolic* if $\Delta > 0$, and *parabolic* if $\Delta = 0$.

The proposition below shows that the criteria for deciding whether (2-1) is elliptic, parabolic, or hyperbolic admits a geometric interpretation.

Proposition 2.1. *Let S be an isocurved surface that is the graph of a function $\varphi(u, v)$ defined in a domain of the plane $x_3 = 0$. For points $(u, v, \varphi(u, v))$ such that the normal vector to S is not vertical, i.e., $|\nabla\varphi| \neq 0$, let $\rho(u, v)$ be the Euclidean radius of the circle orthogonal to $x_3 = 0$ that represents the geodesic in hyperbolic* *geometry passing through* $(u, v, \varphi(u, v))$ *and orthogonal to* S *at this point. Then the solution* φ *of* (2-2) *is*

$$\begin{aligned} \textit{elliptic} &\iff \textit{either } \rho > 1 \textit{ or } |\nabla\varphi| = 0,\\ \textit{parabolic} &\iff \rho = 0,\\ \textit{hyperbolic} &\iff \rho < 1. \end{aligned}$$

Proof. Writing (2-2) in the form (2-3) yields $A = (1 - \varphi^2)$, $B = -\varphi(1 + \varphi_v)$, $C = \varphi\varphi_v\varphi_u$, $D = -\varphi(1+\varphi_u)$, and $E = (1+\varphi_u^2+\varphi_v^2)(\varphi_u^2+\varphi_v^2)$. Direct computation shows that

$$\Delta = -\big(1 + |\nabla\varphi|^2\big)\big(\varphi^2(1 + |\nabla\varphi|^2) - |\nabla\varphi|^2\big).$$

Thus φ is elliptic at points where $|\nabla\varphi| = 0$. For points such that $|\nabla\varphi| \neq 0$, an elementary computation shows that ρ is given by

$$\rho = \frac{\varphi\sqrt{1 + |\nabla\varphi|^2}}{|\nabla\varphi|}.$$

To finish the proof just use the above expressions for ρ and Δ. □

2C. ***An invariance property of isocurved surfaces.*** Even though the Euclidean and hyperbolic metrics enter on equal footing in the definition of isocurved surfaces, the following property suggests that these surfaces might have some alternative geometric description only in terms of hyperbolic geometry.

Theorem 2.2. *Let* S *be an isocurved surface and* S^t *be the h-parallel surface at distance* t. *If* S^t *is smooth, then it is also an isocurved surface.*

Proof. The equation that defines an isocurved surface can be rewritten as

$$K_{\text{ext}}(1 - x_3^2) - 2H_h n_3 + n_3^2 + x_3^2 = 0, \tag{2-4}$$

where K_{ext} and H_h denote the extrinsic and mean curvatures with respect to the hyperbolic metric, respectively. So, we have to prove that if (2-4) holds for S, then

$$K^t_{\text{ext}}(1 - (x^t_3)^2) - 2H^t_h n^t_3 + (n^t_3)^2 + (x^t_3)^2 = 0, \tag{2-5}$$

holds for S^t, where the quantities with upper index t are relative to S^t. To do this, it is convenient to start by writing S and S^t in the hyperboloid model to obtain manageable expressions in the upper half-space model.

Let $P = (P_1, P_2, P_3, P_4)$ and $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)$ be the vector position and unit normal field of S in the hyperboloid model of $\mathbb{H}^3$. Then the parallel surface S^t in this model has vector position and unit normal field given by

$$P^t = \cosh tP + \sinh t\eta, \quad \eta^t = \sinh tP + \cosh t\eta.$$

Now, consider the map

$$\Phi(u_1, u_2, u_3, u_4) = \Big(\frac{u_2}{u_1 - u_4}, \frac{u_3}{u_1 - u_4}, \frac{1}{u_1 - u_4}\Big),$$

that maps the upper sheet of the hyperboloid $-u_1^2 + u_2^2 + u_3^2 + u_4^4 = -1$ onto the upper half-space. After some computations we obtain the following expressions for the third coordinates of the position and Euclidean unit normal vector of S and S^t:

$$x_3 = \frac{1}{P_1 - P_4}, \qquad n_3 = -\frac{\eta_1 - \eta_4}{P_1 - P_4}, \tag{2-6}$$

$$x_3^t = \frac{1}{c(P_1 - P_4) + s(\eta_1 - \eta_4)}, \qquad n_3^t = -\frac{s(P_1 - P_4) + c(\eta_1 - \eta_4)}{c(P_1 - P_4) + s(\eta_1 - \eta_4)}, \tag{2-7}$$

where $c = \cosh t$ and $s = \sinh t$.

We also need expressions for K_{ext}^t and H_h^t as functions of K_{ext} and H_h:

$$K_{\text{ext}}^t = \frac{c^2 K_{\text{ext}} - 2H_h sc + s^2}{s^2 K_{\text{ext}} - 2H_h sc + c^2}, \tag{2-8}$$

$$H_h^t = \frac{(c^2 + s^2)H_h - sc(K_{\text{ext}} + 1)}{s^2 K_{\text{ext}} - 2H_h sc + c^2}. \tag{2-9}$$

See (2-8) and (2-9) for more details about [Tenenblat 1998, Proposition 3.2, page 24].

Now, using (2-6), Equation (2-4) is equivalent to

$$K_{\text{ext}}((P_1 - P_4)^2 - 1) + 2H_h(P_1 - P_4)(\eta_1 - \eta_4) + (\eta_1 - \eta_4)^2 + 1 = 0. \tag{2-10}$$

Using (2-7), (2-8), and (2-9), and performing some computations, we conclude that the left hand side of (2-5) vanishes if and only if

$$\begin{aligned} K_{\text{ext}}\big[(P_1 - P_4)^2(c^2 - s^2)^2 - c^2 + s^2\big] \\ + 2H_h(P_1 - P_4)(\eta_1 - \eta_4)(c^2 - s^2)^2 \\ + (c^2 - s^2)\big[(c^2 - s^2)(\eta_1 - \eta_4)^2 + 1\big] = 0. \end{aligned}$$

Using the fact that $c^2 - s^2 = 1$, we see that the expression above coincides with (2-10) and so S^t is isocurved. □

3. Isocurved surfaces from minimal surfaces

We are now in position to present a geometric method that allows us to construct infinitely many nontrivial examples of isocurved surfaces. Actually, our method of construction works only for elliptic and hyperbolic isocurved surfaces and we don't have a general method to construct examples of parabolic isocurved surfaces.

The basic ingredient of our method is a simply connected minimal surface in $\mathbb{R}^3$ (for elliptic isocurved surfaces) or simply connected timelike minimal surface in $\mathbb{L}^3$ (for hyperbolic isocurved surfaces). For both cases, starting with a minimal surface, we will show how to construct a congruence of geodesics of hyperbolic space (h-geodesics) that has isocurved surfaces as orthogonal surfaces. Since there are slight variations between the two cases, we will discuss them separately.

3A. ***Elliptic isocurved surfaces.*** We first introduce a process to induce from a given simply connected immersed oriented surface Σ in $\mathbb{R}^3$ a congruence of h-geodesics C_Σ whose elements are represented either by circles orthogonal to the plane $x_3 = 0$ or by vertical lines in $\mathbb{R}^3_+$.

A circle orthogonal to the plane $x_3 = 0$ is determined by its (Euclidean) center σ that lies in the plane, its radius R, and a horizontal unit vector e_1 in the Euclidean sense that together with $e_3 = (0, 0, 1)$ defines the vertical plane where the circle lies. The case of vertical lines can be treated as a limiting case.

For simplicity, we will consider only the h-geodesics that are circles. In other words, from now on we will assume that for every $p \in \Sigma$, the tangent plane $T_p\Sigma$ is not horizontal. From the surface Σ with unit normal vector field N, we will define the congruence of h-geodesics as follows.

Let P_{hor} be the orthogonal projection onto the horizontal plane $x_3 = 0$, let $J : \mathbb{R}^2 \to \mathbb{R}^2$ be the $\frac{\pi}{2}$ counterclockwise rotation in this plane, and let $N = (n_1, n_2, n_3)$ be a unit vector field normal to Σ.

For $p \in \Sigma$ we define the center of the circle $\sigma(p)$ as

$$\sigma(p) = P_{\mathrm{hor}}(p), \tag{3-1}$$

the direction $e_1(p)$ as

$$e_1(p) = \frac{J\bigl(P_{\mathrm{hor}}(N(p))\bigr)}{\bigl|P_{\mathrm{hor}}(N(p))\bigr|}, \tag{3-2}$$

and the radius $R(p)$ as

$$R(p) = \frac{1}{\bigl|P_{\mathrm{hor}}(N(p))\bigr|}. \tag{3-3}$$

We now pose the question: what is the condition on the surface Σ such that the congruence of h-geodesics C_Σ admits orthogonal surfaces?

If we knew in advance that C_Σ defines a distribution of planes (the planes of the distribution being the ones orthogonal to the h-geodesics), then we could use the geometric version of the Frobenius theorem to verify the condition for the existence of integral surfaces to this distribution.

However, even though the statements that follow become a bit clumsy, it is interesting to consider the general situation where C_Σ does not necessarily define a distribution of planes. As we shall see, this procedure allows the consideration of self-intersections and singularities for isocurved surfaces in a natural way.

So we ask a weaker question, namely, about the existence of a differentiable map $Y : \Sigma \to \mathbb{R}^3_+$ defined by

$$Y = \sigma + R(\cos\theta\, e_1 + \sin\theta\, e_3), \tag{3-4}$$

where σ, e_1, and R are defined by (3-1), (3-2), and (3-3), respectively, and $\theta : \Sigma \to \mathbb{R}$ is an unknown differentiable function such that, for any point $p \in \Sigma$ where Y is an immersion, the tangent plane of Y at $Y(p)$ is orthogonal to the geodesic of C_Σ associated to p. If Y satisfies the condition above we say that Y has the *orthogonal property.*

The geometric condition above can be rephrased in terms of a system of Frobenius type for the unknown function θ. The next proposition shows that, up to a degenerate case, the function θ exists if and only if Σ is a minimal surface.

Theorem 3.1. *Let Σ be an oriented simply connected surface in the upper half-plane such that $T_p\Sigma$ is not horizontal for every $p \in \Sigma$. The function θ that appears in the expression of the map Y given by* (3-4) *and such that Y has the orthogonal property exists if and only if Σ is either a minimal surface or vertical cylinder over a planar curve in the plane $\partial\mathbb{R}^3_+$.*

Proof. First suppose that Σ is the graph of a function ψ defined in a domain $\Omega \subset \partial\mathbb{R}^3_+$. A local chart for Σ is given by

$$X(u, v) = (u, v, \psi(u, v)).$$

The function θ can be written in these local coordinates as a function $\theta : \Omega \to \mathbb{R}$. Since we are assuming that Y has the orthogonal property, we have

$$\langle dY, -\sin\theta e_1 + \cos\theta e_3\rangle = 0,$$

where $\langle\,\cdot\,,\,\cdot\,\rangle$ is the usual inner product of Euclidean space.

Writing $\eta = -\sin\theta\, e_1 + \cos\theta\, e_3$, we find

$$\begin{aligned}
0 &= \langle Y_u, \eta\rangle \\
&= \langle \sigma_u + R_u(\cos\theta\, e_1 + \sin\theta\, e_3) + R(\cos\theta\, e_1 + \sin\theta\, e_3)_u, \eta\rangle \\
&= \langle \sigma_u + (R_u\cos\theta - R\theta_u\sin\theta)e_1 + R\cos\theta(e_1)_u + R\theta_u\cos\theta e_3, \eta\rangle \\
&= R\theta_u - \sin\theta\langle\sigma_u, e_1\rangle,
\end{aligned}$$

and in the same fashion,

$$0 = \langle Y_v, \eta\rangle = R\theta_v - \sin\theta\langle\sigma_v, e_1\rangle.$$

We conclude that

$$\theta_u = \frac{\sin\theta}{R}\langle\sigma_u, e_1\rangle, \quad \theta_v = \frac{\sin\theta}{R}\langle\sigma_v, e_1\rangle. \tag{3-5}$$

The change of variables $\sin\theta = 1/\cosh\beta$ and $\cos\theta = \tanh\beta$, transforms (3-5) into

$$\beta_u = -\frac{\langle \sigma_u, e_1\rangle}{R}, \quad \beta_v = -\frac{\langle \sigma_v, e_1\rangle}{R}, \tag{3-6}$$

and the Frobenius condition for the integrability of the system is

$$\left(\frac{\langle \sigma_u, e_1\rangle}{R}\right)_v = \left(\frac{\langle \sigma_v, e_1\rangle}{R}\right)_u. \tag{3-7}$$

Since

$$\sigma(u, v) = (u, v, 0), \tag{3-8}$$

$$R = \frac{\sqrt{1 + |\nabla\psi|^2}}{|\nabla\psi|}, \tag{3-9}$$

$$e_1 = \frac{(-\psi_v, \psi_u)}{|\nabla\psi|}, \tag{3-10}$$

Equation (3-7) implies that

$$(1 + \psi_v^2)\psi_{uu} - 2\psi_u\psi_v\psi_{uv} + (1 + \psi_v^2)\psi_{uu} = 0,$$

that is, $\theta(u, v)$ exists if and only if Σ is a minimal surface.

Given a general parametrization for Σ, a long but straightforward computation shows that the integrability condition for θ is always satisfied for points where the normal vector N is horizontal, and — as we have seen — if N is not horizontal, then the mean curvature must vanish. So there two possibilities for Σ: either it is a minimal surface or it is a vertical cylinder over a plane curve in $\partial\mathbb{R}^3_+$. □

Remark 3.2. If Σ in our construction is a vertical cylinder over a planar curve, then it is easy to check that map Y is not an immersion anywhere. On the other hand, as we shall see later on, if Σ is a minimal surface, we cannot guarantee in general that the Y is an immersion everywhere.

From Theorem 3.1, we may start from a minimal surface Σ and construct C_Σ that admits orthogonal surfaces. But it is not clear if these orthogonal surfaces have any geometric property. We will now show that, if the map Y associated to C_Σ is an immersion, then $Y(\Sigma)$ is an isocurved surface.

To prove this, we will start with an arbitrary immersed oriented surface S in $\mathbb{R}^3_+$ and find the conditions on S such that S is orthogonal to the h-geodesics of C_Σ induced by some surface Σ via our geometric method. This leads us to our next result.

Theorem 3.3. *Let S be an immersed surface in $\mathbb{R}^3_+$. If S is orthogonal to the h-geodesics of the congruence C_Σ induced by an oriented immersed surface Σ, then S is an isocurved surface.*

Proof. Let $p \in S$. Without loss of generality, we can assume that T_pS is not vertical. In fact, if we had T_pS vertical, then we could replace S with S_t, i.e., the parallel surface to S at distance t with respect to the hyperbolic metric. For t small enough, S_t is also an immersed surface and the tangent plane $T_{p_t}S_t$ is not vertical. For the following argument, we will also assume that T_pS is not horizontal, and we will treat this situation as a limiting case at the end of the proof.

In a neighborhood of p, the surface S is the graph of a function φ defined in a domain Ω in the plane $x_3 = 0$. We will use u and v as coordinates and write this graph as

$$Y(u, v) = (u, v, \varphi(u, v)).$$

Now we reverse the steps in our geometric construction that induces a congruence of h-geodesics from a given oriented surface. In other words, we will compute the center σ, the radius R, and the direction e_1 for the family of h-geodesics that are orthogonal to S.

The direction e_1 is the projection of the normal vector of S in the plane $x_3 = 0$:

$$e_1 = \frac{(-\varphi_u, -\varphi_v)}{|\nabla\varphi|}.$$

To find the center σ we proceed as follows. The normal field to S is given by

$$\eta = (-\varphi_u, -\varphi_v, 1) = |\nabla\varphi| e_1 + e_3.$$

Let $\tilde{J} : \mathbb{R}^2 \to \mathbb{R}^2$ be the counterclockwise rotation of $\frac{\pi}{2}$ radians in the plane spanned by e_1 and e_3. Consider the line starting from $(u, v, \varphi(u, v))$ in the direction of $\tilde{J}\eta$, parametrized by $\ell(t) = Y + t\tilde{J}\eta$. This line intersects $\partial\mathbb{R}^3_+$ at σ, so we obtain

$$\sigma(u, v) = \left(u - \frac{\varphi\varphi_u}{|\nabla\varphi|^2}, v - \frac{\varphi\varphi_v}{|\nabla\varphi|^2}, 0\right). \tag{3-11}$$

Finally, the radius R is the Euclidean distance from σ to $Y(u, v)$, that is,

$$R = \frac{\varphi\sqrt{1 + |\nabla\varphi|^2}}{|\nabla\varphi|}. \tag{3-12}$$

If S is orthogonal to the h-geodesics of C_Σ, then there is a differentiable function ψ defined in Ω such that Σ is locally parametrized by the map X given by

$$X = \sigma + \psi e_3. \tag{3-13}$$

Let N be the normal field of Σ and θ be the angle between N and the horizontal plane. From our geometric construction, we have that the projection of N into the horizontal plane is $-Je_1$ and $R = (\cos\theta)^{-1}$. Thus,

$$N = \cos\theta(-Je_1) + \sin\theta e_3 = -\frac{1}{R}Je_1 + \frac{\sqrt{R^2 - 1}}{R}e_3.$$

The orthogonality condition $\langle dX, N\rangle = 0$ yields the following system for ψ:

$$\psi_u = \frac{\langle Je_1, \sigma_u\rangle}{\sqrt{R^2-1}}, \quad \psi_v = \frac{\langle Je_1, \sigma_v\rangle}{\sqrt{R^2-1}}. \tag{3-14}$$

The Frobenius integrability condition for this system is given by

$$\left(\frac{\langle Je_1, \sigma_u\rangle}{\sqrt{R^2-1}}\right)_v - \left(\frac{\langle Je_1, \sigma_v\rangle}{\sqrt{R^2-1}}\right)_u = 0. \tag{3-15}$$

Using equations (3-2), (3-11), (3-12), and performing some computations, we obtain

$$\frac{\langle Je_1, \sigma_u\rangle}{\sqrt{R^2-1}} = -\frac{\varphi_v\varphi_u^2 + \varphi\varphi_u\varphi_{uv} - \varphi\varphi_v\varphi_{uu} + \varphi_v^3}{\sqrt{\varphi^2(1+|\nabla\varphi|^2) - |\nabla\varphi|^2|\nabla\varphi|^2}}, \tag{3-16}$$

$$\frac{\langle Je_1, \sigma_v\rangle}{\sqrt{R^2-1}} = \frac{\varphi_u\varphi_v^2 + \varphi\varphi_v\varphi_{uv} - \varphi\varphi_u\varphi_{vv} + \varphi_u^3}{\sqrt{\varphi^2(1+|\nabla\varphi|^2) - |\nabla\varphi|^2|\nabla\varphi|^2}}. \tag{3-17}$$

After the substitution of the expressions given by (3-16) and (3-17) into (3-15) and some straightforward computations, we see that (3-15) is in fact equivalent to (2-2). So, the function ψ exists if and only if S is an isocurved elliptic surface.

To finish the proof, recall that we have assumed that $p \in S$ was such that T_pS was not horizontal. Suppose now that T_pS is horizontal, then either there is a neighborhood U of p in S such that the tangent plane is horizontal for any point of U, or there is a sequence of points $\{p_n\}$ of S converging to p such that $T_{p_n}S$ is not horizontal. For the first situation U is part of a horizontal plane, so it is an isocurved surface. For the second case, since the isocurved condition is satisfied for all the points in $\{p_n\}$, by continuity, it must also be satisfied at p, so S is an isocurved surface. □

Remark 3.4. The proof of Theorem 3.3 shows that if we start with an elliptic isocurved surface, then the map (3-13) is well defined but it is not necessarily an immersion. In the case where X is an immersion, it is also minimal due to Theorem 3.1.

A careful analysis shows that if we define

$$g_{11} = \langle X_u, X_u\rangle, \quad g_{12} = \langle X_u, X_v\rangle, \quad g_{22} = \langle X_v, X_v\rangle,$$
$$h_{11} = \langle X_{uu}, N\rangle, \quad h_{12} = \langle X_{uv}, N\rangle, \quad h_{22} = \langle X_{vv}, N\rangle,$$

then

$$g_{11}h_{22} - 2g_{12}h_{12} + g_{22}h_{11} = 0.$$

In other words, if we start with an elliptic isocurved surface S, we can always locally associate to S a *generalized* minimal surface in the following sense: we have the two maps X and N satisfying the equation above, but not necessarily $g_{11}g_{22} - g_{12}g_{21} > 0$. We will exhibit an example of this situation in the next subsection.

3B. ***Examples of elliptic isocurved surfaces.*** We now wish to apply the relation between minimal and isocurved surfaces to generate explicit examples of isocurved surfaces. Some apparent difficulties arise if we stick to the local analysis using graphs that we have adopted until now. Using graphs, we would have to start with an explicit minimal graph and then have the trouble to integrate the system (3-6). Since there aren't many explicit minimal graphs, our method would not be very useful to generate new examples.

Fortunately, the function β that appears in (3-6) has a nice geometric interpretation and this frees us from the graph representation. In fact, if we consider a simply connected minimal surface Σ, we can associate to it the so-called conjugate surface Σ^*. It turns out that β is the height function (i.e., x_3 coordinate) of Σ^*. This can be seen quite easily by using the expressions (3-8), (3-9), and (3-10) to rewrite (3-6) as follows:

$$\beta_u = \frac{\varphi_v}{\sqrt{1+|\nabla\varphi|^2}}, \quad \beta_v = -\frac{\varphi_u}{\sqrt{1+|\nabla\varphi|^2}}. \tag{3-18}$$

It is well-known that the system above coincides with the one for the conjugate height function; see, for instance, [Mazet et al. 2007].

Now that we don't need graphs anymore, we will generate examples with minimal surfaces constructed using the classic Weierstrass representation. For the examples below, let (f, g) be the Weierstrass data for the starting simply connected minimal surface Σ. The conjugate surface Σ^* is defined by the data (if, g).

Example (Rotational isocurved surfaces). Let $z \in \mathbb{C}$, $z = x + iy$ and consider $f(z) = i\mathrm{e}^z$ and $g(z) = c\mathrm{e}^{-z}$, where $c \in \mathbb{R}$, $c \neq 0$. The corresponding minimal surface is a helicoid, and the associated isocurved surface is a surface of revolution (with respect to the x_3-axis). For instance, choosing $c = 2$ yields the following parametrized surface.

$$X(x, y) = (\alpha(x)\sin y, \alpha(x)\cos y, \gamma(x)),$$

where

$$\alpha(x) = \frac{1}{4}\frac{(-\mathrm{e}^{5x} + 8\mathrm{e}^{3x} + 45\mathrm{e}^{x} - 8\mathrm{e}^{-x} + 16\mathrm{e}^{-3x})\sin y}{(4\mathrm{e}^{-2x}+1)(\mathrm{e}^{4x}+1)},$$

and

$$\gamma(x) = \frac{1}{2}\frac{\mathrm{e}^{3x}\sqrt{1+8\mathrm{e}^{-2x}+16\mathrm{e}^{-4x}}}{\mathrm{e}^{4x}+1}.$$

This surface is seen in Figure 2. Note that even though we have started with a smooth minimal surface, the corresponding isocurved surfaces has singularities.

Example (The isocurved surface associated to a point). If we choose $f(z) = 0$ and $g(z) = z$ and use the Weierstrass representation we don't get a minimal surface; the result is just the point $(0, 0, 0)$ together with the Gauss map defined by $g(z)$.

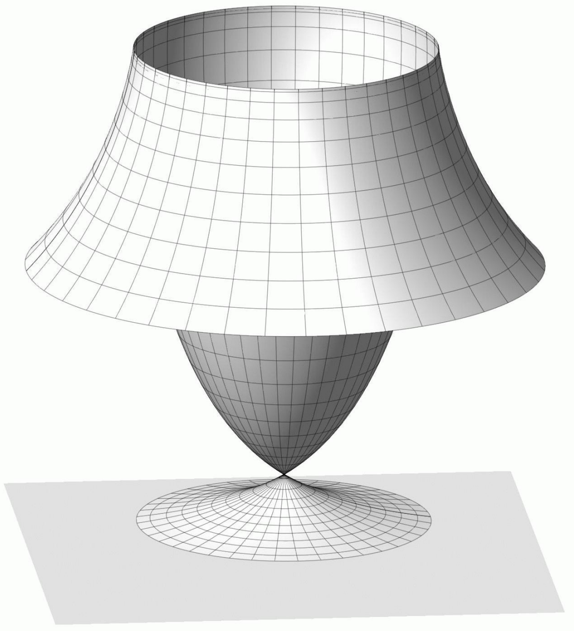


Figure 2. An isocurved surface of revolution.

However, this data, a surface in the space of contact elements of $\mathbb{R}^3$, is enough to use our machinery. For this simple example the associated isocurved surface is the elliptic region of the vertical circular right cone parametrized as

$$X(r,t)=\left(\frac{\tanh k(1+r^2)}{2r}\sin t,\, -\frac{\tanh k(1+r^2)}{2r}\cos t,\, \frac{1+r^2}{2r\cosh k}\right),$$

where $z=r\,\mathrm{e}^{it}$ and $k\in\mathbb{R}$ is a parameter used to describe the family of parallel surfaces in hyperbolic geometry.

Example (A 1-periodic isocurved surface). With $f(z)=z$ and $g(z)=1/z$, the associated isocurved surface is parametrized by

$$X(r,t)=X_1(r,t)+X_2(r,t),$$

where

$$X_1(r,t)=\frac{(r^2+1)(\mathrm{e}^{2r\sin t}-1)}{2r(\mathrm{e}^{2r\sin t}+1)}(\sin t,\cos t,0),$$

and

$$X_2(r,t)=\left(-\frac{2\ln r+r^2(1-2(\cos t)^2)}{4},\, -\frac{t+r^2\sin t\cos t}{2},\, \frac{(r^2+1)\mathrm{e}^{r\sin t}}{r(\mathrm{e}^{2r\sin t}+1)}\right).$$

This example is invariant under Euclidean translations in the x_2 direction by a multiple of π. A view of part of this surface is shown in Figure 3.

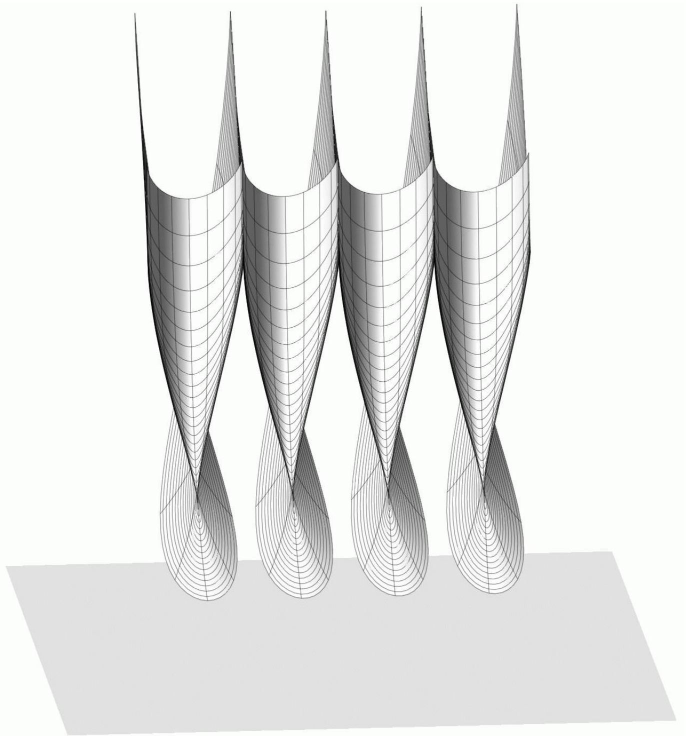


Figure 3. A piece of a 1-periodic isocurved surface.

Example (A Scherk-type isocurved surface). We end our list of examples showing the isocurved surface obtained from Scherk's minimal surface that can be written as the graph of the function

$$\varphi(x, y) = \ln \frac{\cos y}{\cos x}.$$

It is known — see [Nitsche 1989] — that the conjugate function in this case is

$$\varphi^*(x, y) = \arcsin(\sin x \, \sin y).$$

Using our geometric method we obtain the map

$$X(x, y) = (x - \Lambda_1 \sin y \cos x, \, y - \Lambda_1 \sin x \cos y, \, \Lambda_2),$$

where

$$\Lambda_1 = \frac{\sqrt{\cos^2 x + \cos^2 y - \cos^2 x \cos^2 y} \, \tanh(\arcsin(\sin x \, \sin y))}{\sin^2 x \cos^2 y + \sin^2 y \cos^2 x},$$

$$\Lambda_2 = \frac{\sqrt{\cos^2 x + \cos^2 y - \cos^2 x \cos^2 y}}{\cosh(\arcsin(\sin x \, \sin y)) \sqrt{\cos^2 x + \cos^2 y - 2\cos^2 x \cos^2 y}}.$$

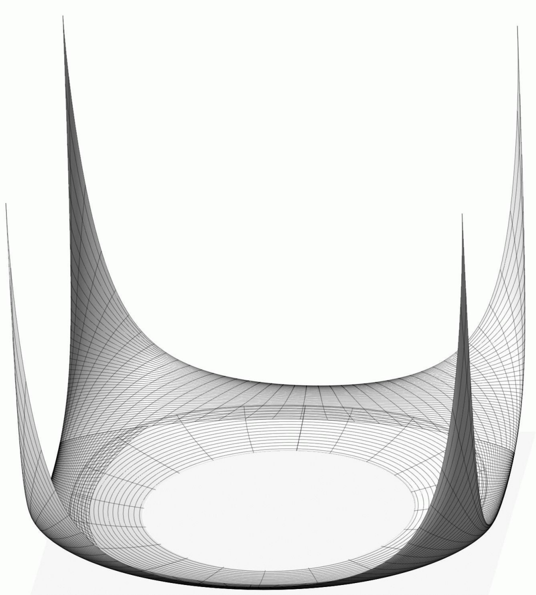


Figure 4. A 2-periodic isocurved surface obtained from Scherk's surface.

Extending this fundamental domain to the whole plane minus the lattice

$$\left\{\left(\tfrac{\pi}{2}+m\pi,\ \tfrac{\pi}{2}+n\pi\right)\in\mathbb{R}^2 \mid m,n\in\mathbb{Z}\right\}$$

in the obvious way, we obtain a 2-periodic surface, with singular curves that project into lines in the $x_3 = 0$ plane having the form $\{(m\pi/2, s) \mid m \in \mathbb{Z}, s \in \mathbb{R}\}$ or $\{(s, m\pi/2) \mid m \in \mathbb{Z}, s \in \mathbb{R}\}$ and, surprising, circular holes. Part of this surface is depicted in Figure 4.

3C. ***Hyperbolic isocurved surfaces.*** A slight variation of the geometric method used to construct elliptic isocurved surfaces from a minimal surface in $\mathbb{R}^3$, allows us to construct hyperbolic isocurved surfaces from timelike minimal surfaces in Minkowski space $\mathbb{L}^3$, that is, $\mathbb{R}^3$ with the Lorentzian metric $ds_L^2 = dx_1^2 + dx_2^2 - dx_3^2$.

For a given oriented timelike surface Σ immersed in $\mathbb{L}^3$ with unit normal N (in the Minkowski metric), we induce a congruence of h-geodesics, C_Σ in the following way. We define the center σ and the vertical plane containing the circle exactly as we did before. The only difference is the choice of the radius. Again, we consider the radius as the inverse of the size of the horizontal projection of N. But since now we are using the Minkowski metric, the size of this projection is bigger or equal to one. In this way, our radius function R will be smaller or equal to one.

Since the ideas and techniques are the same as the ones in Section 3A, we will limit ourselves to state the results without giving detailed proofs.

As before, we will search for a differentiable map $Y : \Sigma \to \mathbb{R}^3_+$ defined by

$$Y = \sigma + R(\cos\theta\, e_1 + \sin\theta\, e_3), \tag{3-19}$$

where σ, e_1, and R are defined, respectively, by (3-1), (3-2), and (3-3), where N is now a unit (spacelike) normal field to Σ with respect to the Minkowski metric, and $\theta : \Sigma \to \mathbb{R}$ is, as before, an unknown differentiable function such that, for any point $p \in \Sigma$ where Y is an immersion, the tangent plane of Y at $Y(p)$ is orthogonal to the h-geodesic of C_Σ associated to p. If Y satisfies the condition above, then we say that Y has the *orthogonal property*.

Our next theorem is analogous to Theorem 3.1.

Theorem 3.5. *Let Σ be an oriented surface in $\mathbb{L}^3$ with unit normal vector field N and such that $T_p\Sigma$ is not horizontal for every $p \in \Sigma$. The function θ that appears in the expression of the map Y given by* (3-19) *and such that Y has the orthogonal property exists if and only if Σ is either a minimal timelike surface or a vertical cylinder over a planar curve in the plane $x_3 = 0$.*

Proof. The proof follows exactly the same lines of the proof given for Theorem 3.1. The only difference is that if we write Σ locally as the graph of a function ψ, then the radius function is given by

$$R = \frac{\sqrt{|\nabla\psi|^2 - 1}}{|\nabla\psi|^2}.$$

The integrability condition associated to the existence of θ now becomes

$$(1 + \psi_v^2)\psi_{uu} - 2\psi_u\psi_v\psi_{uv} + (1 + \psi_v^2)\psi_{uu} = 0,$$

and this is the PDE associated to a minimal surface in $\mathbb{L}^3$. □

Now we state the theorem analogous to Theorem 3.3.

Theorem 3.6. *Let S be an immersed surface in $\mathbb{L}^3_+$. If S is orthogonal to the h-geodesics of the congruence C_Σ induced by an oriented immersed surface Σ, then S is an isocurved surface.*

Proof. The proof follows the lines of the proof of Theorem 3.3. We only note that in the present case, the relation between the radius function R and the (Euclidean) angle formed by N and the horizontal plane is

$$R = \sqrt{1 - \tan^2\theta}.$$ □

3D. ***Examples of hyperbolic isocurved surfaces.*** We now apply the geometric method to construct examples of hyperbolic isocurved surfaces. As in the elliptic case, the function β that appears in (3-6) has a geometric interpretation and we do not need minimal graphs to apply our method.

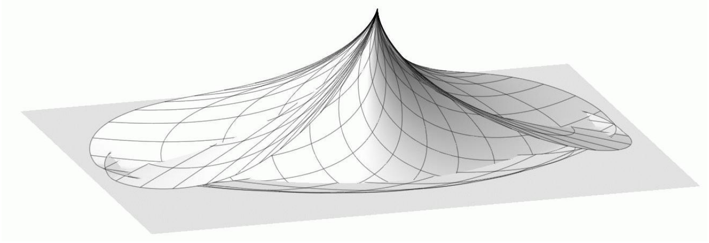


Figure 5. An Enneper-type hyperbolic isocurved surface.

For a simply connected minimal timelike surface Σ in $\mathbb{L}^3$, there is a notion of conjugate surface Σ^* and also a sort of Weierstrass representation; see [Milnor 1990]. It turns out that β in (3-6) is now given by $\beta = -x_3$, i.e., the negative of the height function of Σ^*.

We now use some known examples of timelike minimal surfaces to construct hyperbolic isocurved surfaces.

Example (Enneper-type). Following [Inoguchi and Toda 2004], we consider the Enneper-type timelike minimal surface given by

$$X(x, y) = A(x) + B(y),$$

where

$$A(x) = \frac{1}{2}\left(x^2, nx - \frac{x^3}{3}, x + \frac{x^3}{3}\right) \quad \text{and} \quad B(y) = \frac{1}{2}\left(-y^2, y - \frac{y^3}{3}, -y - \frac{y^3}{3}\right).$$

A piece of this surface appears in Figure 5.

Example. Our last example is induced from a timelike minimal surface that is the graph of the function

$$\psi(x, y) = \frac{y}{\tanh x},$$

that appears in [Milnor 1990]. The conjugate height function ψ^* in this case can be found by direct integration and is given by

$$\psi^* = -\sqrt{y^2 - 1 + \cosh^2 x}.$$

Our method yields the isocurved immersion that is illustrated in Figure 6. The actual expression of the immersion is rather complicated and we will omit it.

4. Final remarks

4A. ***Parabolic isocurved surfaces.*** We have shown how to locally construct examples of elliptic and hyperbolic isocurved surfaces from minimal surfaces in $\mathbb{R}^3$ and $\mathbb{L}^3$. As far as parabolic isocurved surfaces go, we have not found a

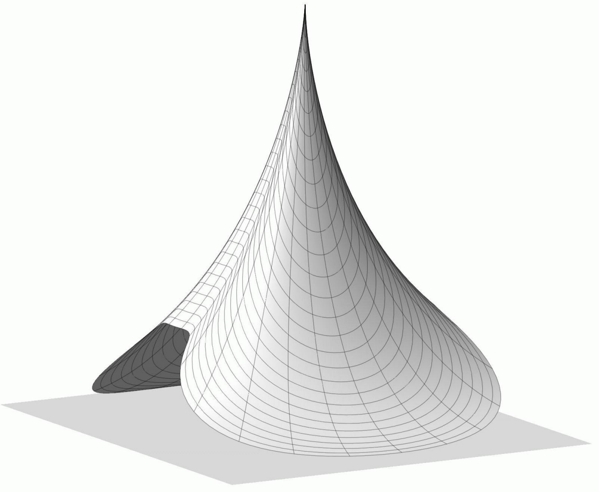


Figure 6. Hyperbolic isocurved surface.

geometric way to generate them. However, we note that there is an interesting family of examples of parabolic isocurved surfaces that was presented by Robert Bryant in his answer to a MathOverflow question posed by the second author; see http://mathoverflow.net/a/108813/53193. Bryant's examples have the form

$$X(s,t) = (a(s) + \cos s\,(t - \tanh t),\, b(s) + \sin s\,(t - \tanh t),\, \operatorname{sech} t)),$$

where $a(s)$ and $b(s)$ are functions such that

$$a'(s)\cos s + b'(s)\sin s = 0.$$

The simplest choice, $a(s) = b(s) = 0$, yields a surface of revolution with respect to the x_3-axis that has a tractrix as the profile curve.

It is also worth mentioning that there are smooth isocurved surfaces that change their type (from elliptic to hyperbolic) and have a smooth curve where the surface is parabolic. The simplest example of such surface is provided by the right circular cones with vertical axis.

4B. ***Anti-isocurved surfaces.*** In our study of hyperbolic isocurved surfaces, we have used timelike minimal surfaces in $\mathbb{L}^3$. One could ask what happens if instead of a timelike surface we start with a spacelike surface in $\mathbb{L}^3$ with vanishing mean curvature (these are known as maximal surfaces). The answer is that the associated congruence of h-geodesics admits orthogonal surfaces and it turns out that for these surfaces $K_h = -K_e$, a class of surfaces that could be called *anti-isocurved surfaces*.

In this context, we recall that there are smooth surfaces in $\mathbb{L}^3$ such as the graph of the function

$$\varphi(x, y) = \ln \frac{\cosh x}{\cosh y},$$

that have zero mean curvature and change their type (i.e., from spacelike to timelike), and each element of the induced family of orthogonal surfaces will be divided into a region where it is isocurved and another region where it is anti-isocurved.

References

[Inoguchi and Toda 2004] J. Inoguchi and M. Toda, "Timelike minimal surfaces via loop groups", *Acta Appl. Math.* **83**:3 (2004), 313–355. MR 2005g:53011 Zbl 1185.53008

[Ivey and Landsberg 2003] T. A. Ivey and J. M. Landsberg, *Cartan for beginners: differential geometry via moving frames and exterior differential systems*, Graduate Studies in Mathematics **61**, American Mathematical Society, Providence, RI, 2003. MR 2004g:53002 Zbl 1105.53001

[López 2001] R. López, "Graphs of constant mean curvature in hyperbolic space", *Ann. Global Anal. Geom.* **20**:1 (2001), 59–75. MR 2002e:53009 Zbl 0996.53039

[Mazet et al. 2007] L. Mazet, M. M. Rodríguez, and M. Traizet, "Saddle towers with infinitely many ends", *Indiana Univ. Math. J.* **56**:6 (2007), 2821–2838. MR 2009c:53009 Zbl 05254001

[Milnor 1990] T. K. Milnor, "Entire timelike minimal surfaces in $E^{3,1}$", *Michigan Math. J.* **37**:2 (1990), 163–177. MR 92b:53007 Zbl 0729.53015

[Nitsche 1989] J. C. C. Nitsche, *Lectures on minimal surfaces: Introduction, fundamentals, geometry and basic boundary value problems*, vol. 1, Cambridge University Press, 1989. MR 90m:49031 Zbl 1209.53002

[Tenenblat 1998] K. Tenenblat, *Transformations of manifolds and applications to differential equations*, Pitman Monographs and Surveys in Pure and Applied Mathematics **93**, Longman, Harlow, 1998. MR 2002b:37111 Zbl 0912.58045

Received May 16, 2013. Revised July 28, 2014.

NILTON BARROSO
DEPARTAMENTO DE MATEMÁTICA
UNIVERSIDADE DE BRASÍLIA
70910-900 BRASILIA, DF
BRAZIL

n.m.b.neto@mat.unb.br

PEDRO ROITMAN
DEPARTAMENTO DE MATEMÁTICA
UNIVERSIDADE DE BRASÍLIA
70910-900 BRASILIA, DF
BRAZIL

roitman@mat.unb.br

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.39

COHOMOLOGY OF LOCAL SYSTEMS ON THE MODULI OF PRINCIPALLY POLARIZED ABELIAN SURFACES

DAN PETERSEN

Let $\mathcal{A}_2$ be the moduli stack of principally polarized abelian surfaces. Let $\mathbb{V}$ be a smooth ℓ-adic sheaf on $\mathcal{A}_2$ associated to an irreducible rational finite-dimensional representation of Sp(4). We give an explicit expression for the cohomology of $\mathbb{V}$ in any degree in terms of Tate-type classes and Galois representations attached to elliptic and Siegel cusp forms. This confirms a conjecture of Faber and van der Geer. As an application we prove a dimension formula for vector-valued Siegel cusp forms for Sp(4, $\mathbb{Z}$) of weight three, which had been conjectured by Ibukiyama.

1. Introduction

Let $Y = \Gamma\backslash\mathfrak{H}$ be a modular curve, given by the quotient of the upper half-plane by a congruence subgroup $\Gamma \subset \mathrm{SL}(2, \mathbb{Z})$. An irreducible rational representation $\mathbb{V}$ of SL(2) defines a local system on Y, since $\mathbb{V}$ is in particular a representation of $\pi_1(Y) \cong \Gamma \subset \mathrm{SL}(2)$. After work of Eichler, Shimura, Ihara, Deligne, and many others after them, we understand extremely well the cohomology groups $H^\bullet(Y, \mathbb{V})$. The cohomology classes can be described group-theoretically in terms of modular forms for the group Γ, and the mixed Hodge structure on the cohomology groups has a natural splitting in which the pure part corresponds to cusp forms and its complement to Eisenstein series. We can think of $\mathbb{V}$ also as a smooth ℓ-adic sheaf (and Y as defined over a number field, or a deeper arithmetic base), in which case the étale cohomology $H^\bullet(Y, \mathbb{V})$ can be expressed in terms of Galois representations attached to the same modular forms [Deligne 1971].

There is a vast theory describing the generalization of the above to moduli spaces of higher-dimensional abelian varieties with some extra structure (polarization, endomorphism, and level), and to more general Shimura varieties. But there is not a single example where our understanding is as complete as in genus one.

This work was carried out at KTH Royal Institute of Technology, supported by the Göran Gustafsson foundation, and in the group of Pandharipande at ETH Zürich, supported by grant ERC-2012-AdG-320368-MCSK.
MSC2010: 11F46, 11F67, 11F75, 11G18, 14K10.
Keywords: abelian surfaces, zeta functions, Galois representations, cohomology of Shimura varieties.

In this article we consider one of the simplest higher-genus examples and give a quite explicit description of the cohomology in this case. Namely, consider the moduli space $\mathcal{A}_2$ of principally polarized abelian surfaces, and let $\mathbb{V}$ be a smooth ℓ-adic sheaf associated to an irreducible representation of $\mathrm{Sp}(4)$. The main theorem of this article is an explicit expression for the (semisimplification of the) ℓ-adic Galois representation $H^k_c(\mathcal{A}_2, \mathbb{V})$ for any k and any $\mathbb{V}$ in terms of Tate-type classes and Galois representations attached to level-1 elliptic/Siegel cusp forms.

These cohomology groups are natural objects of study for algebraic geometers, in particular because of applications to moduli of curves. The results of this paper are used in [Petersen 2013] to prove that the Gorenstein conjecture fails for the tautological rings of the spaces $\mathcal{M}^{\mathsf{ct}}_{2,n}$ for $n \geq 8$. There is some history of algebraic geometers studying the cohomology of $\mathbb{V}_{a,b}$ for small values of $a+b$ by ad hoc methods for such applications; see, e.g., [Getzler 1998, Section 8; Bergström 2009; Petersen and Tommasi 2014, Section 3]. Let us also mention [Faber and van der Geer 2004], who used point counts over finite fields to conjecture an expression for the virtual ℓ-adic Galois representation

$$\sum_k (-1)^k [H^k_c(\mathcal{A}_2, \mathbb{V})] \in K_0(\mathsf{Gal})$$

for any $\mathbb{V}_{a,b}$; see also [Bergström et al. 2014, Section 6] for a more detailed description. The results in this paper confirm Faber and van der Geer's conjecture. When $\mathbb{V}$ has regular highest weight, their conjecture was proven in [Weissauer 2009b] (and later independently in [Tehrani 2013]).

Using the BGG-complex of Faltings, one can relate the results of this paper to the coherent cohomology of the bundles of Siegel modular forms for $\mathrm{Sp}(4, \mathbb{Z})$, as we explain at the end of Theorem 2.1. A direct consequence of our main theorem is a proof of a dimension formula for vector-valued Siegel modular forms for $\mathrm{Sp}(4, \mathbb{Z})$ of weight 3, which had been conjectured in [Ibukiyama 2007b]. This result has been independently obtained in [Taïbi 2014] using Arthur's trace formula.

The strategy of our proof is as follows. Up to semisimplification, the cohomology is the direct sum of the *Eisenstein cohomology* and the *inner cohomology*. The Eisenstein cohomology on $\mathcal{A}_2$ of an arbitrary local system was determined in [Harder 2012], so we need only to find the inner cohomology. Now we use that the inner cohomology contains the cuspidal cohomology and is contained in the intersection cohomology, and both of these can be understood in terms of data attached to discrete spectrum automorphic representations for $\mathrm{GSp}(4)$. There is a very large body of work dealing with automorphic representations on $\mathrm{GSp}(4)$ (due to Piatetski-Shapiro, Soudry, Arthur, Weissauer, Taylor, Hales, Waldspurger and many others) since it is one of the first test cases for the general Langlands program. Since we will only work in level 1, we can work with $\mathrm{PGSp}(4)$, in which case all necessary information on the discrete spectrum automorphic representations is worked out and described very explicitly in [Flicker 2005]. These results allow us to determine both the cuspidal and the intersection cohomology of these local systems, and to deduce after comparing with Harder's results that the inner cohomology coincides with the cuspidal cohomology in these cases.

In Section 2, I state the main theorem and explain the applications to vector-valued Siegel cusp forms. Section 3 contains a brief review of automorphic representations and the cohomology of Shimura varieties. I hope that this will help make the arguments accessible for algebraic geometers without this background. Section 4 specializes to PGSp(4) and contains the proof of the main theorem.

2. Statement of results

Let $\mathcal{A}_2$ denote the moduli stack of principally polarized abelian surfaces. Let $f\colon \mathscr{X} \to \mathcal{A}_2$ be the universal family. We have a local system (smooth ℓ-adic sheaf) $\mathbb{V} = \mathrm{R}^1 f_* \mathbb{Q}_\ell$ on $\mathcal{A}_2$ of rank 4 and weight 1, and there is a symplectic pairing

$$\textstyle\bigwedge^2 \mathbb{V} \to \mathbb{Q}_\ell(-1).$$

Here $\mathbb{Q}_\ell(-1)$ denotes the Tate twist of the constant local system on $\mathcal{A}_2$. Recall Weyl's construction of the irreducible representations of Sp(4) [Fulton and Harris 1991, Section 17.3]: if V is the standard 4-dimensional symplectic vector space, then the irreducible representation with highest weight $a \geq b \geq 0$ is a constituent of $V^{\otimes(a+b)}$, where it is "cut out" by Schur functors and by contracting with the symplectic form. For instance, the representation of highest weight (2, 0) is $\mathrm{Sym}^2(V)$, and the representation (1, 1) is the complement of the class of the symplectic form inside $\bigwedge^2 V$. Weyl's construction works equally well in families, and so for each $a \geq b \geq 0$ we obtain a local system $\mathbb{V}_{a,b}$ which is a summand in $\mathbb{V}^{\otimes(a+b)}$. In this paper we determine the cohomology of $\mathbb{V}_{a,b}$ considered as an ℓ-adic Galois representation up to semisimplification.

Note that every point of $\mathcal{A}_2$ has the automorphism (-1), given by inversion on the abelian variety. This automorphism acts as multiplication by $(-1)^{a+b}$ on the fibers of $\mathbb{V}_{a,b}$. This shows that the local system has no cohomology when $a+b$ is odd. Hence, we restrict our attention to the case when $a+b$ is even.

Before we can state our main results we need to introduce some notation. For any k, let s_k denote the dimension of the space of cusp forms for $\mathrm{SL}(2, \mathbb{Z})$ of weight k. Similarly, for any $j \geq 0$, $k \geq 3$, we denote by $s_{j,k}$ the dimension of the space of vector-valued Siegel cusp forms for $\mathrm{Sp}(4, \mathbb{Z})$, transforming according to the representation $\mathrm{Sym}^j \otimes \det^k$.

To each normalized cusp eigenform f for $\mathrm{SL}(2, \mathbb{Z})$ of weight k is attached a 2-dimensional ℓ-adic Galois representation ρ_f of weight $k-1$ [Deligne 1971]. We define $\mathsf{S}_k = \bigoplus_f \rho_f$ to be the direct sum of these Galois representation for fixed k. By the main theorem of [Weissauer 2005], there are also 4-dimensional Galois representations attached to vector-valued Siegel cusp eigenforms for $\mathrm{Sp}(4, \mathbb{Z})$ of type $\mathrm{Sym}^j \otimes \det^k$ with $k \geq 3$, and we define $\mathsf{S}_{j,k}$ analogously. So $\dim \mathsf{S}_k = 2s_k$ and $\dim \mathsf{S}_{j,k} = 4s_{j,k}$.

Moreover, we introduce s'_k: this is the cardinality of the set of normalized cusp eigenforms f of weight k for $\mathrm{SL}(2, \mathbb{Z})$ for which the central value $L\big(f, \frac{1}{2}\big)$ vanishes. In this paper all L-functions will be normalized to have a functional equation relating s and $1-s$. The functional equation shows that the order of $L(f, s)$ at $s = \frac{1}{2}$ is always odd if $k \equiv 2 \pmod 4$ and is even if $k \equiv 0 \pmod 4$. Hence, in the former case, $s_k = s'_k$; in the latter case, $0 \leq s'_k \leq s_k$. In our results, the quantity s'_k will only occur in the case $k \equiv 0 \pmod 4$, and in this case it is conjectured that $s'_k = 0$. Indeed, [Conrey and Farmer 1999] proved that this vanishing is implied by Maeda's conjecture; Maeda's conjecture has been verified numerically for weights up to 14000 [Ghitza and McAndrew 2012].

Finally we define $\bar{\mathsf{S}}_{j,k} = \mathfrak{gr}^W_{j+2k-3} \mathsf{S}_{j,k}$; in other words, we consider only the part of $\mathsf{S}_{j,k}$ which satisfies the Ramanujan conjecture. Counterexamples to the Ramanujan conjecture arise from the Saito–Kurokawa lifting: for a cusp eigenform f of weight $2k$ for $\mathrm{SL}(2, \mathbb{Z})$, where k is odd, there is attached a scalar-valued Siegel cusp form of weight $k+1$ for $\mathrm{Sp}(4, \mathbb{Z})$ whose attached ℓ-adic Galois representation has the form

$$\mathbb{Q}_\ell(-k+1) \oplus \rho_f \oplus \mathbb{Q}_\ell(-k),$$

where ρ_f is the Galois representation of weight $2k-1$ attached to f. By [Weissauer 2009b, Theorem 3.3], these are in fact the only Siegel cusp forms violating the Ramanujan conjecture. Thus, $\bar{\mathsf{S}}_{j,k} = \mathsf{S}_{j,k}$ unless $j = 0$ and k is even, in which case $\bar{\mathsf{S}}_{j,k}$ is obtained from $\mathsf{S}_{j,k}$ by removing the two summands of Tate type from each Saito–Kurokawa lift.

Note that the definitions of s_k, S_k, $s_{j,k}$ and $\mathsf{S}_{j,k}$ used in [Faber and van der Geer 2004] are different from ours: theirs is not only a sum over cusp forms, but includes in the case $k = 2$ (resp. $j = 0$, $k = 3$) the contribution from the trivial automorphic representation. This allows for a compact expression for the virtual Galois representation $\sum_i (-1)^i [H^i_c(\mathcal{A}_2, \mathbb{V}_{a,b})]$ but will not be used here.

Theorem 2.1. *Suppose $(a, b) \neq (0, 0)$, and that $a+b$ is even. Then*:

(1) *$H^k_c(\mathcal{A}_2, \mathbb{V}_{a,b})$ vanishes for $k \notin \{2, 3, 4\}$.*

(2) *In degree 4 we have*

$$H^4_c(\mathcal{A}_2, \mathbb{V}_{a,b}) = \begin{cases} s_{a+b+4}\mathbb{Q}_\ell(-b-2) & \textit{if } a = b \textit{ even}, \\ 0 & \textit{otherwise.} \end{cases}$$

(3) *In degree* 3 *we have, up to semisimplification,*

$$H^3_c(\mathcal{A}_2, \mathbb{V}_{a,b}) = \bar{\mathsf{S}}_{a-b,b+3} + s_{a+b+4}\mathsf{S}_{a-b+2}(-b-1) + \mathsf{S}_{a+3} + \begin{cases} s'_{a+b+4}\mathbb{Q}_\ell(-b-1) & \textit{if } a=b \textit{ even},\\ s_{a+b+4}\mathbb{Q}_\ell(-b-1) & \textit{otherwise}, \end{cases} + \begin{cases} \mathbb{Q}_\ell & \textit{if } a=b \textit{ odd},\\ 0 & \textit{otherwise}, \end{cases} + \begin{cases} \mathbb{Q}_\ell(-1) & \textit{if } b=0,\\ 0 & \textit{otherwise.} \end{cases}.$$

(4) *In degree* 2 *we have, again up to semisimplification, that*

$$H^2_c(\mathcal{A}_2, \mathbb{V}_{a,b}) = \mathsf{S}_{b+2} + s_{a-b+2}\mathbb{Q}_\ell + \begin{cases} s'_{a+b+4}\mathbb{Q}_\ell(-b-1) & \textit{if } a=b \textit{ even},\\ 0 & \textit{otherwise}, \end{cases} + \begin{cases} \mathbb{Q}_\ell & \textit{if } a>b>0 \textit{ and } a,\ b \textit{ even},\\ 0 & \textit{otherwise.} \end{cases}$$

To exemplify the notation: $s_{a+b+4}\mathsf{S}_{a-b+2}(-b-1)$ means a direct sum of s_{a+b+4} copies of the Galois representation S_{a-b+2}, Tate twisted $b+1$ times.

As remarked earlier, it is conjectured that both occurrences of s'_k in the above theorem can be replaced by 0.

Remark 2.2. It will be clear from the proof that the result is valid (and even a bit easier) also in the category of mixed Hodge structures. Harder's computation of the Eisenstein cohomology is valid in this category, and our computation of the inner cohomology identifies it with the cuspidal cohomology, which obtains a natural Hodge structure from the bigrading on $(\mathfrak{g}, K)$-cohomology. This bigrading is compatible with the one obtained using the "filtration bête" and the BGG-complex of [Faltings and Chai 1990, Theorem VI.5.5.].

Remark 2.3. It is conjectured that the Galois representations $H^k_c(\mathcal{A}_2, \mathbb{V}_{a,b})$ are not semisimple in general. Suppose that $a=b=2k-1$. Then our expression for the semisimplification of $H^3_c(\mathcal{A}_2, \mathbb{V}_{a,b})$ contains the terms $s_{4k+2}\mathbb{Q}_\ell(-2k)$ and S_{4k+2}, the latter being the "Saito–Kurokawa" summand of $\bar{\mathsf{S}}_{0,2k+2}$. Harder [1993, pp. 81–82] has conjectured that they form a nontrivial extension

$$0 \to s_{4k+2}\mathbb{Q}_\ell(-2k) \to M \to \mathsf{S}_{4k+2} \to 0.$$

Note that if f is a Hecke eigenform of weight $4k+2$ and ρ_f is the attached Galois representation (or "motive"), then conjectures of Deligne, Bloch and Beilinson [Gross 1994, Section 1] predict that

$$\dim \operatorname{Ext}^1(\mathbb{Q}_\ell(-2k), \rho_f) = \operatorname{ord}_{s=\frac{1}{2}} L(f,s),$$

and the functional equation for $L(f, s)$ forces it to vanish at $s = \frac{1}{2}$. Here the Ext group is computed either in the category of ℓ-adic Galois representations, or (even better) in the category of mixed motives. I do not know whether there exists a cusp form for the full modular group whose L-function vanishes to more than first order at the central point.

Application to dimension formulas for Siegel modular forms. A consequence of Remark 2.2 is that our main theorem can be applied to produce dimension formulas for vector-valued Siegel modular forms. Let $i : \mathcal{A}_2 \hookrightarrow \tilde{\mathcal{A}}_2$ be a toroidal compactification. Let $\mathcal{V}_{j,k}$ for $j, k \in \mathbb{Z}$, $j \geq 0$, be the vector bundle on $\tilde{\mathcal{A}}_2$ whose global sections are vector-valued Siegel modular forms of type $\mathrm{Sym}^j \otimes \det^k$. Similarly, let $\mathcal{V}_{j,k}(-D_\infty)$ be the vector bundle of Siegel cusp forms. The *BGG-complex* (resp. the *dual BGG-complex*) is a resolution of $i_* \mathbb{V}_{a,b} \otimes \mathbb{C}$ (resp. $i_! \mathbb{V}_{a,b} \otimes \mathbb{C}$) in terms of the vector bundles $\mathcal{V}_{j,k}$ (resp. $\mathcal{V}_{j,k}(-D_\infty)$). Then [Faltings and Chai 1990, Theorem VI.5.5] asserts that the hypercohomology spectral sequence of the BGG-complex degenerates, and that the Hodge filtration on the cohomology of $\mathbb{V}_{a,b}$ can be defined in terms of a filtration of the BGG-complex. There is also an analogous statement for the dual BGG-complex and the compactly supported cohomology. Specialized to our case, their theorem (in the case of the dual BGG-complex) asserts the following (see [Getzler 1998, Theorem 17]):

Theorem 2.4 (Faltings–Chai). *The cohomology groups $H^\bullet_c(\mathcal{A}_2, \mathbb{V}_{a,b} \otimes \mathbb{C})$ have a Hodge filtration with Hodge numbers in the set $\{a+b+3, a+2, b+1, 0\}$. The associated graded pieces satisfy*

$$\begin{aligned}
\mathfrak{gr}^0_F H^\bullet_c(\mathcal{A}_2, \mathbb{V}_{a,b} \otimes \mathbb{C}) &\cong H^\bullet(\tilde{\mathcal{A}}_2, \mathcal{V}_{a-b,-a}(-D_\infty)),\\
\mathfrak{gr}^{b+1}_F H^\bullet_c(\mathcal{A}_2, \mathbb{V}_{a,b} \otimes \mathbb{C}) &\cong H^{\bullet-1}(\tilde{\mathcal{A}}_2, \mathcal{V}_{a+b+2,-a}(-D_\infty)),\\
\mathfrak{gr}^{a+2}_F H^\bullet_c(\mathcal{A}_2, \mathbb{V}_{a,b} \otimes \mathbb{C}) &\cong H^{\bullet-2}(\tilde{\mathcal{A}}_2, \mathcal{V}_{a+b+2,1-b}(-D_\infty)),\\
\mathfrak{gr}^{a+b+3}_F H^\bullet_c(\mathcal{A}_2, \mathbb{V}_{a,b} \otimes \mathbb{C}) &\cong H^{\bullet-3}(\tilde{\mathcal{A}}_2, \mathcal{V}_{a-b,b+3}(-D_\infty)).
\end{aligned}$$

We record three immediate consequences of this theorem combined with our main theorem. The first of these is a proof of a conjecture of Ibukiyama, whereas the second two are new proofs of results which are already known (by admittedly much more direct arguments).

(1) The bundles $\mathcal{V}_{j,k}(-D_\infty)$ have no higher cohomology for any $j \geq 0$, $k \geq 3$, with the sole exception of $H^3(\tilde{\mathcal{A}}_2, \mathcal{V}_{0,3}(-D_\infty)) \cong \mathbb{C}$. (To prove this, consider $\mathfrak{gr}^{a+b+3}_F$.) An explicit formula for the Euler characteristic of the vector bundles $\mathcal{V}_{j,k}(-D_\infty)$ was calculated in [Tsushima 1983] using Hirzebruch–Riemann–Roch; thus, we obtain a dimension formula for vector-valued Siegel cusp forms for all $j \geq 0$, $k \geq 3$. Tsushima himself proved that these bundles have no higher cohomology when $k \geq 5$ using the Kawamata–Viehweg vanishing theorem, and conjectured that it can be

improved to $k \geq 4$. The fact that this vanishing result can be extended to $k \geq 3$ is particular to the case of the full modular group and was conjectured in [Ibukiyama 2007b, Conjecture 2.1]. The resulting dimension formula for $k = 3$ can be stated as

$$\sum_{j\geq 0} s_{j,3}x^j = \frac{x^{36}}{(1-x^6)(1-x^8)(1-x^{10})(1-x^{12})}.$$

This result has also been proven in [Taïbi 2014, Section 5].

(2) There are no vector-valued Siegel modular forms of weight 1 for the full modular group. (Put $b = 0$ and consider $\mathfrak{gr}_F^{a+2}$ to prove the case $\mathrm{Sym}^j \otimes \det$ with $j \geq 2$; the cases $j < 2$ require a separate (easy) argument.) This result was previously known by [Ibukiyama 2007a, Theorem 6.1].

(3) The Siegel Φ-operator is surjective for any $j \geq 0, k \geq 3$. Recall that the Φ-operator maps Siegel modular forms of type $\mathrm{Sym}^j \otimes \det^k$ to elliptic modular forms of weight $j+k$, and that the image of Φ consists only of cusp forms if $j > 0$. Now, the dimension of the part of $\mathfrak{gr}_F^{a+b+3} H^3(\mathcal{A}_2, \mathbb{V}_{a,b} \otimes \mathbb{C})$ given by Eisenstein cohomology is exactly the dimension of the image of the Φ-operator for $\mathrm{Sym}^{a-b} \otimes \det^{b+3}$, since the part given by inner cohomology coincides with the dimension of the space of cusp forms. But the dimension of this part of Eisenstein cohomology is s_{a+3} unless $a = b$ is odd, in which case it is $s_{a+3} + 1$. The result follows from this. Surjectivity of the Φ-operator is known more generally for arbitrary level when $k \geq 5$ and $j > 0$ by [Arakawa 1983]. The scalar-valued case is a classical theorem of Satake. The case $k = 4$ (and $k = 2$) is [Ibukiyama and Wakatsuki 2009, Theorem 5.1].

Only Siegel modular forms of weight two are inaccessible via the cohomology of local systems. In a sequel to this paper we will use similar arguments to derive dimensional results for Siegel modular forms with nontrivial level.

3. Résumé of automorphic representations

In this section I briefly recall some (mostly standard) facts from the theory of automorphic representations that are needed for this paper. Rather than providing detailed references everywhere, I will give general references at the beginning of each subsection.

Automorphic representations. See [Borel and Jacquet 1979; Cogdell et al. 2004]. Let G be a reductive connected group over $\mathbb{Q}$. Let $\mathbb{A} = \mathbb{A}_{\mathrm{fin}} \times \mathbb{R}$ be the ring of (rational) adèles. Let Z be the center of G, and ω a unitary character of $Z(\mathbb{A})/Z(\mathbb{Q})$. We define $L^2(G(\mathbb{Q})\backslash G(\mathbb{A}), \omega)$ to be the space of measurable functions f on $G(\mathbb{Q})\backslash G(\mathbb{A})$ which are square-integrable with respect to a translation-invariant measure, and which satisfy $f(zg) = \omega(z)f(g)$ for any $z \in Z(\mathbb{A})$. The group $G(\mathbb{A})$ acts on this space by right translation. A representation of $G(\mathbb{A})$ is called *automorphic* if it is a subquotient of $L^2(G(\mathbb{Q})\backslash G(\mathbb{A}), \omega)$ for some ω. We call ω the *central character* of the automorphic representation.

The space $L^2(G(\mathbb{Q})\backslash G(\mathbb{A}), \omega)$ contains a maximal subspace which is a direct sum of irreducible representations. This subspace is called the *discrete spectrum*, and an automorphic representation occurring here is called *discrete*. The orthogonal complement of this subspace is the *continuous spectrum*. Langlands identified the continuous spectrum with "Eisenstein series"; it is the direct integral of families of representations induced from parabolic subgroups of $G(\mathbb{A})$. The discrete spectrum, in turn, also decomposes as the direct sum of the *cuspidal* and the *residual* spectrum. The cuspidal spectrum is defined as the subspace spanned by functions f such that the integral over $N(\mathbb{Q})\backslash N(\mathbb{A})$ of f, and all its translates under $G(\mathbb{A})$, vanishes for N the unipotent radical of any proper parabolic subgroup. Langlands proved that the residual spectrum is spanned by the residues of Eisenstein series, and that all residual representations are quotients of representations induced from a parabolic subgroup.

Any irreducible automorphic representation π of $G(\mathbb{A})$ is a completed (restricted) tensor product of local representations π_p of $G(\mathbb{Q}_p)$, where p ranges over the prime numbers, and an archimedean component π_∞. Let $K_p \subset G(\mathbb{Q}_p)$ be a special maximal compact subgroup. We say that π is *spherical* at p if π_p contains a nonzero vector fixed by K_p, in which case this vector will be unique up to a nonzero scalar. The representation π is spherical at all but finitely many primes. The word "restricted" in the first sentence of this paragraph means that the component of the representation at p should be equal to the spherical vector for all but finitely many p.

The archimedean component π_∞ can be identified with an irreducible $(\mathfrak{g}, K_\infty)$-module, where $\mathfrak{g}$ is the Lie group of $G(\mathbb{R})$ and $K_\infty \subset G(\mathbb{R})$ is a maximal compact subgroup. The center of the universal enveloping algebra of $\mathfrak{g}$ acts by a scalar on π_∞. The resulting map $Z(\mathcal{U}\mathfrak{g}) \to \mathbb{C}$ is called the *infinitesimal character* of π.

Local factors. See [Borel 1979]. Suppose π is spherical at p. We define the *spherical Hecke algebra* $\mathscr{H}_{G,K_p}$ to be the convolution algebra of K_p-bi-invariant, $\mathbb{Q}$-valued functions on $G(\mathbb{Q}_p)$. This algebra acts on the one-dimensional space of spherical vectors, and π_p is uniquely determined by this action. Hence, specifying a spherical representation is equivalent to specifying a homomorphism $\mathscr{H}_{G,K_p} \to \mathbb{C}$. We should therefore understand the ring $\mathscr{H}_{G,K_p}$, and this we can do via the *Satake isomorphism*. For this we need the notion of the *dual group*. If G is defined by a root datum, then the dual group $\widehat{G}$ is obtained by switching roots and coroots, and characters and 1-parameter subgroups. The Satake isomorphism states that the Hecke algebra $\mathscr{H}_{G,K_p}$ and the ring of virtual representations $K_0(\mathrm{Rep}(\widehat{G}))$ become isomorphic after an extension of scalars: one has

$$\mathscr{H}_{G,K_p} \otimes \mathbb{C} \cong K_0(\mathrm{Rep}(\widehat{G})) \otimes \mathbb{C}.$$

In particular, a homomorphism $\mathscr{H}_{G,K_p} \to \mathbb{C}$ is identified with a homomorphism $K_0(\mathrm{Rep}(\widehat{G})) \to \mathbb{C}$. But the latter is determined by a semisimple conjugacy class c_p in $\widehat{G}(\mathbb{C})$. (You evaluate such a class on a representation V via $\mathrm{Tr}(c_p|V)$.)

Now suppose instead that we have an ℓ-adic (or λ-adic) representation

$$\rho\colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \widehat{G}(\overline{\mathbb{Q}}_\ell).$$

For all but finitely many primes, ρ is going to be unramified, which means in particular that the expression $\rho(\mathrm{Frob}_p)$ is well defined up to conjugacy. If we choose an isomorphism $\mathbb{C} \cong \overline{\mathbb{Q}}_\ell$, then it makes sense to ask whether c_p and $\rho(\mathrm{Frob}_p)$ are conjugate for almost all p. If this holds, then we say that ρ is *attached* to the automorphic representation π. We remark that this definition would make sense also if we replaced $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ by the absolute Weil group, or the conjectural global Langlands group, as in both cases $\rho(\mathrm{Frob}_p)$ should be well defined up to conjugacy at unramified primes.

By the Chebotarev density theorem, there is at most one Galois representation attached to a given automorphic representation. The strong multiplicity one theorem shows the converse when $G = \mathrm{GL}(n)$, but in general there will be several automorphic representations with the same attached Galois representation. Conjecturally, two automorphic representations will have the same attached Galois representation if and only if they lie in the same "L-packet". However, the notion of a packet has not been rigorously defined in general.

One of many conjectures within the Langlands program says roughly that there should in fact be a bijection between packets of automorphic representations for G and ℓ-adic Galois representations into the dual group. As stated, this conjecture is however false, and making the conjecture precise is a rather delicate matter. For a formulation in terms of the hypothetical Langlands group, see [Arthur 2002], and for a more restrictive formulation only in terms of Galois representations, see [Buzzard and Gee 2014].

Often one fixes once and for all $r\colon \widehat{G} \hookrightarrow \mathrm{GL}(n)$. Then the conjugacy class c_p can be described by specifying an $n \times n$ diagonal matrix $\mathrm{diag}(t_1, \dots, t_n)$. The numbers t_i are called the *Langlands parameters* of π at p. Moreover, one can then attach an L-function to any automorphic representation. At a prime p where π is spherical, the local L-factor is given by

$$\det(\mathbb{1}_n - p^{-s} r(c_p))^{-1}.$$

On the other hand, given r we also obtain from ρ an n-dimensional ℓ-adic Galois representation, which also has an attached L-function. Thus, the Langlands parameters can be identified with the Frobenius eigenvalues of the attached Galois representations. Usually the notion of ρ being attached to π is defined in terms of an equality of L-functions, but L-functions will play only a minor role in this paper.

Shimura varieties. See [Deligne 1979; Harder 1993, Kapitel II]. For G as above, suppose that $h\colon \operatorname{Res}_{\mathbb{C}/\mathbb{R}} \mathbb{G}_m \to G_{/\mathbb{R}}$ is a homomorphism satisfying axioms 2.1.1.1–2.1.1.3 of [Deligne 1979]. Let K_∞ be the stabilizer of h in $G(\mathbb{R})$. Let K_{fin} be any compact open subgroup of $G(\mathbb{A}_{\text{fin}})$. For $K = K_{\text{fin}} \times K_\infty$, we can consider the quotient

$$S_K = G(\mathbb{Q})\backslash G(\mathbb{A})/K = G(\mathbb{Q})\backslash X \times G(\mathbb{A}_{\text{fin}})/K_{\text{fin}},$$

the *Shimura variety* associated to K. Here $X = G(\mathbb{R})/K_\infty$. For K_{fin} small enough, S_K is, in fact, a smooth algebraic variety which is naturally defined over a number field (the reflex field), but in the case we will consider in this paper we will actually need to think of S_K as an orbifold or Deligne–Mumford stack.

Example 3.1. Siegel modular varieties are Shimura varieties. Let $G = \mathrm{GSp}(2g)$ and put

$$h(x+iy) = \begin{bmatrix} xI_g & yI_g \\ -yI_g & xI_g \end{bmatrix}.$$

Then $X = \mathfrak{H}_g \sqcup \overline{\mathfrak{H}}_g$ is the union of Siegel's upper half-space and its complex conjugate. If we choose $K_{\text{fin}} = G(\widehat{\mathbb{Z}})$, then $G(\mathbb{A}_{\text{fin}}) = G(\mathbb{Q}) \cdot K_{\text{fin}}$ and

$$S_K = G(\mathbb{Q})\backslash X \times G(\mathbb{A}_{\text{fin}})/K_{\text{fin}} \cong (G(\mathbb{Q}) \cap K_{\text{fin}})\backslash X = G(\mathbb{Z})\backslash X.$$

Now $G(\mathbb{Z})\backslash X$ is naturally isomorphic to the stack $\mathcal{A}_g$ parametrizing principally polarized abelian varieties of dimension g. Had we chosen K_{fin} smaller, S_K would instead be a disjoint union of finite covers of $\mathcal{A}_g$, parametrizing abelian varieties with "K_{fin}-level structure".

Let $\mathbb{V}$ be an irreducible finite-dimensional rational representation of G. To $\mathbb{V}$ we can attach a local system on S_K, which we also denote by $\mathbb{V}$. As the reader may already have noticed, we (sloppily) use "local system" as a catch-all term to describe several different structures: we obtain a locally constant sheaf of $\mathbb{Q}$-vector spaces on the topological space $S_K(\mathbb{C})$ which in a natural way underlies a variation of Hodge structure; moreover, $\mathbb{V} \otimes \mathbb{Q}_\ell$ can (for any ℓ) be identified with the base change of a smooth ℓ-adic sheaf on S_K over the reflex field. The étale cohomology groups of said ℓ-adic sheaves are (after base changing to $\mathbb{C}$) related to the ordinary singular cohomology groups by a comparison isomorphism, and we may think informally of $\mathbb{V}$ as a "motivic sheaf" and $H^\bullet(S_K, \mathbb{V})$ as a "mixed motive" with a compatible system of ℓ-adic and Hodge-theoretic realizations.

Decomposing cohomology. See [Arthur 1996]. In this subsection we will find the need to compare several different cohomology theories. We will use the phrase "ordinary cohomology" to refer to the usual cohomology of the topological space $S_K(\mathbb{C})$.

The spectral decomposition of $L^2(G(\mathbb{Q})\backslash G(\mathbb{A}))$ contains much information about the cohomology of Shimura varieties for G. The connection to automorphic representations is most transparent if we work transcendentally and consider the sheaf $\mathbb{V} \otimes \mathbb{C}$ on $S_K(\mathbb{C})$. Then, instead of the usual de Rham complex, one can consider the complex of forms ω such that ω and $d\omega$ are square-integrable; the cohomology of this complex is called the *L^2-cohomology*. The L^2-cohomology has an interpretation in terms of $(\mathfrak{g}, K_\infty)$-cohomology:

$$H^\bullet_{(2)}(S_K(\mathbb{C}), \mathbb{V} \otimes \mathbb{C}) \cong H^\bullet\big(\mathfrak{g}, K_\infty; \mathbb{V} \otimes L^2(G(\mathbb{Q})\backslash G(\mathbb{A}))^{K_{\mathrm{fin}}}\big).$$

According to [Borel and Casselman 1983, Section 4], the contribution from the continuous spectrum to the $(\mathfrak{g}, K_\infty)$-cohomology vanishes in many natural cases (including all Shimura varieties); in fact, the contribution is nonzero if and only if the L^2-cohomology is infinite-dimensional. In particular, we may in our case replace $L^2(G(\mathbb{Q})\backslash G(\mathbb{A}))$ by the direct sum $\bigoplus_\pi m(\pi)\pi$ over the discrete spectrum, giving instead the expression

$$H^\bullet_{(2)}(S_K(\mathbb{C}), \mathbb{V} \otimes \mathbb{C}) \cong \bigoplus_{\pi \text{ disc.}} m(\pi)\pi_{\mathrm{fin}}^{K_{\mathrm{fin}}} \otimes H^\bullet(\mathfrak{g}, K_\infty; \mathbb{V} \otimes \pi_\infty).$$

In this decomposition, each $\pi_{\mathrm{fin}}^{K_{\mathrm{fin}}}$ is a module over the Hecke algebra, giving the cohomology a Hecke action. Each $H^\bullet(\mathfrak{g}, K_\infty; \mathbb{V} \otimes \pi_\infty)$ has a natural (p, q)-decomposition, defining a pure Hodge structure on each cohomology group $H^k_{(2)}(S_K(\mathbb{C}), \mathbb{V})$.

We say that an automorphic representation π is *cohomological* if there exists a representation $\mathbb{V}$ for which $H^\bullet(\mathfrak{g}, K_\infty; \mathbb{V} \otimes \pi_\infty) \neq 0$. Wigner's lemma gives a necessary condition for this nonvanishing of $(\mathfrak{g}, K_\infty)$-cohomology, namely that π_∞ and $\mathbb{V}^\vee$ (denoting the contragredient) have the same infinitesimal character. For cohomological representations, the infinitesimal character is the bookkeeping device that tells you to which local system the automorphic representation will contribute L^2-cohomology.

The natural map from L^2-cohomology to ordinary cohomology is in general neither injective nor surjective. One can however also define the *cuspidal cohomology* as the direct summand

$$H^\bullet_{\mathrm{cusp}}(S_K(\mathbb{C}), \mathbb{V} \otimes \mathbb{C}) \cong \bigoplus_{\pi \text{ cusp.}} m(\pi)\pi_{\mathrm{fin}}^{K_{\mathrm{fin}}} \otimes H^\bullet(\mathfrak{g}, K_\infty; \mathbb{V} \otimes \pi_\infty)$$

of the L^2-cohomology, and it injects naturally into the ordinary cohomology [Borel 1981, Corollary 5.5].

Finally, one can consider the *inner cohomology*, which is defined as

$$H^\bullet_!(S_K, \mathbb{V}) = \mathrm{Image}(H^\bullet_c(S_K, \mathbb{V}) \to H^\bullet(S_K, \mathbb{V})).$$

When we extend scalars to $\mathbb{C}$, the inner cohomology is sandwiched between the cuspidal and the L^2-cohomology. Indeed, the map from compactly supported cohomology to ordinary cohomology always factors through the L^2-cohomology, since the orthogonal projection of a closed, compactly supported form to the space of harmonic forms is square-integrable. This shows that the inner cohomology is a subquotient of the L^2-cohomology. On the other hand, the aforementioned result of Borel shows that the cuspidal cohomology injects into the inner cohomology.

The "complement" of the inner cohomology is called the *Eisenstein cohomology*. Formally, it is defined as the cokernel of $H^\bullet_c(S_K, \mathbb{V}) \to H^\bullet(S_K, \mathbb{V})$. One could also consider the kernel, which gives the *compactly supported Eisenstein cohomology*. We denote these $H^\bullet_{\mathrm{Eis}}$ and $H^\bullet_{c,\mathrm{Eis}}$, respectively. We will consider $G = \mathrm{GSp}(2g)$, in which case each local system $\mathbb{V}$ is isomorphic to its dual, up to a twist by the multiplier. Indeed, the restriction of the representation $\mathbb{V}$ to $\mathrm{Sp}(2g)$ satisfies $\mathbb{V} \cong \mathbb{V}^\vee$, with the symplectic pairing providing the isomorphism. In this case, we see that either one of $H^\bullet_{\mathrm{Eis}}$ and $H^\bullet_{c,\mathrm{Eis}}$ determines the other via Poincaré duality.

The Zucker conjecture, proven independently in [Looijenga 1988; Saper and Stern 1990], gives an isomorphism between the L^2-cohomology of S_K and the intersection cohomology of the Baily–Borel–Satake compactification $\overline{S}_K$:

$$H^\bullet_{(2)}(S_K(\mathbb{C}), \mathbb{V} \otimes \mathbb{C}) \cong H^\bullet(\overline{S}_K(\mathbb{C}), j_{!*}\mathbb{V} \otimes \mathbb{C}),$$

where $j : S_K \to \overline{S}_K$ is the inclusion and $j_{!*}$ denotes the intermediate extension. This isomorphism is compatible with the Hecke algebra action. But the intersection cohomology makes sense algebraically, and we can decompose the intersection cohomology of $\mathbb{V}$ into irreducible Hecke modules already over some number field F. We thus get a decomposition

$$H^\bullet(\overline{S}_K, j_{!*}\mathbb{V} \otimes F) = \bigoplus_{\pi_{\mathrm{fin}}} \pi_{\mathrm{fin}}^{K_{\mathrm{fin}}} \otimes H^\bullet(\pi_{\mathrm{fin}}).$$

Here the sum runs over the finite parts of all discrete automorphic representations, and $H^\bullet(\pi_{\mathrm{fin}}) \otimes_F \mathbb{C}$ is isomorphic to $\bigoplus_{\pi_\infty} m(\pi_{\mathrm{fin}} \otimes \pi_\infty) H^\bullet(\mathfrak{g}, K_\infty; \mathbb{V} \otimes \pi_\infty)$. For any nonarchimedean place λ of F we also get a structure of λ-adic Galois representation on each $H^\bullet(\pi_{\mathrm{fin}}) \otimes F_\lambda$ by a comparison isomorphism with the étale intersection cohomology. If we do not insist on a decomposition into absolutely irreducible Hecke modules we can take $F = \mathbb{Q}$, as in Theorem 2.1, where we, for example, consider a summand corresponding to all cusp forms of given weight, instead of a decomposition into Galois representations attached to individual cusp forms. See [Blasius and Rogawski 1994, Conjecture 5.2] for a conjectural formula expressing $H^\bullet(\pi_{\mathrm{fin}}) \otimes F_\lambda$ in terms of Galois representations attached to π.

4. The case of $\mathcal{A}_2$

Consider again the stack $\mathcal{A}_2$ of principally polarized abelian surfaces. As in Example 3.1, we may think of it as a Shimura variety for GSp(4). However, we would prefer to work with $G = \mathrm{PGSp}(4)$, and there is a minor issue here. If we put $K_{\mathrm{fin}} = G(\widehat{\mathbb{Z}})$, then the corresponding Shimura variety is

$$S_K = \mathrm{PGSp}(4, \mathbb{Z})\backslash(\mathfrak{H}_2 \amalg \bar{\mathfrak{H}}_2),$$

which fails to be isomorphic to $\mathcal{A}_2$ as a *stack*. Indeed every point of $\mathcal{A}_2$ has ± 1 in its isotropy group, but a general point of S_K has trivial isotropy. The projection $\mathrm{GSp}(4) \to \mathrm{PGSp}(4)$ defines a map $\pi : \mathcal{A}_2 \to S_K$ which induces an isomorphism on coarse moduli spaces, but which is a μ_2-gerbe in the sense of stacks.

The finite-dimensional irreducible representations of G are indexed by integers $a \geq b \geq 0$ for which $a + b$ is even. The local systems on S_K obtained in this way are strongly related to the local systems $\mathbb{V}_{a,b}$ that we defined in Section 2. Specifically, if $a+b$ is even, then we may Tate twist the local system $\mathbb{V}_{a,b}$ on $\mathcal{A}_2$ to be a weight-zero variation of Hodge structure/ℓ-adic sheaf; its pushforward under π is the one that is naturally attached to an irreducible representation of PGSp(4). Since $\mathrm{R}\pi_*\mathbb{V}_{a,b} = \pi_*\mathbb{V}_{a,b}$, it will suffice to compute the cohomology of the local systems on S_K. From now on we tacitly identify the local systems on $\mathcal{A}_2$ and on S_K with each other.

In this section we will see how the results in [Flicker 2005] allow the computation of the cuspidal and intersection cohomology of these local systems on $\mathcal{A}_2$. Let me emphasize that, as mentioned in the above paragraph, by our definition the $\mathbb{V}_{a,b}$ are Weil sheaves of weight $a + b$; this is the *cohomological normalization*, which is the most natural from the point of view of algebraic geometry. There is also the *unitary normalization*, where $\mathbb{V}_{a,b}$ has weight 0, which is used in Flicker's work. If $a + b$ is even, as in our case, then the two differ only by a Tate twist. We will from now on always make this Tate twist whenever we quote results from Flicker's book, without explicitly mentioning it.

Since $\mathcal{A}_2$ is the complement of a normal crossing divisor in a smooth proper stack over $\mathrm{Spec}(\mathbb{Z})$, and the local systems $\mathbb{V}_{a,b}$ are also defined over $\mathrm{Spec}(\mathbb{Z})$, the cohomology groups $H^\bullet(\mathcal{A}_2, \mathbb{V}_{a,b} \otimes \mathbb{Q}_\ell)$ must define Galois representations of a very special kind: they are unramified at every prime $p \neq \ell$ and crystalline at ℓ. The same phenomenon is clear also on the automorphic side. If $\pi_{\mathrm{fin}}^{K_{\mathrm{fin}}} \neq 0$ and $K_{\mathrm{fin}} = G(\widehat{\mathbb{Z}})$, then π_{fin} must be spherical at all primes by definition, since $G(\mathbb{Z}_p)$ is a special maximal compact subgroup of $G(\mathbb{Q}_p)$. Conversely, if π_{fin} is spherical everywhere then $\pi_{\mathrm{fin}}^{K_{\mathrm{fin}}}$ is exactly one-dimensional.

Considering PGSp(4) rather than GSp(4) is the same as only considering automorphic representations of GSp(4) with trivial central character. The reason we can do this is that we are considering only the completely unramified case (i.e., the case of the full modular group); in general, the image of a congruence subgroup of GSp(4) in PGSp(4) will no longer be a congruence subgroup. We restrict ourselves to PGSp(4) in this paper as this is the situation considered in [Flicker 2005].

We note that Flicker's work assumes that all automorphic representations π occuring are elliptic at at least three places. This is explained in Section I.2g of Part 1 of the book. This assumption is present in order to replace Arthur's trace formula with the simple trace formula of [Flicker and Kazhdan 1988]. However, he also notes that this assumption is only present in order to simplify the exposition — the same results can be derived assuming only that π is elliptic at a single real place, using the same ideas used to derive the simple trace formula in [Flicker and Kazhdan 1988], as detailed in [Laumon 1997; 2005]. In particular, Flicker's classification of the cohomological part of the discrete spectrum carries through (an archimedean component which is cohomological is elliptic).

We begin by determining $H^\bullet(\overline{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b})$. This amounts to determining all representations in the discrete spectrum of PGSp(4) which are spherical at every finite place and cohomological, and the corresponding Galois representation $H^\bullet(\pi_f)$ for each of them. All these things are described very precisely by Flicker. Then we shall see that $H^\bullet_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b})$ is well defined as a subspace of the étale intersection cohomology, and that it coincides with the inner cohomology.

The Vogan–Zuckerman classification. Recall that an automorphic representation $\pi_{\mathrm{fin}} \otimes \pi_\infty$ is cohomological if π_∞ has nonzero $(\mathfrak{g}, K_\infty)$-cohomology with respect to some finite-dimensional representation $\mathbb{V}$. If π_∞ is in the discrete series, then π is always cohomological. The cohomological representations which are not in the discrete series can be determined by [Vogan and Zuckerman 1984]. We recall from [Taylor 1993, p. 293] the result for GSp(4):

In the regular case there are no cohomological ones apart from the two discrete series representations, which we denote by π^H and π^W (we omit the infinitesimal character from the notation). The former is in the holomorphic discrete series and the latter has a Whittaker model. Both have 2-dimensional $(\mathfrak{g}, K_\infty)$-cohomology, concentrated in degree 3: their Hodge numbers are $(3, 0)$ and $(0, 3)$, and $(2, 1)$ and $(1, 2)$, respectively.

The representations π with $\pi_\infty = \pi^H$ correspond bijectively to cuspidal Siegel modular eigenforms. If F is a holomorphic modular form on Siegel's upper half-space of genus g, then by the strong approximation theorem it defines a function on $G(\mathbb{A})$, where $G = \mathrm{GSp}(2g)$. If it is modular for the full modular group, then we obtain a function with trivial central character. The subspace spanned by all right translates of this function is the sought-for automorphic representation (or a sum of several copies of it). Conversely, any automorphic representation π with archimedean component in the holomorphic discrete series uniquely determines a holomorphic vector-valued cusp form by considering the one-dimensional space of lowest K_∞-type in π_∞, and $\pi_{\text{fin}}^{K_{\text{fin}}}$ being one-dimensional forces it to be an eigenvector for all Hecke operators. See [Asgari and Schmidt 2001] for more details.

For singular weights there are further possibilities. If $b = 0$ there is a unitary representation π^1 whose $(\mathfrak{g}, K_\infty)$-cohomology is 2-dimensional in degrees 2 and 4, with Hodge types $(2, 0)$, $(0, 2)$, $(3, 1)$ and $(1, 3)$.

If $a = b$ there are two unitary representations π^{2+} and π^{2-}. One is obtained from the other by tensoring with the sign character. Both have one-dimensional $(\mathfrak{g}, K_\infty)$-cohomology in degrees 2 and 4, with Hodge types $(1, 1)$ and $(2, 2)$.

Finally if $a = b = 0$ we must in addition consider one-dimensional representations, which have cohomology in degrees 0, 2, 4 and 6; we will ignore this case.

Packets and multiplicities. In Flicker's book, the discrete spectrum of PGSp(4) is partitioned into "packets" and "quasipackets", and he conjectures that these coincide with the conjecturally defined L-packets and A-packets. However, in the totally unramified case the situation simplifies. In general, the (conjectural) A-packets are products of *local* A-packets, which specify the possible local components π_v. The local packets at nonarchimedean v are expected to have exactly one spherical member. Since we are only going to consider representations which are spherical at *every* finite place, we thus see that π and π' will be in the same A-packet if and only if they are in the same L-packet if and only if $\pi_{\text{fin}} \cong \pi'_{\text{fin}}$. For this reason we simply write *packet* everywhere in what follows.

In Flicker's classification there are five types of automorphic representations in the discrete spectrum. In the first three types, the corresponding packets are *stable*: each representation in the packet occurs with multiplicity exactly 1 in the discrete spectrum. Types 4 and 5, however, are *unstable*. This means that the multiplicities are not constant over the packets: in general, some representations in the packet occur with multiplicity 0 and others with multiplicity 1. Flicker [2005, Section 2.II.4] gives explicit formulas for the multiplicities of the representations in the packet.

In general there are local packets at each prime p in the unstable case, which consist of either one or two elements. We write such local packets as $\{\Pi_p^+\}$ and $\{\Pi_p^+, \Pi_p^-\}$, respectively. An element of the global packet is specified by choosing an element of the local packet at each p. All but finitely many of the local packets will be singletons, so each packet is finite. When $p = \infty$ we always have $\Pi_p^- = \pi^H$. If π lies in an unstable packet, its multiplicity in the discrete spectrum depends only on the parity of the number of places p where $\pi_p = \Pi_p^-$.

However, as we have already mentioned, the local packet contains only one element for a prime p where π is spherical. More generally, certain local representations need to be discrete series in order for Π_p^- to be nonzero. Since we are in the level 1 case, this means that the representations in the packet can differ only in their archimedean component, and the multiplicity formulas simplify significantly: they depend only on whether or not $\pi_\infty = \pi^H$.

To each discrete spectrum automorphic representation π, one can attach a 4-dimensional Galois representation whose Frobenius eigenvalues at p are given by the Langlands parameters at p of π. (Here we fix the 4-dimensional spin representation of Spin(5), the dual group of PGSp(4).) If π is in a stable packet, then $H^\bullet(\pi_{\text{fin}})$ is 4-dimensional and coincides with this attached representation. In the unstable case, the attached Galois representation is always a sum of two 2-dimensional pieces, and $H^\bullet(\pi_{\text{fin}})$ is given by one of these two summands. Which of the two halves contributes nontrivially is decided by a formula similar to the multiplicity formula; see [Flicker 2005, Part 2, Section V.2]. In particular it again has the feature that it depends on the parity of the number of places where $\pi_p = \Pi_p^-$, and simplifies significantly in the completely unramified case.

***The discrete spectrum of* PGSp(4).** The discrete spectrum automorphic representations which can contribute nontrivially to $H^\bullet_{(2)}(\mathcal{A}_2, \mathbb{V}_{a,b})$ have an archimedean component with infinitesimal character $(a, b) + (2, 1)$. A complete classification into five types is given in [Flicker 2005, Theorem 2, pp. 213–216]. We deal with each type separately. This classification is the same as the one announced by Arthur [2004] for GSp(4), except that the ones of Howe–Piatetski-Shapiro-type do not appear.

We write in parentheses the names assigned to these families by Arthur.

Type 1 (general). These are exactly the ones that lift to cuspidal representations of PGL(4).

Each of these lies in a packet of cardinality 2, where the elements in the packet are distinguished by their archimedean component: one is in the holomorphic discrete series and the other has a Whittaker model. Both elements of the packet occur with multiplicity 1 in the discrete spectrum. Packets of this type correspond bijectively to vector-valued cuspidal Siegel eigenforms which are neither endoscopic (a Yoshida-type lifting) nor CAP (a Saito–Kurokawa-type lifting). The contribution from this part of the discrete spectrum to $H^\bullet(\overline{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b})$ is concentrated in degree 3 and is the sum of the Galois representations attached to the Siegel cusp forms. We shall see that the Yoshida-type liftings do not occur in level 1. We denote this contribution to the cohomology by $\mathsf{S}^{\text{gen}}_{a-b,b+3}$.

Type 2 (Soudry). These packets are singletons, and the archimedean component is π^1, and will therefore not occur unless $b = 0$. Every packet is obtained by a lifting from a cuspidal representation Π of GL(2), corresponding to a cusp eigenform of weight $a + 1$ whose central character ξ is quadratic, $\xi \neq 1$, and $\xi\Pi = \Pi$. This is obviously impossible in level 1 for several reasons: for one, a must be even, and there are no modular forms of odd weight for $\mathrm{SL}(2, \mathbb{Z})$.

Type 3 (one-dimensional). These are the representations with π_∞ one-dimensional and will only occur when $a = b = 0$; for our purposes this case can clearly be ignored.

Type 4 (Yoshida). This is the first unstable case. All these π have $\pi_\infty \in \{\pi^H, \pi^W\}$ and their L-function is the product of L-functions attached to cusp forms for GL(2). For each pair of cuspidal automorphic representations Π_1 and Π_2 of PGL(2) whose weights are $a+b+4$ and $a-b+2$, respectively, there is a packet $\{\pi\}$ of Yoshida type. As explained earlier, the fact that we are in the unramified case implies that members of the packet can only differ in their archimedean component, so we should consider only $\pi_{\text{fin}} \otimes \pi^H$ and $\pi_{\text{fin}} \otimes \pi^W$. The multiplicity formula simplifies (since we are in the unramified case) to

$$m(\pi_{\text{fin}} \otimes \pi^H) = 0, \quad m(\pi_{\text{fin}} \otimes \pi^W) = 1,$$

and so $\pi_{\text{fin}} \otimes \pi^W$ will contribute a 2-dimensional piece of the cohomology in degree 3. The trace of Frobenius on this part of cohomology is also calculated by Flicker and we find the Galois representation $\rho_{\Pi_2} \otimes \mathbb{Q}_\ell(-b-1)$, where ρ_{Π_2} is the 2-dimensional representation attached to Π_2. Summing over all Π_1 and Π_2, this part therefore contributes

$$s_{a+b+4}\mathsf{S}_{a-b+2}(-b-1)$$

to $H^3(\bar{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b})$.

We note in particular that there are no Yoshida-type liftings to Siegel cusp forms in level 1: these would correspond to a π with $\pi_\infty = \pi^H$ and multiplicity 1.

The required liftings and multiplicity formulas for the endoscopic case have also been established for GSp(4) in [Weissauer 2009a, Theorem 5.2].

Type 5 (Saito–Kurokawa). This case appears only when $a = b$. Here there are four possible archimedean components: π^H, π^W, π^{2+} and π^{2-}. Every packet contains precisely one of π^{2+} and π^{2-}. For each cuspidal automorphic representation Π of PGL(2) of weight $a+b+4$ and for $\xi \in \{1, \text{sgn}\}$, we get a Saito–Kurokawa packet $\{\pi\}$. Since we are in level 1, we can ignore the character ξ (it must be trivial), which means that π^{2-} will not appear.

I should also say that there is a minor error at this place in Flicker's book. Flicker states that the Langlands parameters at a place u are (his notation)

$$\text{diag}(\xi_u q_u^{1/2} z_{1u}, \xi_u q_u^{1/2} z_{2u}, \xi_u q_u^{-1/2} z_{2u}, \xi_u q_u^{-1/2} z_{1u}),$$

when they should be

$$\text{diag}(z_{1u}, \xi_u q_u^{1/2}, \xi_u q_u^{-1/2}, z_{2u}).$$

Let us then consider the multiplicities, which again simplify since we are in the level 1 case: we find

$$m(\pi_{\mathrm{fin}} \otimes \pi_\infty) = \tfrac{1}{2}\big(1 + \varepsilon\big(\Pi, \tfrac{1}{2}\big) \cdot (-1)^n\big),$$

where $n = 1$ if $\pi_\infty = \pi^H$ and $n = 0$ otherwise, and $\varepsilon\big(\Pi, \tfrac{1}{2}\big) = (-1)^k$ if Π is attached to a cusp form of weight $2k$.

We thus see that, if $a = b$ is odd, the only representation in the packet with nonzero multiplicity is $\pi_{\mathrm{fin}} \otimes \pi^H$, which should correspond to a Siegel modular form. The Siegel modular forms obtained in this way are precisely the classical Saito–Kurokawa liftings, and the contribution in this case is exactly S_{a+b+4}.

For $a = b$ even we could a priori have both $\pi_{\mathrm{fin}} \otimes \pi^W$ and $\pi_{\mathrm{fin}} \otimes \pi^{2+}$ with nonzero multiplicity. But we can see by studying the Frobenius eigenvalues that π^W will not appear. Indeed, the representation $\pi_{\mathrm{fin}} \otimes \pi^W$ would contribute to the intersection cohomology in degree 3, as we see from the $(\mathfrak{g}, K_\infty)$-cohomology of π^W. Then its Frobenius eigenvalues are pure of weight $a + b + 3$. But the Frobenius eigenvalues at p will be p^{b+1} and p^{b+2}, as determined by Flicker, a contradiction. On the other hand, we know that $\pi_{\mathrm{fin}} \otimes \pi^{2+}$ is automorphic: it is the Langlands quotient of

$$\mathrm{Ind}_{P(\mathbb{A})}^{\mathrm{PGSp}(4,\mathbb{A})}(\Pi \otimes 1),$$

where P is the Siegel parabolic (whose Levi component is $\mathrm{PGL}(2) \times \mathrm{GL}(1)$), and the multiplicity formula shows that it has multiplicity 1 in the discrete spectrum. The representations of this form will contribute a term $s_{a+b+4}\mathbb{Q}_\ell(-b-1)$ to $H^2(\overline{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b})$ and $s_{a+b+4}\mathbb{Q}_\ell(-b-2)$ to $H^4(\overline{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b})$.

Remark 4.1. That $\pi_\infty = \pi^W$ does not occur in the Saito–Kurokawa case is mentioned as a conjecture of Blasius and Rogawski in [Tilouine 2009, Section 6]. The argument above will prove this conjecture for $\mathrm{PGSp}(4)$. Probably a proof for $\mathrm{GSp}(4)$ in general can be obtained by a similar argument, or by considering the possible Hodge numbers of $H^\bullet(\pi_{\mathrm{fin}})$.

The inner cohomology and the proof of the main theorem. From what we have seen so far, we can completely write down the L^2-cohomology and the intersection cohomology of any local system on $\mathcal{A}_2$. Summing up the contributions from all parts of the discrete spectrum, we see that

$$H^3(\overline{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b}) \cong \mathsf{S}^{\mathrm{gen}}_{a-b,b+3} + s_{a+b+4}\mathsf{S}_{a-b+2} + \begin{cases} \mathsf{S}_{a+b+4} & \text{if } a = b \text{ odd,} \\ 0 & \text{otherwise.} \end{cases}$$

The cohomology vanishes outside the middle degree in all cases except when $a = b$ is even, when we have

$$H^2(\overline{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b}) \cong \begin{cases} s_{a+b+4}\mathbb{Q}_\ell(-b-1) & \text{if } a = b \text{ even,} \\ 0 & \text{otherwise,} \end{cases}$$

and $H^4(\bar{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b}) \cong H^2(\bar{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b})(-1)$.

Note that the sum

$$\mathsf{S}^{\mathrm{gen}}_{a-b,b+3} + \begin{cases} \mathsf{S}_{a+b+4} & \text{if } a = b \text{ odd,} \\ 0 & \text{otherwise,} \end{cases}$$

is exactly what was denoted $\bar{\mathsf{S}}_{a-b,b+3}$ in Theorem 2.1, since there are no Yoshida-type liftings in our case.

If we wish to determine in addition the cuspidal cohomology, then we need to understand which of the above representations are in the residual spectrum. The residual spectrum of GSp(4) is completely described in [Kim 2001, Section 7]. We see that there is exactly one case above where the representation is residual: namely, the Langlands quotient of $\mathrm{Ind}_{P(\mathbb{A})}^{\mathrm{PGSp}(4,\mathbb{A})}(\Pi \otimes 1)$ is residual if and only if $L\big(\Pi, \frac{1}{2}\big)$ is nonzero. We deduce that the cuspidal cohomology coincides with the L^2-cohomology except in degrees 2 and 4 when $a = b$ is even, where we have

$$H^2_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b}) \cong s'_{a+b+4}\mathbb{Q}(-b-1)$$

(so, conjecturally, it vanishes) and similarly for H^4_{cusp}. We also observe that, for all packets, either all discrete representations are cuspidal or all are residual, so that the cuspidal cohomology makes sense also as a summand of the étale intersection cohomology (a priori it is only a summand in the L^2-cohomology), and we can talk about the Galois representation on the cuspidal cohomology.

The Eisenstein cohomology of any local system on $\mathcal{A}_2$ has been completely determined in any degree, considered as an ℓ-adic Galois representation up to semisimplification, in [Harder 2012]. From that paper and the above discussion we may deduce the following:

Proposition 4.2. *The natural map $H^\bullet_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b}) \to H^\bullet_!(\mathcal{A}_2, \mathbb{V}_{a,b})$ is an isomorphism for any a, b.*

Proof. Recall that one has

$$H^k(\bar{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b}) = H^k_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b}) \oplus H^k_{\mathrm{res}}(\mathcal{A}_2, \mathbb{V}_{a,b})$$

and

$$W_{k+a+b}H^k(\mathcal{A}_2, \mathbb{V}_{a,b}) = H^k_!(\mathcal{A}_2, \mathbb{V}_{a,b}) \oplus W_{k+a+b}H^k_{\mathrm{Eis}}(\mathcal{A}_2, \mathbb{V}_{a,b}).$$

Moreover, the map $H^k(\bar{\mathcal{A}}_2, j_{!*}\mathbb{V}_{a,b}) \to W_{k+a+b}H^k(\mathcal{A}_2, \mathbb{V}_{a,b})$ is surjective and maps the cuspidal cohomology into the inner cohomology. Hence, if $H^k_{\mathrm{res}}(\mathcal{A}_2, \mathbb{V}_{a,b})$ and $W_{k+a+b}H^k_{\mathrm{Eis}}(\mathcal{A}_2, \mathbb{V}_{a,b})$ have the same dimension, $H^k_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b}) \to H^k_!(\mathcal{A}_2, \mathbb{V}_{a,b})$ is an isomorphism.

We have seen that $H^k_{\mathrm{res}}(\mathcal{A}_2, \mathbb{V}_{a,b})$ is nonzero only for $k \in \{2, 4\}$ and $a = b$ even, so these are the only cases where $H^\bullet_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b}) \to H^\bullet_!(\mathcal{A}_2, \mathbb{V}_{a,b})$ is not automatically an isomorphism. The dimension of $H^k_{\mathrm{res}}(\mathcal{A}_2, \mathbb{V}_{a,b})$ is $s_{a+b+4} - s'_{a+b+4}$

in these cases. From [loc. cit.] we see that $W_{k+a+b}H^k_{\mathrm{Eis}}(\mathcal{A}_2, \mathbb{V}_{a,b}) \neq 0$ only for $k = 2$ and $a = b$ even, in which case its dimension, too, is $s_{a+b+4} - s'_{a+b+4}$. But then $H^2_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b}) \to H^2_!(\mathcal{A}_2, \mathbb{V}_{a,b})$ is an isomorphism by the preceding paragraph, and then it is an isomorphism also in degree 4 since both the cuspidal and the inner cohomology satisfy Poincaré duality. □

Remark 4.3. The equality of dimensions above is not surprising, since Harder explicitly constructs these pure Eisenstein cohomology classes as residues of Eisenstein series associated to cusp forms for SL(2, $\mathbb{Z}$) with nonvanishing central value. So, in a sense, the dimension argument in the preceding theorem is unnecessarily convoluted. See also [Schwermer 1995], which describes in general all possible contributions from the residual spectrum to the Eisenstein cohomology of a Siegel threefold.

The main theorem of the paper follows from this result, as we now explain.

Proof of Theorem 2.1. Up to semisimplification we have

$$H^\bullet_c(\mathcal{A}_2, \mathbb{V}_{a,b}) = H^\bullet_!(\mathcal{A}_2, \mathbb{V}_{a,b}) \oplus H^\bullet_{c,\mathrm{Eis}}(\mathcal{A}_2, \mathbb{V}_{a,b}),$$

and $H^\bullet_{c,\mathrm{Eis}}(\mathcal{A}_2, \mathbb{V}_{a,b})$ was — as already remarked — determined in [Harder 2012]. By the preceding proposition we have $H^\bullet_!(\mathcal{A}_2, \mathbb{V}_{a,b}) = H^\bullet_{\mathrm{cusp}}(\mathcal{A}_2, \mathbb{V}_{a,b})$, and the latter has been determined already in this section. Summing up the Eisenstein cohomology and the cuspidal contribution gives the result. □

Acknowledgements

I am grateful to Jonas Bergström for many useful discussions on these topics and for his interest in this work, and Tomoyoshi Ibukiyama for several helpful pointers to the literature.

References

[Arakawa 1983] T. Arakawa, "Vector-valued Siegel's modular forms of degree two and the associated Andrianov L-functions", *Manuscripta Math.* **44**:1-3 (1983), 155–185. MR 84j:10030 Zbl 0517.10024

[Arthur 1996] J. Arthur, "L^2-cohomology and automorphic representations", pp. 1–17 in *Canadian Mathematical Society 1945–1995, vol. 3: Invited papers*, edited by J. B. Carrell and M. R. Murty, Canadian Mathematical Society, Ottawa, ON, 1996. MR 2000a:22026 Zbl 1205.22010

[Arthur 2002] J. Arthur, "A note on the automorphic Langlands group", *Canad. Math. Bull.* **45**:4 (2002), 466–482. MR 2004a:11120 Zbl 1031.11066

[Arthur 2004] J. Arthur, "Automorphic representations of GSp(4)", pp. 65–81 in *Contributions to automorphic forms, geometry, and number theory* (Baltimore, MD, 2002), edited by H. Hida et al., Johns Hopkins University Press, Baltimore, MD, 2004. MR 2005d:11074 Zbl 1080.11037

[Asgari and Schmidt 2001] M. Asgari and R. Schmidt, "Siegel modular forms and representations", *Manuscripta Math.* **104**:2 (2001), 173–200. MR 2002a:11044 Zbl 0987.11037

[Bergström 2009] J. Bergström, "Equivariant counts of points of the moduli spaces of pointed hyperelliptic curves", *Doc. Math.* **14** (2009), 259–296. MR 2011c:14083 Zbl 1211.14030

[Bergström et al. 2014] J. Bergström, C. Faber, and G. van der Geer, "Siegel modular forms of degree three and the cohomology of local systems", *Selecta Math.* (*N.S.*) **20**:1 (2014), 83–124. MR 3147414 Zbl 06261572

[Blasius and Rogawski 1994] D. Blasius and J. D. Rogawski, "Zeta functions of Shimura varieties", pp. 525–571 in *Motives* (Seattle, WA, 1991), vol. 2, edited by U. Jannsen et al., Proceedings of Symposia in Pure Mathematics **55**, American Mathematical Society, Providence, RI, 1994. MR 95e:11051 Zbl 0827.11033

[Borel 1979] A. Borel, "Automorphic L-functions", pp. 27–61 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 2, edited by A. Borel and W. Casselman, Proceedings of Symposia in Pure Mathematics **33**, American Mathematical Society, Providence, RI, 1979. MR 81m:10056 Zbl 0412.10017

[Borel 1981] A. Borel, "Stable real cohomology of arithmetic groups, II", pp. 21–55 in *Manifolds and Lie groups* (Notre Dame, IN, 1980), edited by J. Hano et al., Progress in Mathematics **14**, Birkhäuser, Boston, 1981. MR 83h:22023 Zbl 0483.57026

[Borel and Casselman 1983] A. Borel and W. Casselman, "L^2-cohomology of locally symmetric manifolds of finite volume", *Duke Math. J.* **50**:3 (1983), 625–647. MR 86j:22015 Zbl 0528.22012

[Borel and Jacquet 1979] A. Borel and H. Jacquet, "Automorphic forms and automorphic representations", pp. 189–207 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 1, edited by A. Borel and W. Casselman, Proceedings of Symposia in Pure Mathematics **33**, American Mathematical Society, Providence, RI, 1979. MR 81m:10055 Zbl 0414.22020

[Buzzard and Gee 2014] K. Buzzard and T. Gee, "The conjectural connections between automorphic representations and Galois representations", pp. 135–187 in *Automorphic forms and Galois representations*, vol. 1, edited by F. Diamond et al., London Mathematical Society Lecture Note Series **414**, Cambridge University Press, 2014.

[Cogdell et al. 2004] J. W. Cogdell, H. H. Kim, and M. Ram Murty, *Lectures on automorphic L-functions*, Fields Institute Monographs **20**, American Mathematical Society, Providence, RI, 2004. MR 2005h:11104 Zbl 1066.11021

[Conrey and Farmer 1999] J. B. Conrey and D. W. Farmer, "Hecke operators and the nonvanishing of L-functions", pp. 143–150 in *Topics in number theory* (University Park, PA, 1997), edited by S. D. Ahlgren et al., Mathematics and its Applications **467**, Kluwer, Dordrecht, 1999. MR 2000f:11055 Zbl 0943.11029

[Deligne 1971] P. Deligne, "Formes modulaires et représentations l-adiques", pp. 139–172 in *Séminaire Bourbaki. Vol. 1968/69: Exposé 355*, Lecture Notes in Mathematics **175**, Springer, Berlin, 1971. MR 3077124 Zbl 0206.49901

[Deligne 1979] P. Deligne, "Variétés de Shimura: interprétation modulaire, et techniques de construction de modèles canoniques", pp. 247–289 in *Automorphic forms, representations and L-functions* (Corvallis, OR, 1977), vol. 2, edited by A. Borel and W. Casselman, Proceedings of Symposia in Pure Mathematics **33**, American Mathematical Society, Providence, RI, 1979. MR 81i:10032 Zbl 0437.14012

[Faber and van der Geer 2004] C. Faber and G. van der Geer, "Sur la cohomologie des systèmes locaux sur les espaces de modules des courbes de genre 2 et des surfaces abéliennes, I", *C. R. Math. Acad. Sci. Paris* **338**:5 (2004), 381–384. MR 2005a:14033a Zbl 1062.14034

[Faltings and Chai 1990] G. Faltings and C.-L. Chai, *Degeneration of abelian varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) **22**, Springer, Berlin, 1990. MR 92d:14036 Zbl 0744.14031

[Flicker 2005] Y. Z. Flicker, *Automorphic forms and Shimura varieties of* PGSp(2), World Scientific, Hackensack, NJ, 2005. MR 2006g:11120 Zbl 1103.11018

[Flicker and Kazhdan 1988] Y. Z. Flicker and D. A. Kazhdan, "A simple trace formula", *J. Analyse Math.* **50** (1988), 189–200. MR 90e:11078 Zbl 0666.10018

[Fulton and Harris 1991] W. Fulton and J. Harris, *Representation theory: a first course*, Graduate Texts in Mathematics **129**, Springer, New York, 1991. MR 93a:20069 Zbl 0744.22001

[Getzler 1998] E. Getzler, "Topological recursion relations in genus 2", pp. 73–106 in *Integrable systems and algebraic geometry* (Kobe/Kyoto, 1997), edited by M.-H. Saito et al., World Scientific, River Edge, NJ, 1998. MR 2000b:14028 Zbl 1021.81056

[Ghitza and McAndrew 2012] A. Ghitza and A. McAndrew, "Experimental evidence for Maeda's conjecture on modular forms", *Tbil. Math. J.* **5**:2 (2012), 55–69. MR 3055515 Zbl 1280.11023

[Gross 1994] B. H. Gross, "L-functions at the central critical point", pp. 527–535 in *Motives* (Seattle, WA, 1991), vol. 1, edited by U. Jannsen et al., Proceedings of Symposia in Pure Mathematics **55**, American Mathematical Society, Providence, RI, 1994. MR 95a:11060 Zbl 0807.14015

[Harder 1993] G. Harder, *Eisensteinkohomologie und die Konstruktion gemischter Motive*, Lecture Notes in Mathematics **1562**, Springer, Berlin, 1993. MR 95g:11043 Zbl 0795.11024

[Harder 2012] G. Harder, "The Eisenstein motive for the cohomology of $\mathrm{GSp}_2(\mathbb{Z})$", pp. 143–164 in *Geometry and arithmetic* (Island of Schiermonnikoog, 2010), edited by C. Faber et al., European Mathematical Society, Zürich, 2012. MR 2987659

[Ibukiyama 2007a] T. Ibukiyama, "Dimension formulas of Siegel modular forms of weight 3 and supersingular abelian varieties", pp. 39–60 in *Siegel modular forms and abelian varieties: proceedings of the 4th Spring Conference on Modular Forms and Related Topics* (Hamana Lake, 2007), edited by T. Ibukiyama, Ryushido, Kobe, 2007.

[Ibukiyama 2007b] T. Ibukiyama, "Siegel modular forms of weight three and conjectural correspondence of Shimura type and Langlands type", pp. 55–69 in *The Conference on L-Functions* (Fukuoka, 2006), edited by L. Weng and M. Kaneko, World Scientific, Hackensack, NJ, 2007. MR 2008d:11044 Zbl 1183.11025

[Ibukiyama and Wakatsuki 2009] T. Ibukiyama and S. Wakatsuki, "Siegel modular forms of small weight and the Witt operator", pp. 189–209 in *Quadratic forms: algebra, arithmetic, and geometry*, edited by R. Baeza et al., Contemporary Mathematics **493**, American Mathematical Society, Providence, RI, 2009. MR 2010f:11072 Zbl 1244.11049

[Kim 2001] H. H. Kim, "Residual spectrum of odd orthogonal groups", *Internat. Math. Res. Notices* **2001**:17 (2001), 873–906. MR 2002k:11071 Zbl 1035.11020

[Laumon 1997] G. Laumon, "Sur la cohomologie à supports compacts des variétés de Shimura pour $\mathrm{GSp}(4)_{\mathbb{Q}}$", *Compositio Math.* **105**:3 (1997), 267–359. MR 2000a:11097 Zbl 0877.11037

[Laumon 2005] G. Laumon, "Fonctions zêtas des variétés de Siegel de dimension trois", pp. 1–66 in *Formes automorphes, II: Le cas du groupe* GSp(4), edited by J. Tilouine et al., Astérisque **302**, Société Mathématique de France, Paris, 2005. MR 2008e:22019 Zbl 1097.11021

[Looijenga 1988] E. Looijenga, "L^2-cohomology of locally symmetric varieties", *Compositio Math.* **67**:1 (1988), 3–20. MR 90a:32044 Zbl 0658.14010

[Petersen 2013] D. Petersen, "The tautological ring of the space of pointed genus two curves of compact type", preprint, 2013. arXiv 1310.7369

[Petersen and Tommasi 2014] D. Petersen and O. Tommasi, "The Gorenstein conjecture fails for the tautological ring of $\overline{\mathcal{M}}_{2,n}$", *Invent. Math.* **196**:1 (2014), 139–161. MR 3179574 Zbl 1295.14030

[Saper and Stern 1990] L. Saper and M. Stern, "L_2-cohomology of arithmetic varieties", *Ann. of Math.* (2) **132**:1 (1990), 1–69. MR 91m:14027 Zbl 0722.14009

[Schwermer 1995] J. Schwermer, "On Euler products and residual Eisenstein cohomology classes for Siegel modular varieties", *Forum Math.* **7**:1 (1995), 1–28. MR 96d:11062 Zbl 0829.11031

[Taïbi 2014] O. Taïbi, "Dimensions of spaces of level one automorphic forms for split classical groups using the trace formula", preprint, 2014. arXiv 1406.4247

[Taylor 1993] R. Taylor, "On the l-adic cohomology of Siegel threefolds", *Invent. Math.* **114**:2 (1993), 289–310. MR 95j:11054 Zbl 0810.11034

[Tehrani 2013] S. S. Tehrani, "On the strict endoscopic part of modular Siegel threefolds", preprint, 2013. arXiv 1305.4313

[Tilouine 2009] J. Tilouine, "Cohomologie des variétés de Siegel et représentations Galoisiennes associées aux représentations cuspidales cohomologiques de $\mathrm{GSp}_4(\mathbb{Q})$", pp. 99–114 in *Acte du coloque "Cohomologie l-adique et corps de nombres"* (CIRM, 2007), Publ. Math. Besançon: Algèbre et théorie des nombres 2007–2009, 2009. MR 2010i:11078 Zbl 1219.11081

[Tsushima 1983] R. Tsushima, "An explicit dimension formula for the spaces of generalized automorphic forms with respect to $\mathrm{Sp}(2, \mathbb{Z})$", *Proc. Japan Acad. Ser. A Math. Sci.* **59**:4 (1983), 139–142. MR 85a:11011 Zbl 0513.10025

[Vogan and Zuckerman 1984] D. A. Vogan, Jr. and G. J. Zuckerman, "Unitary representations with nonzero cohomology", *Compositio Math.* **53**:1 (1984), 51–90. MR 86k:22040 Zbl 0692.22008

[Weissauer 2005] R. Weissauer, "Four dimensional Galois representations", pp. 67–150 in *Formes automorphes, II: Le cas du groupe* $GSp(4)$, edited by J. Tilouine et al., Astérisque **302**, Société Mathématique de France, Paris, 2005. MR 2007f:11057 Zbl 1097.11027

[Weissauer 2009a] R. Weissauer, *Endoscopy for* $\mathrm{GSp}(4)$ *and the cohomology of Siegel modular threefolds*, Lecture Notes in Mathematics **1968**, Springer, Berlin, 2009. MR 2010h:11086 Zbl 1273.11089

[Weissauer 2009b] R. Weissauer, "The trace of Hecke operators on the space of classical holomorphic Siegel modular forms of genus two", preprint, 2009. arXiv 0909.1744

Received December 20, 2013. Revised October 16, 2014.

DAN PETERSEN
DEPARTEMENT MATHEMATIK
ETH ZÜRICH
RÄMISTRASSE 101
CH-8092 ZURICH
SWITZERLAND

pdan@math.ethz.ch

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.63

ON CERTAIN DUAL q-INTEGRAL EQUATIONS

OLA A. ASHOUR, MOURAD E. H. ISMAIL AND ZEINAB S. MANSOUR

We consider three different systems of dual q-integral equations where the kernel is the third Jackson q-Bessel functions. We solve the first system by applying the multiplying factor method (ansatz solution) and the second by employing the fractional q-calculus, and we use the q-Mellin transform to reduce the third system to a Fredholm q-integral equation of the second kind. Examples are included.

1. Introduction

Dual integral equations arise in a natural way while solving certain mixed boundary value problems. See [Sneddon 1966; Sneddon and Lowengrub 1969; Titchmarsh 1986]. Many of the dual integral equations are of the form

$$\int_0^\infty w(u)A(u)K(u,x)\,du=\lambda(x),\qquad 0<x<a,$$
$$\int_0^\infty A(u)K(u,x)\,du=\mu(x),\qquad a<x<\infty,$$

where $w(u)$ is the weight function, $K(x,u)$ is the kernel function. Several authors have described various methods to solve dual integral equations, especially when the kernel is a Bessel function. Busbridge [1938], Tranter [1951], Noble [1955; 1963], Sneddon [1960], Copson [1961], Peters [1961], Williams [1961], Erdélyi and Sneddon [1962], Nasim [1986] and others have described different methods to solve

MSC2010: primary 45F10; secondary 34B30, 31B10, 26A33, 35R11.

Keywords: q-Fourier transform, fractional q-integral operator, q-special functions, q-dual integral equations.

dual integral equations. An account of these methods is given in the introduction of [Erdélyi and Sneddon 1962] and, at greater length, in [Sneddon 1966, Chapter IV].

We now briefly mention three methods whose q-analogs will be treated in this work. The first approach, developed by Noble [1955] and Copson [1961], is the multiplying factor method. This approach involves the application of a certain multiplying factor, and after some manipulations we can solve the dual integral equations. The second approach uses fractional calculus to solve dual integral equations, and was developed by Erdélyi and Kober [1940] and Erdélyi [1951]. Their technique became a standard tool for solving dual integral equations. For example, see [Erdélyi and Sneddon 1962; Love 1963; Kesarwani 1967]. Finally, the third approach uses the Mellin transform to reduce the dual integral equations to a Fredholm equation of the second kind which can then be solved numerically. See [Williams 1961; Nasim 1986; Titchmarsh 1986].

In this paper, we are interested in solving dual q-integral equations when the kernel is the third Jackson q-Bessel function defined in (2-14) below and the q-integral is Jackson's q-integral. This paper is organized as follows. Section 2 includes the main notions and terminology from q-analysis which we need in our investigations. It also includes some q-integrals involving the third Jackson q-Bessel function. Section 3 includes the fractional q-integral operators and their calculus, which we need in our analysis. In Section 4, we apply the multiplying factor method to solve certain dual q-integral equations. In Section 5, we solve certain dual q-integral equations by using the fractional q-calculus method. In the last section, Section 6, we solve dual q-integral equations by using the q-Mellin transform introduced in [Fitouhi et al. 2006].

2. Preliminaries; q-notation

In the following, q is a positive number which is less than one. Let A_q, B_q, and $\mathbb{R}_{q,+}$ be the sets defined by

$$A_q := \{q^m : m \in \mathbb{N}_0\}, \quad B_q := \{q^{-m} : m \in \mathbb{N}\}, \quad \mathbb{R}_{q,+} := A_q \cup B_q, \tag{2-1}$$

where $\mathbb{N}_0 = \{0, 1, \dots\}$ and $\mathbb{N} = \{1, 2, \dots\}$. We introduce some of the needed q-notation and results. The q-shifted factorial, see [Gasper and Rahman 2004], and the multiple q-shifted factorial are defined by

$$\begin{gathered}(a;q)_0 := 1, \quad (a;q)_n := \prod_{k=0}^{n-1}(1-aq^k),\\ (a_1, a_2, \dots, a_k; q)_n := \prod_{j=1}^{k}(a_j;q)_n.\end{gathered} \tag{2-2}$$

The limit $\lim_{n\to\infty}(a;q)_n$ exists and is denoted by $(a;q)_\infty$. For $\gamma\in\mathbb{C}$, $aq^\gamma\neq q^{-n}$, $n\in\mathbb{N}$, we define $(a;q)_\gamma$ to be

$$(a;q)_\gamma := \frac{(a;q)_\infty}{(aq^\gamma;q)_\infty}. \tag{2-3}$$

The q-hypergeometric series (or basic hypergeometric series) ${}_r\phi_s$ is defined by

$$\begin{aligned}&{}_r\phi_s(a_1,a_2,\dots,a_r;b_1,b_2,\dots,b_s;q,z)\\&\qquad=\sum_{n=0}^{\infty}\frac{(a_1,a_2,\dots,a_r;q)_n}{(q,b_1,b_2,\dots,b_s;q)_n}z^n(-q^{(n-1)/2})^{n(s-r+1)}.\end{aligned}\tag{2-4}$$

The series representation for ${}_r\phi_s$ converges absolutely for all $z\in\mathbb{C}$ if $r\le s$ and converges only for $|z|<1$ if $r=s+1$.

Lemma 2.1 [Koornwinder and Swarttouw 1992]. *If $|z|<1$, then for $m,n\in\mathbb{Z}$,*

$$\begin{aligned}\sum_{k=-\infty}^{\infty} z^{k+n}\frac{(q^{n+k+1};q)_\infty}{(q;q)_\infty}{}_1\phi_1(0;q^{n+k+1};q,z^2)&\\ \times z^{k+m}\frac{(q^{m+k+1};q)_\infty}{(q;q)_\infty}{}_1\phi_1(0;q^{m+k+1};q,z^2)&=\delta_{nm}.\end{aligned}$$

If $\mu\in\mathbb{R}$, a subset A of $\mathbb{R}$ is called a μ-geometric set if $\mu z\in A$ for all $z\in A$. Let f be a function, real- or complex-valued, defined on a q-geometric set A. The q-difference operator is defined by

$$D_qf(z):=\frac{f(z)-f(qz)}{z-qz},\quad z\in A\setminus\{0\}.\tag{2-5}$$

If $0\in A$, the q-derivative at zero is defined by

$$D_qf(0):=\lim_{n\to\infty}\frac{f(zq^n)-f(0)}{zq^n},\quad z\in A\setminus\{0\},$$

if the limit exists and does not depend on z. See [Annaby and Mansour 2012]. The nonsymmetric q-product rule is

$$D_q(fg)(x)=g(x)D_qf(x)+f(qx)D_qg(x).\tag{2-6}$$

A right inverse to D_q, the Jackson q-integration [Jackson 1910], is

$$\int_0^z f(t)\,d_qt:=z(1-q)\sum_{n=0}^{\infty}q^nf(zq^n),\quad z\in A,\tag{2-7}$$

provided that the series converges, and

$$\int_a^b f(t)\,d_qt:=\int_0^b f(t)\,d_qt-\int_0^a f(t)\,d_qt,\quad a,b\in A.$$

If A is q^{-1}-geometric, then the q-integration over $[z, \infty)$, $z \in A$, is defined by

$$\int_z^\infty f(t)\, d_qt := \sum_{n=1}^{\infty} zq^{-n}(1-q)f(zq^{-n}), \tag{2-8}$$

and defined on $(0, \infty)$ by

$$\int_0^\infty f(t)\, d_qt := \sum_{n=-\infty}^{\infty} q^n(1-q)f(q^n). \tag{2-9}$$

The q-integration by parts rule is

$$\int_0^a f(qt)D_qg(t)\, d_qt = f(a)g(a) - \lim_{n\to\infty} f(q^n)g(q^n) - \int_0^a D_qf(t)g(t)\, d_qt. \tag{2-10}$$

For $\eta \in \mathbb{C}$ and a function f defined on $\mathbb{R}_{q,+}$, we define the spaces

$$\begin{aligned} L_{q,\eta}(\mathbb{R}_{q,+}) &:= \left\{ f : \|f\|_{q,\eta} := \int_0^\infty |t^\eta f(t)|\, d_qt < \infty \right\}, \\ L_{q,\eta}(A_q) &:= \left\{ f : \|f\|_{A_q,\eta} := \int_0^1 |t^\eta f(t)|\, d_qt < \infty \right\}, \\ L_{q,\eta}(B_q) &:= \left\{ f : \|f\|_{B_q,\eta} := \int_1^\infty |t^\eta f(t)|\, d_qt < \infty \right\}, \end{aligned}$$

and

$$L_q(C) := L_{q,0}(C), \quad C \in \{A_q, B_q, \mathbb{R}_{q,+}\}.$$

Clearly, $L_{q,\eta}(\mathbb{R}_{q,+}) = L_{q,\eta}(A_q) \cap L_{q,\eta}(B_q)$.

Lemma 2.2. *For $\alpha \in \mathbb{C}$, we have*

$$\sum_{k=0}^{n} q^{2k\alpha} \frac{(q^{2\alpha}; q^2)_{n-k}(q^{2k+2}; q^2)_{n-k}}{(q^2; q^2)_{n-k}} = 1. \tag{2-11}$$

Proof. The left-hand side of (2-11) is

$$\begin{aligned} &(q^2; q^2)_n \sum_{k=0}^{n} \frac{q^{2k\alpha}}{(q^2; q^2)_k} \frac{(q^{2\alpha}; q^2)_{n-k}}{(q^2; q^2)_{n-k}} \\ &\qquad = (q^{2\alpha}; q^2)_n \lim_{\varepsilon\to 0} {}_2\phi_1(q^{-2n}, \varepsilon; q^{2-2n-2\alpha}; q^2, q^2) \\ &\qquad = (q^{2\alpha}; q^2)_n \lim_{\varepsilon\to 0} \frac{(q^{2-2n-2\alpha}/\varepsilon; q^2)_n}{(q^{2-2n-2\alpha}; q^2)_n} \varepsilon^n = \lim_{\varepsilon\to 0} (q^{2\alpha}\varepsilon; q^2)_n = 1; \end{aligned}$$

here we used [Gasper and Rahman 2004, (I.9), (II.6)]. This completes the proof. □

The q-gamma function [Jackson 1904; Gasper and Rahman 2004] is defined by

$$\Gamma_q(z) := \frac{(q;q)_\infty}{(q^z;q)_\infty}(1-q)^{1-z}, \quad z \in \mathbb{C},\ |q|<1, \tag{2-12}$$

where we take the principal values of q^z and $(1-q)^{1-z}$. The q-binomial theorem (see [Andrews et al. 1999, p. 488]) takes the form

$$\sum_{n=0}^{\infty} \frac{(a;q)_n}{(q;q)_n} z^n = \frac{(az;q)_\infty}{(z;q)_\infty}, \quad |z|<1. \tag{2-13}$$

The third Jackson q-Bessel function $J_\nu^{(3)}(z;q)$ for $z \in \mathbb{C}$, (see [Jackson 1905; Ismail 2005]) is defined by

$$\begin{aligned} J_\nu(z;q) = J_\nu^{(3)}(z;q) &:= \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty} z^\nu {}_1\phi_1(0;q^{\nu+1};q,qz^2) \\ &= \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty} z^\nu \sum_{n=0}^{\infty} (-1)^n \frac{q^{n(n+1)/2} z^{2n}}{(q;q)_n (q^{\nu+1};q)_n}, \end{aligned} \tag{2-14}$$

and satisfies

$$D_q[(\,\cdot\,)^{-\nu} J_\nu^{(3)}(\,\cdot\,;q^2)](z) = -\frac{q^{1-\nu} z^{-\nu}}{1-q} J_{\nu+1}^{(3)}(qz;q^2), \tag{2-15}$$

$$D_q[(\,\cdot\,)^{\nu} J_\nu^{(3)}(\,\cdot\,;q^2)](z) = \frac{z^{\nu}}{1-q} J_{\nu-1}^{(3)}(z;q^2); \tag{2-16}$$

see [Koornwinder and Swarttouw 1992; Swarttouw 1992]. The q-Bessel function $J_\nu(\,\cdot\,;q^2)$, $\nu > -1$, satisfies

$$|J_\nu(q^n;q^2)| \le \frac{(-q^2;q^2)_\infty(-q^{2\nu+2};q^2)_\infty}{(q^2;q^2)_\infty} \begin{cases} q^{n\nu} & \text{if } n \ge 0, \\ q^{n^2-(\nu+1)n} & \text{if } n < 0. \end{cases} \tag{2-17}$$

See [Koelink 1994]. The following identity, which was introduced by Koornwinder and Swarttouw [1992], is useful in our investigations.

$$\frac{(q^{\alpha-t+1};q^2)_\infty}{(q^{\alpha+t+1};q^2)_\infty} = \sum_{-\infty}^{\infty} q^{n(t+1)} J_\alpha(q^n;q^2), \tag{2-18}$$

where t and α are complex numbers such that $\Re(t) > -\Re(\alpha) - 1$. We recall that the functions $\cos(z;q)$ and $\sin(z;q)$ are defined, for $z \in \mathbb{C}$, by

$$\begin{aligned} \cos(z;q) &:= \frac{(q^2;q^2)_\infty}{(q;q^2)_\infty}(zq^{-1/2}(1-q))^{1/2} J_{-1/2}(z(1-q)/\sqrt{q}\,;q^2), \\ \sin(z;q) &:= \frac{(q^2;q^2)_\infty}{(q;q^2)_\infty}(z(1-q))^{1/2} J_{1/2}(z(1-q);\,q^2). \end{aligned}$$

Proposition 2.3 [Koelink and Swarttouw 1994, p. 694]. *For* $\Re(\nu) > -1$, $x > 0$, *and* $a, b \in \mathbb{C} \setminus \{0\}$, *we have*

$$(a^2 - b^2) \int_0^x t J_\nu(aqt; q^2) J_\nu(bqt; q^2)\, d_q t \\ = (1-q) q^{\nu-1} x [a J_{\nu+1}(aqx; q^2) J_\nu(bx; q^2) - b J_\nu(ax; q^2) J_{\nu+1}(aqx; q^2)]. \tag{2-19}$$

Koornwinder and Swarttouw [1992] introduced the following inverse pair of q-integral transforms under the side condition $f, g \in L_q^2(\mathbb{R}_{q,+})$:

$$g(\lambda) = \int_0^\infty f(x) J_\nu(\lambda x; q^2) x \, d_q x, \qquad f(x) = \int_0^\infty g(\lambda) J_\nu(\lambda x; q^2) \lambda \, d_q \lambda, \tag{2-20}$$

where $\lambda, x \in \mathbb{R}_{q,+}$. This pair of q-integral transforms is a q-analog of the Hankel transform pair

$$g(\lambda) = \int_0^\infty f(x) J_\nu(\lambda x) x \, dx, \qquad f(x) = \int_0^\infty g(\lambda) J_\nu(\lambda x) \lambda \, d\lambda.$$

The following result is a discrete q-analog of the Weber–Schafheitlin integral.

Proposition 2.4 [Koornwinder and Swarttouw 1992, p. 455–456]. *Let* α, β, *and* γ *be complex numbers and* ξ, $\rho \in \mathbb{R}_{q,+}$. *Then*

$$\frac{1}{1-q} \int_0^\infty t^{-\gamma} J_\alpha(\xi t; q^2) J_\beta(\rho t; q^2)\, d_q t \\ = \rho^\beta \xi^{(\gamma-\beta-1)} \frac{(q^{\alpha-\beta+\gamma+1}, q^{2\beta+2}; q^2)_\infty}{(q^{\alpha+\beta-\gamma+1}, q^2; q^2)_\infty} \\ \times {}_2\phi_1\left(q^{\beta-\alpha-\gamma+1}, q^{\beta+\alpha-\gamma+1}; q^{2\beta+2}; q^2, \frac{\rho^2}{\xi^2} q^{-\beta+\alpha+\gamma+1} \right)$$

if

$$\Re(-\beta+\alpha+\gamma+1) \geq 0,\ \rho < \xi \quad \textit{or} \quad \Re(-\beta+\alpha+\gamma+1) > 0,\ \rho \leq \xi;$$

and

$$\frac{1}{1-q} \int_0^\infty t^{-\gamma} J_\alpha(\xi t; q^2) J_\beta(\rho t; q^2)\, d_q t \\ = \xi^\alpha \rho^{(\gamma-\alpha-1)} \frac{(q^{\beta-\alpha+\gamma+1}, q^{2\alpha+2}; q^2)_\infty}{(q^{\beta+\alpha-\gamma+1}, q^2; q^2)_\infty} \\ \times {}_2\phi_1\left(q^{\alpha-\beta-\gamma+1}, q^{\beta+\alpha-\gamma+1}; q^{2\alpha+2}; q^2, \frac{\xi^2}{\rho^2} q^{-\alpha+\beta+\gamma+1} \right)$$

if

$$\Re(-\alpha+\beta+\gamma+1) \geq 0,\ \xi < \rho \quad \textit{or} \quad \Re(-\alpha+\beta+\gamma+1) > 0,\ \xi \leq \rho.$$

Lemma 2.1 gives us the orthogonality relation

$$\int_0^\infty t J_\alpha(\xi t; q^2) J_\alpha(\rho t; q^2)\, d_q t = \frac{1-q}{\xi^2}\delta_{\rho,\xi}, \quad \Re(\alpha) > -1, \tag{2-21}$$

where ρ, ξ are in $\mathbb{R}_{q,+}$ and $\delta_{\rho,\xi}$ is the Kronecker delta. The following is a q-analog of the Sonine–Schafheitlin integral. If in Proposition 2.4 we take $\gamma = 1$ and $\alpha = \beta$, we obtain

$$\int_0^\infty t^{-1} J_\alpha(\xi t; q^2) J_\alpha(\rho\, t; q^2)\, d_q t = \begin{cases} (1-q)/(1-q^{2\alpha})(\rho/\xi)^\alpha & \text{if } \rho < \xi, \\ (1-q)/(1-q^{2\alpha})(\xi/\rho)^\alpha & \text{if } \xi < \rho. \end{cases}$$

Corollary 2.5. *Let* $\alpha, \beta \in \mathbb{C}$, $\rho, t \in \mathbb{R}_{q,+}$. *If* $\Re(\beta) > \Re(\alpha) > -1$, *then*

$$\int_0^\infty t^{\alpha-\beta+1} J_\alpha(\xi\, t; q^2) J_\beta(\rho\, t; q^2)\, d_q t \tag{2-22}$$
$$= \begin{cases} 0 & \textit{if } \xi > \rho, \\ \dfrac{(1-q)(1-q^2)^{1-\beta+\alpha}}{\Gamma_{q^2}(\beta-\alpha)} \xi^\alpha \rho^{\beta-2\alpha-2} (q^2\xi^2/\rho^2; q^2)_{\beta-\alpha-1} & \textit{if } \xi \le \rho. \end{cases}$$

Proof. This follows from Proposition 2.4 by taking $\gamma = \beta - \alpha - 1$. □

Corollary 2.6. *Let* m, n *be nonnegative integers and* $\nu > -n - m - k$. *Then*

$$\int_0^\infty t^{-1} J_{\nu+2n+k}(t; q^2) J_{\nu+2m+k}(t; q^2)\, d_q t = \begin{cases} 0 & \textit{if } m \neq n, \\ \dfrac{1-q}{1-q^{2\nu+4n+2k}} & \textit{if } m = n. \end{cases} \tag{2-23}$$

Proof. This result follows by applying Proposition 2.4 with $\gamma = 1$, $\alpha = \nu + 2m + k$, $\beta = \nu + 2n + k$, and $\xi = \rho = 1$. □

The little q-Jacobi polynomial [Gasper and Rahman 2004, p. 27] is defined by

$$p_n(x; a, b \mid q) := {}_2\phi_1(q^{-n}, abq^{n+1}; aq; q, qx).$$

Corollary 2.7. *Let* $\Re(\nu) > -1$, $\rho, t \in \mathbb{R}_{q,+}$. *Then*

$$\int_0^\infty t^{1-k} J_{\nu+2m+k}(t; q^2) J_\nu(\rho\, t; q^2)\, d_q t \tag{2-24}$$
$$= \begin{cases} 0 & \rho > 1, \\ \dfrac{(1-q)\rho^\nu (q^{2m+2}\rho^2, q^{2m+2k}, q^{2\nu+2}; q^2)_\infty}{(q^{2m+2k}\rho^2, q^{2\nu+2m+2}, q^2; q^2)_\infty} p_m(q^{2m}\rho^2; q^{2k-2}, q^{2\nu} \mid q^2) & \rho < 1. \end{cases}$$

Proof. This result follows by applying Proposition 2.4 with $\gamma = k-1$, $\alpha = \nu+2m+k$, $\beta = \nu$, and $\xi = 1$, in addition to the transformation

$${}_2\phi_1(a, b; c; q, z) = \frac{(abz/c; q)_\infty}{(z; q)_\infty} {}_2\phi_1(c/a, c/b; c; q, abz/c).$$ □

Corollary 2.8. *Let* $\Re(\nu) > -1$, $\rho, t \in \mathbb{R}_{q,+}$. *Then*

$$\begin{aligned}(2\text{-}25)\quad & J_{\nu+2m+k}(t;q^2)\\ &= \frac{(1-q)t^k(q^{2m+2k},q^{2\nu+2};q^2)_\infty}{(q^{2\nu+2m+2};q^2;q^2)_\infty}\\ &\quad\times\int_0^1 \rho^{\nu+1}(q^{2m+2}\rho^2;q^2)_{k-1}J_\nu(\rho t;q^2)p_m(q^{2m}\rho^2;q^{2k-2},q^{2\nu}\mid q^2)\,d_q\rho.\end{aligned}$$

Proof. This follows from Corollary 2.7 and the q-Hankel transform pair (2-20). □

Proposition 2.9. *Let x, u, and α be complex numbers. If* $\Re(\gamma+\beta) > -1$ *and* $\Re(\beta) > -1$, *then*

$$\begin{aligned}(2\text{-}26)\quad & \int_0^x t^\gamma(q^2t^2/x^2;q^2)_\alpha J_\beta(ut;q^2)\,d_qt\\ &= \frac{x^{\gamma+\beta+1}u^\beta(1-q)(q^{2\beta+2},q^{2\alpha+\gamma+\beta+3};q^2)_\infty}{(q^{2\alpha+2},q^{\gamma+\beta+1};q^2)_\infty}\\ &\quad\times {}_2\phi_2(0,q^{\gamma+\beta+1};q^{2\beta+2},q^{2\alpha+\gamma+\beta+3};q^2,q^2x^2u^2).\end{aligned}$$

In particular, if $\gamma=\beta+1$, *then*

$$\begin{aligned}(2\text{-}27)\quad & \int_0^x t^{\beta+1}(q^2t^2/x^2;q^2)_\alpha J_\beta(ut;q^2)\,d_qt\\ &= x^{\beta-\alpha+1}u^{-\alpha-1}(1-q)(1-q^2)^\alpha\Gamma_{q^2}(\alpha+1)J_{\alpha+\beta+1}(xu;q^2).\end{aligned}$$

Proof. According to (2-14), we have

$$\begin{aligned}(2\text{-}28)\quad & \int_0^x t^\gamma(q^2t^2/x^2;q^2)_\alpha J_\beta(ut;q^2)\,d_qt\\ &= \frac{(q^{2\beta+2};q^2)_\infty}{(q^2;q^2)_\infty}\sum_{k=0}^\infty \frac{(-1)^kq^{k(k+1)}u^{2k+\beta}}{(q^2;q^2)_k(q^{2\beta+2};q^2)_k}\int_0^x t^{\gamma+\beta+2k}(q^2t^2/x^2;q^2)_\alpha\,d_qt.\end{aligned}$$

By using (2-3) and the q-binomial theorem with base q^2 instead of q on the inner series in (2-13), we obtain

$$\begin{aligned}(2\text{-}29)\quad & \int_0^x t^{\gamma+\beta+2k}(q^2t^2/x^2;q^2)_\alpha\,d_qt\\ &= \frac{(1-q)x^{2k+\gamma+\beta+1}(q^2;q^2)_\infty(q^{2k+2\alpha+\gamma+\beta+3};q^2)_\infty}{(q^{2\alpha+2};q^2)_\infty(q^{2k+\gamma+\beta+1};q^2)_\infty}.\end{aligned}$$

Substituting (2-29) in (2-28) and using (2-4), the desired result follows. The particular case follows by direct substitution in (2-26) and the definition (2-14). □

Proposition 2.10. *Let ν and α be complex numbers such that $\Re(\nu) > -1$. For $x, u \in \mathbb{R}_{q,+}$,*

$$\int_x^\infty (x^2/t^2; q^2)_{\alpha-1} t^{2\alpha-\nu-1} J_\nu(tu; q^2)\, d_qt = x^{\alpha-\nu} u^{-\alpha} q^\alpha (1-q) \frac{(q^2; q^2)_\infty}{(q^{2\alpha}; q^2)_\infty} J_{\nu-\alpha}(xu/q; q^2). \tag{2-30}$$

Proof. Using the ${}_1\phi_1$ transformation (see [Gasper and Rahman 2004, p. 29])

$$(c; q)_\infty {}_1\phi_1(0; a; q, z) = (z; q)_\infty {}_1\phi_1(0; z; q, c)$$

and (2-14), one can verify that

$$J_\nu(z; q^2) = \frac{(q^2z^2; q^2)_\infty}{(q^2; q^2)_\infty} z^\nu \sum_{j=0}^\infty (-1)^j q^{j(j+1)} \frac{q^{2\nu j}}{(q^2, q^2z^2; q^2)_j}.$$

Hence,

$$\int_x^\infty (x^2/t^2; q^2)_{\alpha-1} t^{2\alpha-\nu-1} J_\nu(tu; q^2)\, d_qt = \frac{u^\nu}{(q^2; q^2)_\infty} \sum_{j=0}^\infty \frac{q^{j^2+j+2\nu j}}{(q^2; q^2)_j} \int_x^\infty (x^2/t^2; q^2)_{\alpha-1} t^{2\alpha-1} (q^{2j+2}t^2u^2; q^2)_\infty\, d_qt. \tag{2-31}$$

Using Lemma 2.2, we can prove that

$$\int_x^\infty (x^2/t^2; q^2)_{\alpha-1} t^{2\alpha-1} (q^{2j+2}t^2u^2; q^2)_\infty\, d_qt = (1-q) u^{-2\alpha} \frac{(q^2; q^2)_\infty}{(q^{2\alpha}; q^2)_\infty} q^{-2\alpha j+2\alpha} (x^2u^2q^{2j}; q^2)_\infty \tag{2-32}$$

for $x, t \in \mathbb{R}_{q,+}$. Substituting (2-32) into (2-31) yields the desired result. □

3. Fractional q-calculus

In this section, we introduce fractional q-integral operators and their properties which we need in our fractional q-calculus approach for solving certain dual q-integral equations. A comprehensive study of the fractional q-calculus and equations is in [Samko et al. 1987; Butzer and Westphal 2000; Annaby and Mansour 2012]. Al-Salam [1966] defined a two-parameter q-fractional operator by

$$K_q^{\eta,\alpha}\phi(x) := \frac{q^{-\eta}x^\eta}{\Gamma_q(\alpha)} \int_x^\infty (x/t; q)_{\alpha-1} t^{-\eta-1} \phi(tq^{1-\alpha})\, d_qt,$$

where $\alpha \neq -1, -2, \dots$. This is a q-analog of the Erdélyi and Sneddon fractional operator (see [Erdélyi 1951; Erdélyi and Sneddon 1962])

$$K^{\eta,\alpha} f(x) = \frac{x^{\eta}}{\Gamma(\alpha)} \int_x^{\infty} (t-x)^{\alpha-1} t^{-\eta-1} f(t)\, dt.$$

The following operator is a slight modification of the operator $K_q^{\eta,\alpha}$, which we found very convenient in our analysis. This operator, denoted by $\mathcal{K}_q^{\eta,\alpha}$, is defined as

$$\mathcal{K}_q^{\eta,\alpha} \phi(x) := \frac{q^{-\eta} x^{\eta}}{\Gamma_q(\alpha)} \int_x^{\infty} (x/t; q)_{\alpha-1} t^{-\eta-1} \phi(qt)\, d_q t, \tag{3-1}$$

where $\alpha \neq -1, -2, \dots$. Using (2-8), the operator $\mathcal{K}_q^{\eta,\alpha}$ has the series representation

$$\mathcal{K}_q^{\eta,\alpha} \phi(x) = (1-q)^{\alpha} \sum_{n=0}^{\infty} q^{n\eta} \frac{(q^{\alpha}; q)_n}{(q; q)_n} \phi(xq^{-n}), \tag{3-2}$$

which is valid for all α.

Proposition 3.1. *Let $\eta \in \mathbb{C}$. If $\phi \in L_{q,-\eta-1}(B_q)$, then $\mathcal{K}_q^{\eta,\alpha}\phi(x)$ exists for all $x \in \mathbb{R}_{q,+}$ and belongs to $L_{q,\mu}(B_q)$ for any $\mu \in \mathbb{C}$ such that $\Re(\mu) < -\Re(\eta) - 1$.*

Proof. Let $m \in \mathbb{Z}$. Since

$$\mathcal{K}_q^{\eta,\alpha} \phi(q^m) = (1-q)^{\alpha} \sum_{n=0}^{\infty} q^{n\eta} \frac{(q^{\alpha}; q)_n}{(q; q)_n} \phi(q^{m-n}),$$

we obtain

$$|\mathcal{K}_q^{\eta,\alpha} \phi(q^m)| \le (1-q)^{\Re(\alpha)} \frac{(-q^{\Re(\alpha)}; q)_{\infty}}{(q; q)_{\infty}} \sum_{n=0}^{\infty} q^{n\Re(\eta)} |\phi(q^{m-n})|. \tag{3-3}$$

Therefore, if $m > 0$, we obtain

$$\begin{aligned} &|\mathcal{K}_q^{\eta,\alpha} \phi(q^m)| \\ &\quad \le (1-q)^{\Re(\alpha)} q^{m\Re(\eta)} \frac{(-q^{\Re(\alpha)}; q)_{\infty}}{(q; q)_{\infty}} \left(\sum_{k=0}^{m} q^{-k\Re(\eta)} |\phi(q^k)| + \frac{1}{1-q} \|\phi\|_{B_q,-1-\eta} \right). \end{aligned} \tag{3-4}$$

If $m < 0$, then

$$|\mathcal{K}_q^{\eta,\alpha} \phi(q^m)| \le (1-q)^{\Re(\alpha)-1} q^{m\Re(\eta)} \frac{(-q^{\Re(\alpha)}; q)_{\infty}}{(q; q)_{\infty}} \|\phi\|_{B_q,-1-\eta}. \tag{3-5}$$

Consequently, $K_q^{\eta,\alpha}(x)$ exists for all $x \in \mathbb{R}_{q,+}$. If $\mu \in \mathbb{C}$ and $\Re(\mu) < -\Re(\eta) - 1$, then

$$\begin{aligned} &\int_1^{\infty} |t^{\mu} \mathcal{K}_q^{\eta,\alpha} \phi(t)|\, d_q t \\ &\quad \le (1-q)^{\Re(\alpha)-1} \frac{(-q^{\Re(\alpha)}; q)_{\infty}}{(q; q)_{\infty}} \|\phi\|_{B_q,-1-\eta} \sum_{j=0}^{\infty} q^{-j\Re(1+\mu+\eta)} < \infty, \end{aligned} \tag{3-6}$$

where $\Re(\mu) < -\Re(\eta) - 1$. Thus $\mathcal{K}_q^{\eta,\alpha}\phi \in L_{q,\mu}(B_q)$. □

Al-Salam [1966, p. 138–139] proved formally the semigroup identity

$$K_q^{\eta,\alpha} K_q^{\eta+\alpha,\beta}\phi(x) = K_q^{\eta,\alpha+\beta}\phi(x), \tag{3-7}$$

where η, α, and β are complex numbers and without imposing any conditions on the function ϕ. Using the same technique we can prove that the semigroup property

$$\mathcal{K}_q^{\eta,\alpha}\mathcal{K}_q^{\eta+\alpha,\beta}\phi(x) = \mathcal{K}_q^{\eta,\alpha+\beta}\phi(x) \tag{3-8}$$

holds for $x \in \mathbb{R}_{q,+}$ whenever $\phi \in L_{q,-\eta-\alpha-1}(B_q)$, η, α, β are complex numbers and $\Re(\alpha) < 0$. Therefore, if we take $\beta = -\alpha$ in (3-8) and note that $\mathcal{K}_q^{\eta,0}$ is the identity operator, we obtain

$$(\mathcal{K}_q^{\eta,\alpha})^{-1}\phi(x) = \mathcal{K}_q^{\eta+\alpha,-\alpha}\phi(x), \quad x \in \mathbb{R}_{q,+}, \tag{3-9}$$

for any $\phi \in L_{q,-\eta-\alpha-1}(B_q)$, and η, α are complex numbers.

Agarwal [1969] defined the two-parameter family

$$I_q^{\eta,\alpha}\phi(x) := \frac{x^{-\eta-1}}{\Gamma_q(\alpha)}\int_0^x (qt/x;q)_{\alpha-1} t^{\eta}\phi(t)\,d_qt, \quad \alpha \neq 0, -1, -2, \dots. \tag{3-10}$$

of fractional q-integral operators, which can be written as

$$I_q^{\eta,\alpha}\phi(x) = (1-q)^{\alpha}\sum_{n=0}^{\infty} q^{(\eta+1)n}\frac{(q^{\alpha};q)_n}{(q;q)_n}\phi(xq^n), \tag{3-11}$$

which is valid for all α. The special case $I_q^{0,\alpha}$ is the q-analog of the Riemann–Liouville fractional operator introduced in [Al-Salam 1966] and is denoted by I_q^{α}. Hence,

$$I_q^{\alpha}\phi(x) := \frac{x^{\alpha-1}}{\Gamma_q(\alpha)}\int_0^x (qt/x;q)_{\alpha-1}\phi(t)\,d_qt, \quad \alpha \neq 0, -1, -2, \dots. \tag{3-12}$$

In [Annaby and Mansour 2012, p. 121], the authors solved the q-analog of the Abel integral equation on a continuous domain of the form $[0, a]$. In the following we state without proof a modified version of [ibid., Theorem 4.7] which holds when the domain of solution is discrete.

Theorem 3.2. *The q-Abel integral equation*

$$\frac{x^{\alpha-1}}{\Gamma_q(\alpha)}\int_0^x (qt/x;q)_{\alpha-1}\phi(t)\,d_qt = f(x) \quad (0 < \Re(\alpha) < 1,\ x \in A_q)$$

has a unique solution $\phi \in L_q(A_q)$, given by

$$\phi(x) = D_{q,x} I_q^{1-\alpha} f(x),$$

if and only if f and $I_q^{1-\alpha} f$ are $L_q(A_q)$ functions with $I_q^{1-\alpha} f(0) = 0$.

Proof. The proof is similar to the proof of [Annaby and Mansour 2012, Theorem 4.7] and is omitted. □

Proposition 3.3. *Let η and α be complex numbers. If $\phi \in L_{q,\eta}(A_q)$, then $I_q^{\eta,\alpha}\phi(x)$ exists for all $x \in \mathbb{R}_{q,+}$ and belongs to $L_{q,\mu}(A_q)$ for all $\mu \in \mathbb{C}$ such that $\Re(\mu-\eta) > 0$.*

Proof. Assume that $\phi \in L_{q,\eta}(A_q)$. Then

$$I_q^{\eta,\alpha}\phi(q^m) = (1-q)^\alpha q^{-(\eta+1)m} \sum_{n=m}^{\infty} q^{(\eta+1)n} \frac{(q^\alpha; q)_{n-m}}{(q;q)_{n-m}} \phi(q^n).$$

Thus, if $m \geq 0$, we obtain

$$|I_q^{\eta,\alpha}\phi(q^m)| \leq (1-q)^{\Re(\alpha)} q^{-(\Re(\eta)+1)m} \frac{(-q^{\Re(\alpha)}; q)_\infty}{(q;q)_\infty} \|\phi\|_{A_q,\eta};$$

and if $m < 0$, we obtain

$$|I_q^{\eta,\alpha}\phi(q^m)| \leq (1-q)^{\Re(\alpha)} q^{-(\Re(\eta)+1)m} \frac{(-q^{\Re(\alpha)}; q)_\infty}{(q;q)_\infty} \times \left(\sum_{k=m}^{-1} q^{k(\Re(\eta)+1)} |\phi(q^k)| + \frac{1}{1-q} \|\phi\|_{A_q,\eta} \right) < \infty.$$

Moreover,

$$\int_0^1 |t^\mu I_q^{\eta,\alpha}\phi(t)| \leq (1-q)^{\Re(\alpha)-1} \frac{(-q^{\Re(\alpha)}; q)_\infty}{(q;q)_\infty} \|\phi\|_{A_q,\eta} \sum_{j=0}^{\infty} q^{j\Re(\mu-\eta)} < \infty,$$

provided that $\Re(\mu-\eta) > 0$. This completes the proof. □

Proposition 3.4. *Let η and α be complex numbers, and $\Re(\alpha) < 0$. If $\phi \in L_{q,\eta+\alpha}(A_q)$ then*

$$(I_q^{\eta,\alpha})^{-1} = I_q^{\eta+\alpha,-\alpha}. \tag{3-13}$$

Proof. This follows by noting that

$$I_q^{\eta,\alpha} I_q^{\eta+\alpha,-\alpha}\phi(x) = \sum_{k=0}^{\infty} q^{(\eta+1)k} \frac{(q^\alpha; q)_k}{(q;q)_k} \sum_{m=0}^{\infty} q^{(\eta+1+\alpha)m} \frac{(q^{-\alpha}; q)_m}{(q;q)_m} \phi(q^{k+m}x). \tag{3-14}$$

Make the substitution $n = m + k$ on the inner series of (3-14). This gives

$$I_q^{\eta,\alpha} I_q^{\eta+\alpha,-\alpha}\phi(x) = \sum_{k=0}^{\infty} q^{-\alpha k} \frac{(q^\alpha; q)_k}{(q;q)_k} \sum_{n=k}^{\infty} q^{(\eta+1+\alpha)n} \frac{(q^{-\alpha}; q)_{n-k}}{(q;q)_{n-k}} \phi(q^n x). \tag{3-15}$$

If $\Re(\alpha) < 0$ and $\phi \in L_{q,\eta+\alpha}(A_q)$, then the double series in (3-15) is absolutely convergent and we can interchange the order of summation to obtain

$$\text{(3-16)}\quad \begin{aligned} I_q^{\eta,\alpha} I_q^{\eta+\alpha,-\alpha}\phi(x) &= \sum_{n=0}^{\infty} q^{(\eta+1+\alpha)n}\phi(q^n x)\sum_{k=0}^{n} q^{-\alpha k}\frac{(q^{\alpha};q)_k}{(q;q)_k}\frac{(q^{-\alpha};q)_{n-k}}{(q;q)_{n-k}} \\ &= \sum_{n=0}^{\infty} q^{(\eta+1+2\alpha)n}\phi(q^n x)\frac{(q^{-\alpha};q)_n}{(q;q)_n}\frac{(q^{1-n};q)_n}{(q^{1+\alpha-n};q)_n} \\ &= \phi(x), \end{aligned}$$

where we applied [Gasper and Rahman 2004, Equation (II.6)] to the inner series. □

A direct calculation gives

$$I_q^{\eta,\alpha} x^{\beta} f(x) = x^{\beta} I_q^{\eta+\beta,\alpha} f(x),$$

where η, α, β are complex numbers, and $f \in L_{q,\eta+\beta}(A_q)$. Agarwal [1969] also proved the following semigroup identity when η, λ, and μ are positive constants:

$$\text{(3-17)}\quad \begin{aligned} I_q^{\eta,\lambda} I_q^{\eta+\lambda,\mu}\phi(x) &= I_q^{\eta,\mu+\lambda}\phi(x) = I_q^{\eta+\lambda,\mu} I_q^{\eta,\lambda}\phi(x) \\ &= I_q^{\eta,\mu} I_q^{\mu+\eta,\lambda}\phi(x) = I_q^{\eta+\mu,\lambda} I_q^{\eta,\mu}\phi(x). \end{aligned}$$

But, using the same technique introduced in [ibid.], we can prove that

$$\text{(3-18)}\quad I_q^{\eta,\lambda} I_q^{\eta+\lambda,\mu}\phi(x) = I_q^{\eta,\mu+\lambda}\phi(x), \quad x \in \mathbb{R}_{q,+},$$

holds for complex numbers η, λ, and μ whenever $\phi \in L_{q,\eta+\lambda}(A_q)$, and $\Re(\lambda) < 0$.

It should be mentioned here that in most of the proofs of the semigroup properties in [Agarwal 1969; Al-Salam 1966], the domain where the fractional integrals and the related properties hold is not determined precisely.

Let $S_q^{\eta,\alpha}$ be the operator defined by

$$\text{(3-19)}\quad \begin{aligned} S_q^{\eta,\alpha}\phi(x) &:= \frac{x^{-\alpha/2}}{(1-q)}\int_0^{\infty} y^{-\alpha/2} J_{2\eta+\alpha}(\sqrt{xy};q)\phi(y)\,d_q y, \\ &= x^{-\alpha/2}\sum_{n=-\infty}^{\infty} q^{n(1-\alpha/2)} J_{2\eta+\alpha}(\sqrt{x}q^{n/2};q)\phi(q^n). \end{aligned}$$

This operator is a q-analog of the modified Hankel transform operator introduced by Erdélyi and Kober [1940]; it is also a modification of the q-Hankel transform operator introduced in [Koornwinder and Swarttouw 1992] and defined by

$$\mathcal{H}_q^{\nu}(f)(x) = \int_0^{\infty} f(t) J_{\nu}(xt;q^2)\,(xt)^{1/2}\,d_q t.$$

Proposition 3.5. *Let η and α be complex numbers such that $\Re(2\eta+\alpha) > -1$. If $\phi \in L_{q^2,\eta}(\mathbb{R}_{q^2,+})$ then $S_{q^2}^{\eta,\alpha}\phi(x)$ exists for all $x \in \mathbb{R}_{q^2,+}$ and belongs to $L_{q^2,\eta+\alpha}(\mathbb{R}_{q^2,+})$.*

Proof. Let $\phi \in L_{q^2,\eta}(\mathbb{R}_{q^2,+})$. From (2-17), if $x \in \mathbb{R}_{q^2,+}$, there exists $M > 0$ such that

$$\left|\sum_{n=-\infty}^{\infty} q^{n(2-\alpha)} J_{2\eta+\alpha}(\sqrt{x}q^n; q^2)\phi(q^{2n})\right| \le M \sum_{n=-\infty}^{\infty} q^{2n(\Re(\eta)+1)}|\phi(q^{2n})| < \infty$$

because $\phi \in L_{q^2,\eta}(\mathbb{R}_{q^2,+})$. Thus, $S_{q^2}^{\eta,\alpha}(\phi)(x)$ exists for all $x \in \mathbb{R}_{q^2,+}$. Now we prove that $S_{q^2}^{\eta,\alpha} \in L_{q^2,\eta+\alpha}(\mathbb{R}_{q^2,+})$. Indeed,

$$\begin{aligned}(3\text{-}20)\quad &\int_0^\infty t^{\eta+\alpha}|S_{q^2}^{\eta,\alpha}\phi(t)|\,d_{q^2}t\\ &= \frac{1}{1-q^2}\int_0^\infty |t^{\eta+\alpha/2}|\int_0^\infty |y^{-\alpha/2}|\,|J_{2\eta+\alpha}(\sqrt{yt}; q^2)|\,|\phi(y)|\,d_{q^2}y\,d_{q^2}t\\ &\le \frac{1}{1-q^2}\|\phi\|_\eta \int_0^\infty |t^{\eta+\alpha/2}| \sup_{y\in\mathbb{R}_{q^2,+}} |y^{-\eta-\alpha/2}J_{2\eta+\alpha}(\sqrt{ty}; q^2)|\,d_{q^2}t.\end{aligned}$$

Using the estimates (2-17), the q-integral on the third line of (3-20) is convergent when $\Re(2\eta+\alpha) > -1$, and the proposition follows. □

Proposition 3.6. *$S_{q^2}^{\eta,\alpha}$ defines a one-to-one linear operator from $L_{q^2,\eta}(\mathbb{R}_{q^2,+})$ into $L_{q^2,\eta+\alpha}(\mathbb{R}_{q^2,+})$. Also,*

$$(3\text{-}21)\qquad (S_{q^2}^{\eta,\alpha})^{-1} = S_{q^2}^{\eta+\alpha,-\alpha}.$$

Proof. Clearly $S_{q^2}^{\eta,\alpha}$ is linear. To prove that it is one-to-one, assume that there is a function $\phi \in L_{q^2,\eta}(\mathbb{R}_{q^2,+})$ such that $S_{q^2}^{\eta,\alpha}\phi(x) = 0$ for all $x \in \mathbb{R}_{q^2,+}$. Hence,

$$(3\text{-}22)\qquad \sum_{n=-\infty}^{\infty} q^{n(2-\alpha)}\phi(q^{2n})J_{2\eta+\alpha}(q^n\xi; q^2) = 0 \quad \text{for all } x \in \mathbb{R}_{q,+}.$$

Multiplying both sides of (3-22) by $\xi J_{2\eta+\alpha}(q^r\xi; q^2)$ $(r \in \mathbb{Z})$, calculating the q-integration for $\xi \in (0, \infty)$, then applying (2-21), we obtain

$$\sum_{n=-\infty}^{\infty} q^{n(2-\alpha)}\phi(q^{2n})q^{-n-r}\delta_{nr} = 0.$$

That is, $\phi(q^{2r}) = 0$ for all $r \in \mathbb{Z}$, and $S_{q^2}^{\eta,\alpha}$ is a one-to-one operator. Now we prove (3-21). From (3-19), for $j \in \mathbb{Z}$, we have

$$S_{q^2}^{\eta,\alpha}S_{q^2}^{\eta+\alpha,-\alpha}\phi(q^{2j}) = q^{-j\alpha}\sum_{n=-\infty}^{\infty} q^{n(2-\alpha)}J_{2\eta+\alpha}(q^{j+n}; q^2)S_{q^2}^{\eta,\alpha}\phi(q^{2n}).$$

Hence,

$$\begin{aligned}(3\text{-}23)\quad &S_{q^2}^{\eta,\alpha}S_{q^2}^{\eta+\alpha,-\alpha}\phi(q^{2j})\\ &= q^{-j\alpha}\sum_{n=-\infty}^{\infty} q^{2n}J_{2\eta+\alpha}(q^{j+n}; q^2)\sum_{k=-\infty}^{\infty} q^{k(2+\alpha)}J_{2\eta+\alpha}(q^{k+n}; q^2)\phi(q^{2k}).\end{aligned}$$

Using (2-17) and $\phi \in L_{q,\eta}(\mathbb{R}_{q,+})$, we can prove that the series on the left-hand side of (3-23) is absolutely convergent. Consequently, we can interchange the order of summation to obtain

$$S_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,-\alpha} \phi(q^{2j}) \\ = q^{-j\alpha} \sum_{k=-\infty}^{\infty} q^{k(2+\alpha)} \phi(q^{2k}) \sum_{n=-\infty}^{\infty} q^{2n} J_{2\eta+\alpha}(q^{j+n}; q^2) J_{2\eta+\alpha}(q^{k+n}; q^2).$$

Therefore, from (2-21),

$$S_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,-\alpha} \phi(q^{2j}) = q^{-j} q^{-j\alpha} \sum_{k=-\infty}^{\infty} q^{k(1+\alpha)} \phi(q^{2k}) \delta_{jk} = \phi(q^{2j}), \tag{3-24}$$

and the desired result follows. □

The next result gives another sufficient condition for the existence of $S_{q^2}^{\eta,\alpha}$.

Proposition 3.7. *Let η and α be complex numbers satisfying $\Re(2\eta+\alpha) > 0$. Let ϕ be a function defined on $\mathbb{R}_{q^2,+}$. If there exists $\mu \in \mathbb{C}$ such that*

$$\phi|_{A_{q^2}} \in L_{q^2,\eta}(A_{q^2}), \quad \phi|_{B_{q^2}} \in L_{q^2,\mu}(B_{q^2}),$$

then $S_{q^2}^{\eta,\alpha}\phi(x)$ exists for all $x \in \mathbb{R}_{q^2,+}$.

Proof. Let $x \in \mathbb{R}_{q^2,+}$. From (2-17), there exists $M > 0$ such that

$$|J_{2\eta+\alpha}(\sqrt{x}q^n; q^2)| \le M q^{n\Re(2\eta+\alpha)} \quad \text{for all } m \in \mathbb{N}_0.$$

Since $\phi|_{A_{q^2}} \in L_{q^2,\eta}(A_{q^2})$, then

$$\left| \sum_{n=0}^{\infty} q^{n(2-\alpha)} J_{2\eta+\alpha}(\sqrt{x}q^n; q^2) \phi(q^{2n}) \right| \le M \sum_{n=0}^{\infty} q^{2n(\Re(\eta)+1)} |\phi(q^{2n})| < \infty. \tag{3-25}$$

From (2-17), there exists $K > 0$ such that

$$|J_{2\eta+\alpha}(\sqrt{x}q^n; q^2)| \le K q^{n^2-(\Re(2\eta+\alpha)+1)n} \quad \text{for all } n \in \mathbb{Z}^-.$$

Since $\phi|_{B_{q^2}} \in L_{q^2,\mu}(B_{q^2})$, then

$$\left| \sum_{n=1}^{\infty} q^{-n(2-\alpha)} J_{2\eta+\alpha}(\sqrt{x}q^{-n}; q^2) \phi(q^{-2n}) \right| \\ \le K \sum_{n=1}^{\infty} q^{-2n(-\Re(\eta+\alpha)+1/2)+n^2} |\phi(q^{-2n})| < \infty. \tag{3-26}$$

Combining (3-25) and (3-26), we can then conclude that $S_{q^2}^{\eta,\alpha}\phi(x)$ exists for all $x \in \mathbb{R}_{q^2,+}$. □

Proposition 3.8. *Let α, β, and λ be complex numbers such that $\Re(2\eta+\alpha) > -1$. If $\phi \in L_{q^2,\eta}(\mathbb{R}_{q^2,+})$, then*

$$I_{q^2}^{\eta+\alpha,\beta} S_{q^2}^{\eta,\alpha}\phi(x) = (1-q^2)^{\beta} S_{q^2}^{\eta,\alpha+\beta}\phi(x) \quad (x \in \mathbb{R}_{q^2,+}).$$

Proof. Let $x \in \mathbb{R}_{q^2,+}$ be fixed. Using definitions (3-10) and (3-19) we get

$$\begin{aligned}(3\text{-}27)\quad I_{q^2}^{\eta+\alpha,\beta} S_{q^2}^{\eta,\alpha}\phi(x) &= \frac{x^{-\eta-\alpha-1}}{(1-q^2)\Gamma_{q^2}(\beta)} \\ &\quad\times \int_0^x (q^2t/x;q^2)_{\beta-1} t^{\eta+\alpha/2} \int_0^\infty y^{-\alpha/2} J_{2\eta+\alpha}(\sqrt{ty};q^2)\phi(y)\,d_{q^2y}\,d_{q^2}t.\end{aligned}$$

From (2-17), there exists $M > 0$ such that

$$|J_{2\eta+\alpha}(\sqrt{ty};q^2)| \le M(ty)^{\Re(\eta+\alpha/2)} \quad \text{for all } y,t \in \mathbb{R}_{q^2,+}.$$

Consequently,

$$\begin{aligned}&\left|\frac{1}{\Gamma_q^2(\beta)}\int_0^x (q^2t/x;q^2)_{\beta-1} t^{\eta+\alpha/2}\int_0^\infty y^{-\alpha/2} J_{2\eta+\alpha}(\sqrt{ty};q^2)\phi(y)\,d_{q^2}y\,d_{q^2}t\right| \\ &\qquad \le \frac{M}{|\Gamma_q^2(\beta)|}\|\phi\|_{q^2,\eta}\left|\int_0^x (q^2t/x;q^2)_{\beta-1} t^{2\eta+\alpha}\,d_{q^2}t\right| \\ &\qquad = \frac{M}{|\Gamma_q^2(\beta)|}\|\phi\|_{q^2,\eta}|x^{2\eta+\alpha+1}B_q(\beta, 2\eta+\alpha+1)| < \infty,\end{aligned}$$

since $\Re(2\eta+\alpha)+1 > 0$. Hence, the series is absolutely convergent, and we can interchange the order of summation to obtain

$$\begin{aligned}(3\text{-}28)\quad I_{q^2}^{\eta+\alpha,\beta} S_{q^2}^{\eta,\alpha}\phi(x) &= \frac{x^{-\eta-\alpha-1}}{\Gamma_{q^2}(\beta)(1-q^2)} \\ &\quad\times \int_0^\infty y^{-\alpha/2}\phi(y)\int_0^x (q^2t/x;q^2)_{\beta-1} t^{\eta+\alpha/2} J_{2\eta+\alpha}(\sqrt{ty};q^2)\,d_{q^2}t\,d_{q^2}y.\end{aligned}$$

Using

$$(3\text{-}29)\qquad \int_0^x f(t)\,d_{q^2}t = \frac{1+q}{x}\int_0^x tf(t^2/x)\,d_qt$$

and (2-27), we obtain

$$\begin{aligned}(3\text{-}30)\quad &\int_0^x (q^2t/x;q^2)_{\beta-1} t^{\eta+\alpha/2} J_{2\eta+\alpha}(\sqrt{ty};q^2)\,d_{q^2}t\,d_{q^2}y \\ &\qquad = \frac{1+q}{x^{1+\eta+\alpha/2}}\int_0^x (q^2t^2/x^2;q^2)_{\beta-1} t^{2\eta+\alpha+1} J_{2\eta+\alpha}\left(t\sqrt{\frac{y}{x}};q^2\right)d_qt \\ &\qquad = (1-q^2)^{\beta} y^{-\beta/2} x^{1+\eta+(\alpha-\beta)/2}\Gamma_{q^2}(\beta) J_{2\eta+\alpha+\beta}(\sqrt{xy};q^2).\end{aligned}$$

Substituting (3-30) into (3-28) yields the desired result and completes the proof. □

Proposition 3.9. *Let α, β, and η be complex numbers. If $\phi \in L_{q^2,\eta+\alpha-\gamma}(\mathbb{R}_{q^2,+})$ for some $\gamma \in \mathbb{C}$, $\Re(\gamma) > \max\{0, \Re(\alpha)\}$, then*

$$\mathcal{K}_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,\beta}\phi(x) = (1-q^2)^{\alpha} S_{q^2}^{\eta,\alpha+\beta}\phi(x) \tag{3-31}$$

for all $x \in \mathbb{R}_{q^2,+}$.

Proof. Using definitions (3-2), (3-19), (2-8), and (2-9) we get

$$\begin{aligned}\mathcal{K}_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,\beta}\phi(x) &= \frac{q^{-2\eta}x^{\eta}}{\Gamma_{q^2}(\alpha)}\int_x^{\infty}(x/t;q^2)_{\alpha-1}(S_{q^2}^{\eta+\alpha,\beta}\phi)(q^2t)\,d_{q^2}t\\ &= \frac{x^{\eta}}{\Gamma_{q^2}(\alpha)}\int_{q^2x}^{\infty}(q^2x/t;q^2)_{\alpha-1}(S_{q^2}^{\eta+\alpha,\beta}\phi)(t)\,d_{q^2}t\\ &= \frac{x^{\eta}}{(1-q^2)\Gamma_{q^2}(\alpha)}\int_{q^2x}^{\infty}\biggl((q^2x/t;q^2)_{\alpha-1}t^{-\eta-1-\beta/2}\\ &\qquad\times\int_0^{\infty}y^{-\beta/2}J_{2\eta+2\alpha+\beta}(\sqrt{yt};q^2)\phi(y)\,d_{q^2}y\biggr)\,d_{q^2}t.\end{aligned} \tag{3-32}$$

Set $c := -1-\Re(2\eta+\beta+2\alpha-2\gamma)$. From (2-18), there exists $C > 0$ such that

$$|J_{2\eta+2\alpha+\beta}(q^r;q^2)| \le Cq^{-r(c+1)} \quad (r \in \mathbb{Z}). \tag{3-33}$$

Hence,

$$\begin{aligned}\biggl|\int_{q^2x}^{\infty}(q^2x/t;q^2)_{\alpha-1}t^{-\eta-1-\beta/2}\int_0^{\infty}y^{-\beta/2}J_{2\eta+2\alpha+\beta}(\sqrt{yt};q^2)\phi(y)\,d_{q^2}y\,d_{q^2}t\biggr|\\ \le C\|\phi\|_{q^2,\eta+\alpha-\gamma}|(q^2x/t;q^2)_{\alpha-1}t^{\alpha-\gamma-1}\,d_{q^2}t| < \infty\end{aligned}$$

whenever $\Re(\gamma) > \Re(\alpha)$. Hence, the double q^2-integration in (3-32) is absolutely convergent, and we can interchange the order of q^2-integration to obtain

$$\begin{aligned}K_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,\beta}\phi(x) &= \frac{x^{\eta}}{\Gamma_{q^2}(\alpha)}\\ &\times\int_0^{\infty}y^{-\beta/2}\phi(y)\int_{q^2x}^{\infty}(q^2x/t;q^2)_{\alpha-1}t^{-\eta-1-\beta/2}J_{2\eta+2\alpha+\beta}(\sqrt{yt};q^2)\,d_{q^2t}\,d_{q^2}y.\end{aligned}$$

Using

$$\int_a^{\infty} f(t)\,d_{q^2}t = \frac{1+q}{a}\int_a^{\infty} tf\Bigl(\frac{t^2}{a}\Bigr)\,d_qt$$

and Proposition 2.10, one finds that

$$\begin{aligned}\int_{q^2x}^{\infty}&(q^2x/t;q^2)_{\alpha-1}t^{-\eta-1-\beta/2}J_{2\eta+2\alpha+\beta}(\sqrt{yt};q^2)\,d_{q^2}t\\ &=(1+q)(q^2x)^{\eta+\beta/2}\int_{q^2x}^{\infty}(q^4x^2/t^2;q^2)_{\alpha-1}t^{-2\eta-1-\beta}J_{2\eta+2\alpha+\beta}\left(\frac{t\sqrt{y}}{q\sqrt{x}};q^2\right)d_{q^2}t\\ &=(1+q)y^{-\alpha/2}x^{-\eta-(\alpha+\beta)/2}\frac{(q^2;q^2)_\infty}{(q^{2\alpha};q^2)_\infty}J_{2\eta+\alpha}(\sqrt{xy};q^2).\end{aligned}$$

Hence,

$$\begin{aligned}K_{q^2}^{\eta,\alpha}S_{q^2}^{\eta+\alpha,\beta}\phi(x)&=(1-q^2)^{\alpha-1}x^{-(\alpha+\beta)/2}\int_0^\infty y^{-(\alpha+\beta)/2}\phi(y)J_{2\eta+\alpha}(\sqrt{xy};q^2)\,d_{q^2}y\\ &=(1-q^2)^{\alpha}S_{q^2}^{\eta,\alpha+\beta}(x).\end{aligned}$$ □

Proposition 3.10. *Let η, α, and β be complex numbers satisfying $\Re(\beta+\alpha)>0$ and $\Re(2\eta+\alpha)>-1$. If $\phi\in L_{q^2,\eta}(\mathbb{R}_{q^2,+})$, then*

$$\mathcal{S}_{q^2}^{\eta+\alpha,\beta}S_{q^2}^{\eta,\alpha}\phi(x)=(1-q^2)^{-\beta-\alpha}I_{q^2}^{\eta,\alpha+\beta}\phi(x)\quad(x\in\mathbb{R}_{q^2,+}).$$

Proof. Let $x\in\mathbb{R}_{q^2,+}$ be fixed. Using definitions (3-10) and (3-19) we get

$$\text{(3-34)}\quad S_{q^2}^{\eta+\alpha,\beta}S_{q^2}^{\eta,\alpha}\phi(x)=\frac{x^{-\beta/2}}{(1-q^2)^2}\int_0^\infty\biggl(t^{-(\alpha+\beta)/2}J_{2\eta+2\alpha+\beta}(\sqrt{tx};q^2)\times\int_0^\infty y^{-\alpha/2}J_{2\eta+\alpha}(\sqrt{ty};q^2)\phi(y)\,d_{q^2}y\biggr)\,d_{q^2}t.$$

From (2-17), there exists $M>0$ such that

$$|J_{2\eta+\alpha}(\sqrt{ty};q^2)|\le M(ty)^{\Re(\eta+\alpha/2)}\quad\text{for all } y,\ t\in\mathbb{R}_{q^2,+}.$$

Consequently,

$$\begin{aligned}\left|\int_0^\infty t^{-(\alpha+\beta)/2}J_{2\eta+2\alpha+\beta}(\sqrt{tx};q^2)\int_0^\infty y^{-\alpha/2}J_{2\eta+\alpha}(\sqrt{ty};q^2)\phi(y)\,d_{q^2}y\,d_{q^2}t\right|&\\ \le M\|\phi\|_{q^2,\eta}\left|\int_0^\infty t^{\eta-\beta/2}J_{2\eta+2\alpha+\beta}(\sqrt{tx};q^2)\,d_{q^2}t\right|&<\infty.\end{aligned}$$

Hence, the series is absolutely convergent, and we can interchange the order of summation to obtain

$$\mathcal{S}_{q^2}^{\eta+\alpha,\beta}S_{q^2}^{\eta,\alpha}\phi(x)=\frac{x^{-\beta/2}(1+q)}{(1-q^2)^2}\int_0^\infty\biggl(y^{-\alpha/2}\phi(y)\times\int_0^\infty t^{1-\beta-\alpha}J_{2\eta+2\alpha+\beta}(\sqrt{x}t;q^2)J_{2\eta+\alpha}(\sqrt{y}t;q^2)\,d_qt\biggr)\,d_{q^2}y.$$

Therefore, applying Corollary 2.5 with $\Re(\beta+\alpha)>0$ and $\Re(2\eta+\alpha)>-1$, we obtain

$$\begin{aligned}\mathscr{S}_{q^2}^{\eta+\alpha,\beta} S_{q^2}^{\eta,\alpha}\phi(x) &= \frac{(1-q^2)^{-\beta-\alpha}x^{-\eta-1}}{\Gamma_{q^2}(\beta+\alpha)}\int_0^x y^{\eta}(q^2y/x;q^2)_{\alpha+\beta-1}\phi(y)\,d_{q^2}y\\ &= (1-q^2)^{-\beta-\alpha} I_{q^2}^{\eta,\alpha+\beta}\phi(x).\end{aligned}$$ □

Proposition 3.11. *Let η, α, and β be complex numbers satisfying $\Re(\beta+\alpha)>0$ and $\Re(2\eta+\alpha)>-1$. If $\phi\in L_{q^2,\eta+\alpha}(\mathbb{R}_{q^2,+})$, then*

$$\mathscr{S}_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,\beta}\phi(x) = (1-q^2)^{-\beta-\alpha}\mathscr{K}_{q^2}^{\eta,\alpha+\beta}\phi(x)\quad (x\in\mathbb{R}_{q^2,+}).$$

Proof. Let $x\in\mathbb{R}_{q^2,+}$ be fixed. Using definitions (3-10) and (3-19) we get

$$\text{(3-35)}\quad S_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,\beta}\phi(x) = \frac{x^{-\alpha/2}}{(1-q^2)^2}\int_0^\infty\biggl(t^{-(\alpha+\beta)/2}J_{2\eta+\alpha}(\sqrt{tx};q^2)\times\int_0^\infty y^{-\beta/2}J_{2\eta+2\alpha+\beta}(\sqrt{ty};q^2)\phi(y)\,d_{q^2y}\biggr)\,d_{q^2}t.$$

From (2-17), there exists $M>0$ such that

$$|J_{2\eta+2\alpha+\beta}(\sqrt{ty};q^2)|\le M(ty)^{\Re(\eta+\alpha+\beta/2)}\quad\text{for all } y,t\in\mathbb{R}_{q^2,+}.$$

Consequently,

$$\begin{aligned}&\left|\int_0^\infty t^{-(\alpha+\beta)/2}J_{2\eta+\alpha}(\sqrt{tx};q^2)\int_0^\infty y^{-\beta/2}J_{2\eta+2\alpha+\beta}(\sqrt{ty};q^2)\phi(y)\,d_{q^2}y\,d_{q^2}t\right|\\ &\qquad\le M\|\phi\|_{q^2,\eta+\alpha}\left|\int_0^\infty t^{\eta+\alpha/2}J_{2\eta+\alpha}(\sqrt{tx};q^2)\,d_{q^2}t\right|<\infty.\end{aligned}$$

Hence, the series is absolutely convergent, and we can interchange the order of summation to obtain

$$\mathscr{S}_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,\beta}\phi(x) = \frac{x^{-\alpha/2}(1+q)}{(1-q^2)^2}\int_0^\infty\biggl(y^{-\beta/2}\phi(y)\times\int_0^\infty t^{1-\beta-\alpha}J_{2\eta+2\alpha+\beta}(\sqrt{y}t;q^2)J_{2\eta+\alpha}(\sqrt{x}t;q^2)\,d_qt\biggr)\,d_{q^2}y.$$

Therefore, applying Corollary 2.5 with $\Re(\beta+\alpha)>0$ and $\Re(2\eta+\alpha)>-1$, we obtain

$$\begin{aligned}\mathscr{S}_{q^2}^{\eta,\alpha} S_{q^2}^{\eta+\alpha,\beta}\phi(x) &= \frac{(1-q^2)^{-\beta-\alpha}q^{-\eta}x^{\eta}}{\Gamma_{q^2}(\beta+\alpha)}\int_x^\infty y^{-\eta-1}(q^2x/y;q^2)_{\alpha+\beta-1}\phi(q^2y)\,d_{q^2}y\\ &= (1-q^2)^{-\beta-\alpha}\mathscr{K}_{q^2}^{\eta,\alpha+\beta}\phi(x).\end{aligned}$$ □

4. The multiplying factor method

Titchmarsh [1986, p. 334–339] solved the dual integral equations

$$\begin{aligned}\int_0^\infty \xi^{2\alpha}\psi(\xi)J_\nu(\rho\xi)\,d\xi &= f(\rho),\ (0<\rho<1),\\ \int_0^\infty \psi(\xi)J_\nu(\rho\xi)\,d\xi &= 0,\ (1<\rho),\end{aligned} \tag{4-1}$$

for $\alpha > 0$ by using difficult analysis involving Mellin transforms. His techniques were extended to the case $\alpha > -1$ by Busbridge [1938]. Sneddon [1960] used the Abel integral formula to solve the system in the case when $\nu = 0$ and $\alpha = \pm\frac{1}{2}$. The technique was generalized by Copson [1961] for any α satisfying $-1<\alpha<1, \alpha \neq 0$, and $\nu > -1$. In this section, we introduce and solve a q-analog of Copson's result.

Theorem 4.1. *Let α and ν be complex numbers such that $\Re(\nu) > -1$ and let f be a function defined on A_q. Consider the dual q-integral equations*

$$\int_0^\infty \xi^{2\alpha}\psi(\xi)J_\nu(\rho\xi;q^2)d_q\xi = f(\rho), \quad \rho\in A_q, \tag{4-2}$$

$$\int_0^\infty \psi(\xi)J_\nu(\rho\xi;q^2)d_q\xi = 0, \quad \rho\in B_q. \tag{4-3}$$

We consider three cases:

Case I. *If $\alpha = 0$, $\Re(\nu) > -1$, then the q-integral equations* (4-2)–(4-3) (*which are now not dual*) *have a solution of the form*

$$\psi(\xi) = \frac{\xi}{1-q}\int_0^1 t^2 J_\nu(\xi t;q^2) f(t)\,d_qt, \tag{4-4}$$

provided that $t^{\nu+1}f(t)\in L_{q^2}(A_{q^2})$.

Case II. *If $0<\Re(\alpha)<1$ and $\Re(\nu+\alpha)>0$, the dual q-integral equations* (4-2)–(4-3) *have a solution of the form*

$$\psi(\xi) = \frac{\xi^{1-\alpha}}{(1-q)^2(1-q^2)^\alpha}\int_0^1 t^{1-\nu-\alpha}J_{\nu+\alpha}(\xi t;q^2)\,I_{q^2}^\alpha[(\,\cdot\,)^{\nu/2}f(\sqrt{\cdot})](t^2)\,d_qt, \tag{4-5}$$

provided that

$$I_{q^2}^\alpha[(\,\cdot\,)^{\nu/2}f(\sqrt{\cdot})](\rho^2) \textit{ and } I_{q^2}^{\alpha-1}[(\,\cdot\,)^{\nu/2}f(\sqrt{\cdot})](\rho^2) \textit{ are } L_{q^2}(A_{q^2}) \textit{ functions.} \tag{4-6}$$

Case III. *If $-1<\Re(\alpha)<0$ and $\Re(\nu+\alpha)>-1$, the dual q-integral equations* (4-2)–(4-3) *have a solution of the form*

$$\psi(\xi) = \frac{\xi^{1-\alpha}}{(1-q)(1-q^2)^\alpha}\int_0^1 t^{1-\nu-\alpha}J_{\nu+\alpha}(\xi t;q^2)I_{q^2}^\alpha[(\,\cdot\,)^{\nu/2}f(\sqrt{\cdot})](t^2)\,d_qt, \tag{4-7}$$

provided that

(4-8) $I^{\alpha}_{q^2}[(\cdot)^{\nu/2} f(\sqrt{\cdot})](\rho^2)$ *and* $I^{\alpha+1}_{q^2}[(\cdot)^{\nu/2} f(\sqrt{\cdot})](\rho^2)$ *are* $L_{q^2}(A_{q^2})$ *functions.*

Proof. Let $\phi \in L_{q,\nu+\alpha-2}(A_q)$ be a function which shall be defined later on. Define the function ψ on $\mathbb{R}_{q,+}$ by

$$\psi(\xi) = \xi^{1-\alpha} \int_0^1 \phi(t) J_{\nu+\alpha}(\xi t; q^2)\, d_q t. \tag{4-9}$$

Hence ψ is a well-defined function. We now prove that ψ satisfies (4-3). Indeed,

$$\begin{aligned}\int_0^\infty \psi(\xi) J_\nu(\rho\,\xi; q^2)\, d_q\xi \\ = \int_0^\infty \xi^{1-\alpha} J_\nu(\rho\xi; q^2) \int_0^1 \phi(t) J_{\nu+\alpha}(\xi t; q^2)\, d_q t\, d_q\xi.\end{aligned} \tag{4-10}$$

From (2-17), there exists $M > 0$ such that

$$|J_{\nu+\alpha}(\xi\, t; q^2)| \le M |(\xi\, t)^{\nu+\alpha}|$$

for all ξ, t in $\mathbb{R}_{q,+}$. Also, $\phi \in L_{q,\nu+\alpha-2}(A_q)$ implies that $\phi \in L_{q,\nu+\alpha}(A_q)$. Consequently,

$$\left| \int_0^\infty \psi(\xi) J_\nu(\rho\,\xi; q^2)\, d_q\xi \right| \le M \|\phi\|_{A_q,\nu+\alpha} \int_0^\infty |\xi^{1+\nu} J_\nu(\rho\,\xi; q^2)|\, d_q\xi < \infty$$

whenever $\Re(\nu) > -1$, where we applied again (2-17). Therefore, the double q-integration in (4-10) is absolutely convergent and we can interchange the order of q-integration to obtain

$$\begin{aligned}\int_0^\infty \psi(\xi) J_\nu(\rho\,\xi; q^2)\, d_q\xi \\ = \int_0^1 \phi(t) \int_0^\infty \xi^{1-\alpha} J_\nu(\rho\xi; q^2) J_{\nu+\alpha}(\xi t; q^2)\, d_q\xi\, d_q\, t.\end{aligned} \tag{4-11}$$

Thus, by replacing α, β by ν, $\nu+\alpha$, respectively, and applying Corollary 2.5, the q-integration in (4-11) vanishes, and this proves (4-3). In the following we prove (4-2). We distinguish among three cases.

Case I. $\alpha = 0$ and $\Re(\nu) > -1$. By (4-11), (2-21), Equation (4-2) is reduced to

$$\frac{1-q}{\rho^2} \int_0^1 \phi(t) \delta_{\rho,t}\, d_q t = f(\rho), \tag{4-12}$$

that is,

$$\phi(\rho) = \frac{\rho^2 f(\rho)}{1-q}. \tag{4-13}$$

Substituting (4-13) in (4-9), we obtain (4-4).

Case II. $0 < \Re(\alpha) < 1$ and $\Re(\nu+\alpha) > 0$. From (2-15), we obtain

$$\psi(\xi) = -q^{\nu+\alpha-1}(1-q)\xi^{-\alpha}\int_0^1 \phi(t)t^{\nu+\alpha-1}D_{q,t}[t^{1-\nu-\alpha}J_{\nu+\alpha-1}(q^{-1}\xi t;q^2)]\,d_qt.$$

Applying the q-integration by parts rule (2-10), we obtain

$$\begin{aligned}\psi(\xi) = (1-q)\xi^{-\alpha}\int_0^1 D_{q,t}[t^{\nu+\alpha-1}\phi(t)]t^{1-\nu-\alpha}J_{\nu+\alpha-1}(\xi t;q^2)\,d_qt&\\ -(1-q)q^{\nu+\alpha-1}\xi^{-\alpha}\phi(1)J_{\nu+\alpha-1}(q^{-1}\xi;q^2)&\\ +(1-q)q^{\nu+\alpha-1}\xi^{-\alpha}\lim_{n\to\infty}\phi(q^n)J_{\nu+\alpha-1}(q^{n-1}\xi;q^2)&.\end{aligned}$$

Since $\phi \in L_{q,\nu+\alpha-2}(A_q)$, we have

$$\lim_{n\to\infty} q^{n(\nu+\alpha-1)}\phi(q^n) = 0.$$

Therefore,

$$\begin{aligned}\psi(\xi) = (1-q)\xi^{-\alpha}\int_0^1 D_{q,t}[t^{\nu+\alpha-1}\phi(t)]t^{1-\nu-\alpha}J_{\nu+\alpha-1}(\xi t;q^2)\,d_qt&\\ -(1-q)q^{\nu+\alpha-1}\xi^{-\alpha}\phi(1)J_{\nu+\alpha-1}(q^{-1}\xi;q^2)&.\end{aligned}\tag{4-14}$$

Substituting (4-14) into (4-2), we obtain

$$\begin{aligned}f(\rho) = -(1-q)q^{\nu+\alpha-1}\phi(1)\int_0^\infty \xi^\alpha J_{\nu+\alpha-1}(q^{-1}\xi;q^2)J_\nu(\rho\xi;q^2)\,d_q\xi&\\ +(1-q)\int_0^\infty \xi^\alpha\int_0^1 \varphi(t)t^{1-\nu-\alpha}J_{\nu+\alpha-1}(\xi t;q^2)J_\nu(\rho\xi;q^2)\,d_qt\,d_q\xi&,\end{aligned}\tag{4-15}$$

where for the convenience of the reader, we set

$$\varphi(t) := D_{q,t}[t^{\nu+\alpha-1}\phi(t)] \quad (t \in A_q).\tag{4-16}$$

Since $\rho \in A_q$, then from Corollary 2.5, the first q-integral in (4-15) vanishes. As for the second double q-integral, the conditions on ϕ imply that $\varphi \in L_q(A_q)$. Therefore, under the conditions on ν and α, the double q-integration is absolutely convergent; and we can interchange the order of q-integration to obtain

$$f(\rho) = (1-q)\int_0^1 \varphi(t)t^{1-\nu-\alpha}\int_0^\infty \xi^\alpha J_{\nu+\alpha-1}(\xi t;q^2)J_\nu(\rho\xi;q^2)\,d_q\xi\,d_qt.$$

But under the conditions that $\Re(\alpha) < 1$ and $\Re(\nu+\alpha) > 0$, we have

$$\int_0^\infty \xi^\alpha J_{\nu+\alpha-1}(t\xi;q^2)J_\nu(\rho\xi;q^2)\,d_q\xi = \begin{cases}\dfrac{(1-q)(1-q^2)^\alpha}{\Gamma_{q^2}(1-\alpha)}t^{\nu+\alpha-1}\rho^{-\nu-2\alpha}(q^2t^2/\rho^2;q^2)_{-\alpha} & \text{if } \rho \ge t,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore,

$$\begin{aligned} f(\rho) &= \frac{(1-q)^2(1-q^2)^\alpha}{\Gamma_{q^2}(1-\alpha)}\rho^{-\nu-2\alpha}\int_0^\rho (q^2t^2/\rho^2;q^2)_{-\alpha}\varphi(t)\,d_qt \\ &= \frac{(1-q)^2(1-q^2)^\alpha\rho^{-(\nu+2\alpha)}}{(1+q)\Gamma_{q^2}(1-\alpha)}\int_0^{\rho^2} (q^2t/\rho^2;q^2)_{-\alpha}\frac{\varphi(\sqrt{t})}{\sqrt{t}}\,d_{q^2}t. \end{aligned} \tag{4-17}$$

Now, we apply the q-Abel integral (Theorem 3.2) under the conditions in (4-6). This gives

$$\frac{\varphi(\rho)}{\rho} = \frac{(1+q)}{(1-q)^2(1-q^2)^\alpha}(D_{q^2}I^\alpha_{q^2}g)(\rho^2), \quad g(\rho^2) := \rho^\nu f(\rho),\ \rho \in A_q. \tag{4-18}$$

From (4-16) and the fact that $\phi \in L_{q,\nu+\alpha-1}(A_q)$, we obtain

$$\phi(t) = t^{-\nu+\alpha-1}\int_0^t \varphi(\rho)\,d_q\rho.$$

Thus, from (4-18),

$$\begin{aligned} \phi(t) &= \frac{(1+q)t^{1-\nu-\alpha}}{(1-q)^2(1-q^2)^\alpha}\int_0^t \rho(D_{q^2}I^\alpha_{q^2}g)(\rho^2)\,d_q\rho \\ &= \frac{t^{1-\nu-\alpha}}{(1-q)^2(1-q^2)^\alpha}\int_0^{t^2} (D_{q^2}I^\alpha_{q^2}g)(\rho)\,d_{q^2}\rho \\ &= \frac{t^{1-\nu-\alpha}}{(1-q)^2(1-q^2)^\alpha}(I^\alpha_{q^2}g)(t^2). \end{aligned} \tag{4-19}$$

Hence, $\psi(\,\cdot\,)$ is given by (4-9).

Case III. $-1 < \Re(\alpha) < 0$ and $\Re(\nu+\alpha) > -1$. We substitute (4-9) in (4-2) to obtain

$$f(\rho) = \int_0^\infty \xi^{\alpha+1}\int_0^1 \phi(t)J_{\nu+\alpha}(\xi t;q^2)J_\nu(\rho\xi;q^2)\,d_qt\,d_q\xi. \tag{4-20}$$

If $\phi \in L_{q,\nu+\alpha}(A_q)$ and $\Re(\nu+\alpha) > -1$, the double q-integration in (4-20) is absolutely convergent. Therefore, we can interchange the order of q-integration to obtain

$$f(\rho) = \int_0^1 \phi(t)\int_0^\infty \xi^{\alpha+1}J_{\nu+\alpha}(\xi t;q^2)J_\nu(\rho\xi;q^2)\,d_q\xi\,d_qt.$$

Since $\Re(\alpha) < 0$, then $\Re(\nu) > \Re(\nu+\alpha) > -1$. Thus applying Corollary 2.5 yields

$$f(\rho) = \frac{\rho^{-\nu-2\alpha-2}(1-q)}{\Gamma_{q^2}(-\alpha)(1-q^2)^{-\alpha}}\int_0^{\rho^2}\left(\frac{q^2t}{\rho^2};q^2\right)_{-\alpha-1} t^{(\nu+\alpha-1)/2}\phi(\sqrt{t})\,d_{q^2}t.$$

Since f satisfies (4-8), we can apply the q-Abel integral (Theorem 3.2) to obtain

$$\begin{aligned}\phi(\rho) &= \frac{\rho^{1-\nu-\alpha}}{(1-q)(1-q^2)^{\alpha}} D_{q^2,\rho^2} I_{q^2}^{\alpha+1}[(\cdot)^{\nu/2} f(\sqrt{\cdot})](\rho^2) \\ &= \frac{\rho^{1-\nu-\alpha}}{(1-q)(1-q^2)^{\alpha}} I_{q^2}^{\alpha}[(\cdot)^{\nu/2} f(\sqrt{\cdot})](\rho^2),\end{aligned} \tag{4-21}$$

where $\rho \in A_q$. Thus, substituting (4-21) into (4-9) gives (4-7) and completes the proof of the theorem. □

Example 4.2. If $\nu = \frac{1}{2}$ and $\alpha = \frac{1}{4}$, then the solution of the system

$$\begin{aligned}\frac{\rho^{-1/2}\sqrt{1-q^2}}{\Gamma_{q^2}(\frac{1}{2})} \int_0^{\infty} \psi(\xi) \sin\left(\frac{\rho\xi}{1-q}; q^2\right) d_q\xi &= f(\rho), \quad \rho \in A_q, \\ \frac{\rho^{-1/2}\sqrt{1-q^2}}{\Gamma_{q^2}(\frac{1}{2})} \int_0^{\infty} \xi^{-\frac{1}{2}} \psi(\xi) \sin\left(\frac{\rho\xi}{1-q}; q^2\right) d_q\xi &= 0, \quad \rho \in B_q,\end{aligned}$$

is

$$\psi(\xi) = \frac{\xi^{3/4}}{(1-q)^2(1-q^2)^{1/4}} \int_0^1 t^{1/4} J_{3/4}(\xi t; q^2) I_{q^2}^{1/4}[(\cdot)^{-1/4} f(\sqrt{\cdot})](t^2)\, d_q t,$$

where $I_{q^2}^{1/4}[(\cdot)^{1/4} f(\sqrt{\cdot})](t^2)$ and $I_{q^2}^{-3/4}[(\cdot)^{1/4} f(\sqrt{\cdot})](t^2)$ are $L_{q^2}(A_{q^2})$ functions.

In particular, if $f(t) = t^{3/2}$, then

$$\psi(\xi) = \frac{\xi^{3/4}\Gamma_{q^2}(3/2)}{(1-q)^2(1-q^2)^{1/4}\Gamma_{q^2}(7/4)} \int_0^1 t^{7/4} J_{3/4}(\xi t; q^2)\, d_q t,$$

and by (2-16), we obtain

$$\psi(\xi) = \frac{\Gamma_{q^2}(3/2)}{(1-q)(1-q^2)^{1/4}\Gamma_{q^2}(7/4)} \xi^{-1/4} J_{7/4}(\xi; q^2).$$

Example 4.3. If $v = 2$ and $\alpha = -\frac{1}{2}$, then the solution of the system

$$\begin{aligned}\int_0^{\infty} \xi^{-1} \psi(\xi) J_2(\rho\xi; q^2)\, d_q\xi &= f(\rho), \quad \rho \in A_q, \\ \int_0^{\infty} \psi(\xi) J_2(\rho\xi; q^2) d_q\xi &= 0, \quad \rho \in B_q,\end{aligned}$$

is

$$\psi(\xi) = \frac{\xi^{3/2}(1-q^2)^{1/2}}{(1-q)} \int_0^1 t^{-1/2} J_{3/2}(\xi t; q^2) I_{q^2}^{-1/2}[(\cdot) f(\sqrt{\cdot})](t^2)\, d_q t,$$

where $I_{q^2}^{-1/2}[(\cdot) f(\sqrt{\cdot})](t^2)$ and $I_{q^2}^{1/2}[(\cdot) f(\sqrt{\cdot})](t^2)$ are $L_{q^2}(A_{q^2})$ functions.

In particular, if we take $f(t)=t^{-1}$ then

$$\psi(\xi)=\frac{\xi^{3/2}(1-q^2)^{1/2}\Gamma_{q^2}(3/2)}{(1-q)}\int_0^1 t^{-1/2}J_{3/2}(\xi t;q^2)\,d_qt,$$

and by applying (2-15), we can see that

$$\psi(\xi)=-\frac{q(1-q^2)\Gamma_{q^2}(3/2)}{\Gamma_{q^2}(1/2)}\sin\biggl(\frac{q^{-1}\xi}{1-q};q\biggr)-\xi.$$

5. The fractional q-calculus approach

In this section, we solve certain dual q-integral equations by using the fractional q-calculus approach. Peters [1961] solved the dual integral equations

$$\int_0^\infty \xi^{-2\alpha}\Psi(\xi)J_\mu(2\rho\xi)d\xi=F(\rho)\quad(0<\rho<1),$$
$$\int_0^\infty \xi^{-2\beta}\Psi(\xi)J_\nu(2\rho\xi)d\xi=G(\rho)\quad(\rho>1),$$

by using fractional calculus. Here we give a q-type analog of Peters' problem.

Theorem 5.1. *Let α, β, μ, and ν be complex numbers and let*

$$\lambda:=\tfrac{1}{2}(\mu+\nu)-(\alpha-\beta)>-1.$$

Assume that

$$\Re(\mu)>-1,\quad \Re(\nu)>-1,\quad \Re(\lambda)>-1,\quad \textit{and}\quad \Re(\lambda-\mu-2\alpha)>0.$$

Let $f\in L_{q^2,\mu/2+\alpha}(A_{q^2})$ and $g\in L_{q^2,-\mu/2+\alpha-1}(B_{q^2})$. Then the dual q^2-integral equations

$$\begin{aligned}\xi^{-\alpha}\int_0^\infty \rho^{-\alpha}\psi(\rho)J_\mu(\sqrt{\rho\xi};q^2)d_{q^2}\rho&=f(\xi)\quad(\xi\in A_{q^2}),\\ \xi^{-\beta}\int_0^\infty \rho^{-\beta}\psi(\rho)J_\nu(\sqrt{\rho\xi};q^2)d_{q^2}\rho&=g(\xi)\quad(\xi\in B_{q^2}),\end{aligned}\tag{5-1}$$

have the solution

$$\begin{aligned}\psi(\xi)=(1-q^2)^{\lambda-\nu+2\alpha-2}\xi^{\lambda/2-\mu/2+\alpha}\int_0^1 J_\lambda(\sqrt{\rho\xi};q^2)I_{q^2}^{\mu/2+\alpha,\lambda-\mu}f(\rho)\,d_{q^2}\rho\\ +(1-q^2)^{\lambda-\nu-2}\xi^{\lambda/2-\mu/2+\alpha}\int_1^\infty J_\lambda(\sqrt{\rho\xi};q^2)\mathcal{K}_{q^2}^{\lambda/2-\nu/2-\beta,\nu-\lambda}g(\rho)\,d_{q^2}\rho,\end{aligned}$$

in $L_{q^2,\mu/2-\alpha}(\mathbb{R}_{q^2,+})\cap L_{q^2,\nu/2-\beta}(\mathbb{R}_{q^2,+})\cap L_{q^2,\nu/2-\beta-\gamma}(\mathbb{R}_{q^2,+})$, for γ satisfying

$$1+\Re(\nu)>\Re(\gamma)>\max\{0,\Re(\nu-\lambda)\}.\tag{5-2}$$

Proof. We shall extend the domain of the functions f and g in (5-1) to be $\mathbb{R}_{q^2,+}$ by introducing the functions f_1 and g_1, where

$$f_1 \equiv f \text{ on } A_{q^2} \quad \text{and} \quad f_1 \equiv 0 \text{ on } B_{q^2},$$
$$g_1 \equiv 0 \text{ on } A_{q^2} \quad \text{and} \quad g_1 \equiv g \text{ on } B_{q^2},$$

respectively. From (3-19), (5-1) can be written as

$$\begin{aligned}(1-q^2)S_{q^2}^{\mu/2-\alpha,2\alpha}\psi(\xi) &= f(\xi) \quad (\xi \in A_{q^2}),\\ (1-q^2)S_{q^2}^{\nu/2-\beta,2\beta}\psi(\xi) &= g(\xi) \quad (\xi \in B_{q^2}).\end{aligned} \tag{5-3}$$

Under the conditions on the functions ψ, f, and g, we can apply Propositions 3.8 and 3.9 to obtain

$$\frac{I_{q^2}^{\mu/2+\alpha,\lambda-\mu} f(\xi)}{(1-q^2)^{\lambda-\mu+1}} = \frac{I_{q^2}^{\mu/2+\alpha,\lambda-\mu} S_{q^2}^{\mu/2-\alpha,2\alpha}\psi(\xi)}{(1-q^2)^{\lambda-\mu+1}} = S_{q^2}^{\mu/2-\alpha,\lambda-\mu+2\alpha}\psi(\xi)$$

and

$$\frac{\mathcal{K}_{q^2}^{\mu/2-\alpha,\nu-\lambda} g(\xi)}{(1-q^2)^{\nu-\lambda+1}} = \frac{\mathcal{K}_{q^2}^{\mu/2-\alpha,\nu-\lambda} S_{q^2}^{\nu/2-\beta,2\beta}\psi(\xi)}{(1-q^2)^{\nu-\lambda+1}} = S_{q^2}^{\mu/2-\alpha,\lambda-\mu+2\alpha}\psi(\xi).$$

Thus, the last two identities can be described by

$$S_{q^2}^{\mu/2-\alpha,\lambda-\mu+2\alpha}\psi(\xi) = h(\xi) \quad (\xi \in R_{q^2,+}), \tag{5-4}$$

where h is the function defined by

$$h(\xi) := \begin{cases}(1-q^2)^{\mu-\lambda-1} I_{q^2}^{\mu/2+\alpha,\lambda-\mu} f(\xi) & \text{if } \xi \in A_{q^2},\\ (1-q^2)^{\lambda-\nu-1}\mathcal{K}_{q^2}^{\mu/2-\alpha,\nu-\lambda} g(\xi) & \text{if } \xi \in B_{q^2}.\end{cases}$$

Thus, applying the inversion formula in Proposition 3.6 yields

$$\psi(\xi) = S_{q^2}^{-\mu/2+\alpha+\lambda,-\lambda+\mu-2\alpha} h(\xi).$$

In other words,

$$\begin{aligned}\psi(\xi) = (1-q^2)^{\lambda-\nu+2\alpha-2}\xi^{\lambda/2-\mu/2+\alpha}\int_0^1 J_\lambda(\sqrt{\rho\xi};q^2) I_{q^2}^{\mu/2+\alpha,\lambda-\mu} f(\rho)\, d_{q^2}\rho\\ + (1-q^2)^{\lambda-\nu-2}\xi^{\lambda/2-\mu/2+\alpha}\int_1^\infty J_\lambda(\sqrt{\rho\xi};q^2)\mathcal{K}_{q^2}^{\lambda/2-\nu/2-\beta,\nu-\lambda} g(\rho)\, d_{q^2}\rho.\end{aligned} \tag{5-5}$$

From (2-18), the q^2-integral on $[1, \infty)$ in (5-5) is absolutely convergent, and from Proposition 3.3, one can verify that if $\Re(\lambda-\mu-2\alpha) > 0$ then the q^2-integral on $[0, 1]$ in (5-5) is absolutely convergent. Hence, the function ψ is well defined. Also, using (2-18),

$$\psi \in L_{q^2,\mu/2-\alpha}(\mathbb{R}_{q^2,+}) \cap L_{q^2,\nu/2-\beta}(\mathbb{R}_{q^2,+}) \cap L_{q^2,\nu/2-\beta-\gamma}(\mathbb{R}_{q^2,+})$$

when γ satisfies the condition in (5-2). □

Example 5.2. If $\nu=\frac{1}{2}$, $\mu=-\frac{1}{2}$, $\alpha=0$, and $\beta=\frac{1}{2}$, then $\lambda=\frac{1}{2}$ and the solution of the system

$$\frac{\sqrt{1-q^2}\rho^{-1/4}}{\Gamma_{q^2}(1/2)}\int_0^\infty \xi^{-1/4}\psi(\xi)\cos\biggl(\frac{\sqrt{q\rho\xi}}{1-q};q\biggr)\,d_{q^2}\xi=f(\rho),\quad \rho\in A_{q^2},$$
$$\frac{\sqrt{1-q^2}\rho^{-3/4}}{\Gamma_{q^2}(1/2)}\int_0^\infty \xi^{-3/4}\psi(\xi)\sin\biggl(\frac{\sqrt{\rho\xi}}{1-q};q\biggr)\,d_{q^2}\xi=g(\rho),\quad \rho\in B_{q^2},$$

is

$$\psi(\xi)=\frac{\xi^{1/4}}{(1-q^2)^{3/2}\Gamma_{q^2}(1/2)}\int_0^1 \rho^{-1/4}\sin\biggl(\frac{\sqrt{\rho\xi}}{1-q};q\biggr)I_{q^2}^{-1/4,1}f(\rho)\,d_{q^2}\rho \\ +\frac{\xi^{1/4}}{(1-q^2)^{3/2}\Gamma_{q^2}(1/2)}\int_1^\infty \rho^{-1/4}\sin\biggl(\frac{\sqrt{\rho\xi}}{1-q};q\biggr)g(\rho)\,d_{q^2}\rho.$$

The solution is in $L_{q^2,-1/4}(\mathbb{R}_{q^2,+})\cap L_{q^2,-1/4-\gamma}(\mathbb{R}_{q^2,+})$, where $0<\Re(\gamma)<\frac{3}{2}$.

In particular, if $f(\rho)=g(\rho)=\rho^{-1/4}$, then

$$\psi(\xi)=\frac{\xi^{1/4}}{(1-q)\sqrt{1-q^2}\Gamma_{q^2}(3/2)}\int_0^1 \sin\biggl(\frac{\sqrt{\xi}\rho}{1-q};q\biggr)\,d_q\rho \\ +\frac{\xi^{1/4}}{(1-q)\sqrt{1-q^2}\Gamma_{q^2}(1/2)}\int_1^\infty \sin\biggl(\frac{\sqrt{\xi}\rho}{1-q};q\biggr)\,d_q\rho;$$

and by applying (2-15) and Proposition 2.10, we obtain

$$\psi(\xi)=\frac{\xi^{-1/4}}{\sqrt{1-q^2}}\biggl[\frac{q^{3/2}}{\Gamma_{q^2}(1/2)}-\frac{1}{\Gamma_{q^2}(3/2)}\biggr]\cos\biggl(\frac{\sqrt{\xi}}{(1-q)\sqrt{q}};q\biggr)+\frac{\xi^{-1/4}}{\sqrt{1-q^2}\Gamma_{q^2}(1/2)}.$$

6. q-Mellin transform method

Nasim [1986] showed that the dual integral equations

$$\int_0^\infty t^{-2\alpha}J_\nu(xt)[1+w(t)]\phi(t)\,dt=f(x),\quad 0<x<1,$$
$$\int_0^\infty t^{-2\beta}J_\mu(xt)\phi(t)\,dt=g(x),\quad 1<x<\infty,$$

where w is an arbitrary weight function, can be reduced to a single (rather complicated) Fredholm equation of the second kind by using the Mellin transform. In this section we give a q-analog of Nasim's problem where we employ the q-Mellin transform to reduce certain dual q-integral equations into a Fredholm q-integral equation of the second kind.

First, we write down some definitions and results which we use later on. Fitouhi et al. [2006] defined a q-analog of the Mellin transform through the identity

$$\mathcal{M}_q(f)(s) := \int_0^\infty t^{s-1} f(t)\, d_qt.$$

Let α and β be real numbers and $\mathcal{M}_q^{\alpha,\beta}$ be the space of all functions defined on $\mathbb{R}_{q,+}$ such that

$$f|_{A_q} \in L_{q,\alpha-1}(A_q) \quad \text{and} \quad f|_{B_q} \in L_{q,\beta-1}(B_q).$$

The next lemma includes a sufficient condition for the existence of the q-Mellin transform which is slightly different from the one introduced in [Fitouhi et al. 2006].

Lemma 6.1. *Let α, β be real numbers such that $\alpha \le \beta$. If $f \in \mathcal{M}_q^{\alpha,\beta}$, then the q-Mellin transform of the function f exists on the strip $\alpha \le \Re(s) \le \beta$.*

Proof. Assume that $f \in \mathcal{M}_q^{\alpha,\beta}$. Then, using that

$$\mathcal{M}_q f(s) = \int_0^1 t^{s-1} f(t)\, d_qt + \int_1^\infty t^{s-1} f(t)\, d_qt,$$

and the inequalities

$$|t^s| \le t^\alpha \quad \text{for } \Re(s) \ge \alpha \text{ and all } t \in [0, 1],$$

and

$$|t^s| \le t^\beta \quad \text{for } \Re(s) \le \beta \text{ and all } t \in [1, \infty),$$

we obtain

$$|\mathcal{M}_q f(s)| \le \int_0^1 t^{\alpha-1} |f(t)|\, d_qt + \int_1^\infty t^{\beta-1} |f(t)|\, d_qt < \infty,$$

and the desired result follows. □

By $(\alpha_{q,f}, \beta_{q,f})$ we mean the fundamental (largest) strip on which the q-Mellin transform exists for $s \in \mathbb{C}$ such that $\alpha_{q,f} < \Re(s) < \beta_{q,f}$.

Lemma 6.2. *Let α, β be real numbers such that $\alpha < \beta$. If $f \in \mathcal{M}_q^{\alpha,\beta}$, then $\mathcal{M}_q f(s)$ is an analytic function on the simply connected domain defined by the strip $(\alpha_{q,f}, \beta_{q,f})$.*

Proof. If we set $F_n(s) := (1-q)\sum_{-n}^{n} q^{ks} f(q^k)$, we can verify that:

(1) $F_n(s)$ is an entire function for each $n \in \mathbb{N}_0$;

(2) $F_n(s)$ tends uniformly to $\mathcal{M}_q f(s)$ as n tends to ∞ for $\Re(s) \in (\alpha_{q,f}, \beta_{q,f})$.

Hence, $\mathcal{M}_q f(s)$ is an analytic function in the domain defined by the strip $\Re(s) \in (\alpha_{q,f}, \beta_{q,f})$. □

A direct consequence of the previous lemma is that

$$\int_C x^{-s} \mathcal{M}_q f(s)\, ds = (1-q) \sum_{-\infty}^{\infty} f(q^k) \int_C x^{-s} q^{ks}\, ds$$

for any contour that lies in the interior of the domain defined by the strip $\Re(s) \in (\alpha_{q,f}, \beta_{q,f})$. Fitouhi et al. [2006] chose C to be the contour that connects the points $c - i\pi/\log(q)$ and $c + i\pi/\log(q)$, where $c \in (\alpha_{q,f}, \beta_{q,f})$, to introduce and prove the q-Mellin inversion formula

$$f(x) = \frac{\log(q)}{2i\pi(1-q)} \int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} \mathcal{M}_q(f)(s) x^{-s}\, ds, \quad x \in \mathbb{R}_{q,+}. \tag{6-1}$$

In addition, they introduced a q-Parseval's formula for the Mellin transform, under suitable conditions on the functions f and g,

$$\int_0^\infty f(xt)g(t)\, d_qt = \frac{\log(q)}{2i\pi(1-q)} \int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} \mathcal{M}_q(f)(s)\mathcal{M}_q(g)(1-s)x^{-s}\, ds, \tag{6-2}$$

where c is in the fundamental strip of defining f, g. They also proved:

Theorem 6.3. *Let K and g be a pair of functions defined on $\mathbb{R}_{q,+}$ such that the strip $I_{K;g} = (\beta_{q,K}, \alpha_{q,K}) \cap (1-\beta_{q,g}, 1-\alpha_{q,g})$ is not empty. If*

$$f(x) = \int_0^\infty g(t)K(xt)\, d_qt,$$

then

$$\mathcal{M}_q(f)(s) = \mathcal{M}_q(K)(s)\mathcal{M}_q(g)(1-s), \quad s \in I_{K;g}.$$

The q-Mellin transform of the third Jackson q-Bessel has been calculated in [Fitouhi et al. 2006] by using the identity of Koornwinder and Swarttouw [1992, p. 449]:

$$\frac{(t^{-1}z;q)_\infty}{(tz;q)_\infty} = \sum_{-\infty}^{\infty} t^n z^n \frac{(z^2;q)_\infty}{(q;q)_\infty} {}_1\phi_1(0; z^2; q, q^{n+1}),$$

where t and z are complex numbers such that $0 < |t| < |z|^{-1}$. Using the same technique, we can prove that

$$\begin{aligned} &\mathcal{M}_q(z^\beta J_\alpha(z;q^2)(s)) \\ &\quad = (1-q)(1-q^2)^{s+\beta-1} \frac{\Gamma_{q^2}(\frac{1}{2}(\alpha+s+\beta))}{\Gamma_{q^2}(\frac{1}{2}(\alpha-s-\beta+2))}, \quad \Re(s) > -\Re(\alpha+\beta). \end{aligned} \tag{6-3}$$

The following two lemmas are needed in the sequel.

Lemma 6.4. *If* $\Re(s) < \nu + 2\alpha$ *and*

$$h_1^* := \mathcal{M}_q^{-1}\left(\frac{\Gamma_{q^2}(\frac{1}{2}\nu - \frac{1}{2}s + \alpha)}{\Gamma_{q^2}(\frac{1}{2}\mu - \frac{1}{2}s + \beta + 1)}\right),$$

then

$$h_1^*(x) = \frac{(1+q)x^{-\nu-2\alpha}}{\Gamma_{q^2}(-\frac{1}{2}\nu+\frac{1}{2}\mu-\alpha+\beta+1)}(q^2/x^2;q^2)_{-\nu/2+\mu/2-\alpha+\beta} \quad (x \in \mathbb{R}_{q,+}). \tag{6-4}$$

Proof. Applying (6-1) we obtain

$$\begin{aligned} h_1^*(x) &= \frac{\log(q)}{2i\pi(1-q)}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} \mathcal{M}_q(h_1^*)(s)x^{-s}\,ds \\ &= \frac{\log(q)}{2i\pi(1-q)}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} x^{-s}\frac{\Gamma_{q^2}(\frac{1}{2}\nu - \frac{1}{2}s + \alpha)}{\Gamma_{q^2}(\frac{1}{2}\mu - \frac{1}{2}s + \beta + 1)}\,ds \\ &= \frac{\log(q)(1-q^2)^{\mu/2-\nu/2-\alpha+\beta+1}}{2i\pi(1-q)}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} x^{-s}\frac{(q^{\mu-s+2\beta+2};q^2)_\infty}{(q^{\nu-s+2\alpha};q^2)_\infty}\,ds. \end{aligned}$$

Applying the q-binomial theorem (2-13) with $z = q^{\nu-s+2\alpha}$ and for $\Re(s) < \nu + 2\alpha$, we obtain

$$\begin{aligned} h_1^*(x) = \frac{\log(q)(1-q^2)^{\mu/2-\nu/2-\alpha+\beta+1}}{2i\pi(1-q)}\sum_{n=0}^{\infty}\frac{(q^{\mu-\nu-2\alpha+2\beta+2};q^2)_n}{(q^2;q^2)_n}q^{n(\nu+2\alpha)} \\ \times\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)}(q^n x)^{-s}\,ds. \end{aligned}$$

Therefore, if $x := q^m$ $(m \in \mathbb{Z})$,

$$\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} q^{-s(n+m)}\,ds = \begin{cases} 2i\pi/\log(q) & \text{if } m = -n, \\ 0 & \text{if } m \neq -n. \end{cases}$$

Hence, $h_1^*(q^m) = 0$ for $m \in \mathbb{N}$; and for $m \in -\mathbb{N}_0$,

$$\begin{aligned} h_1^*(q^m) &= \frac{q^{-m(\nu+2\alpha)}(1-q^2)^{\mu/2-\nu/2-\alpha+\beta+1}}{1-q}\frac{(q^{\mu-\nu-2\alpha+2\beta+2};q^2)_{-m}}{(q^2;q^2)_{-m}} \\ &= \frac{q^{-m(\nu+2\alpha)}(1+q)}{\Gamma_{q^2}(\frac{1}{2}\mu - \frac{1}{2}\nu - \alpha + \beta + 1)}(q^{2-2m};q^2)_{\mu/2-\nu/2-\alpha+\beta}. \end{aligned}$$

Thus $h_1^*(x)$ is given by (6-4) for all $x \in \mathbb{R}_{q,+}$. □

Lemma 6.5. *If* $\Re(s) > -\nu + 2\alpha + 2$ *and*

$$h_2^* := \mathcal{M}_q^{-1}\left(\frac{\Gamma_{q^2}(\frac{1}{2}\nu + \frac{1}{2}s - \alpha - 1)}{\Gamma_{q^2}(\frac{1}{2}\mu + \frac{1}{2}s - \beta)}\right),$$

then, for $x \in \mathbb{R}_{q,+}$,

$$h_2^*(x) = \frac{(1+q)x^{\nu-2\alpha-2}}{\Gamma_{q^2}(-\frac{1}{2}\nu+\frac{1}{2}\mu+\alpha-\beta+1)}(q^2x^2;q^2)_{-\nu/2+\mu/2+\alpha-\beta}. \tag{6-5}$$

Proof. The proof is similar to the proof of Lemma 6.4 and is omitted. □

In the following, we use the q-Mellin transform to reduce the dual q-integral equations

$$\begin{aligned} \int_0^\infty u^{-2\alpha}\psi(u)[1+w(u)]J_\nu(u\rho;q^2)\,d_qu &= f(\rho), \quad \rho \in A_q, \\ \int_0^\infty u^{-2\beta}\psi(u)J_\mu(u\rho;q^2)\,d_qu &= g(\rho), \quad \rho \in B_q, \end{aligned} \tag{6-6}$$

to a Fredholm q-integral equation of the second kind. Before we start our mission, we set the following notation. Let $\{H_1, H_1^*\}$ and $\{H_2, H_2^*\}$ be the pair of functions defined by

$$H_1(s) := \mathcal{M}_q(u^{-2\alpha}J_\nu(u;q^2))(s) = (1-q)(1-q^2)^{s-2\alpha-1}\frac{\Gamma_{q^2}(\frac{1}{2}\nu+\frac{1}{2}s-\alpha)}{\Gamma_{q^2}(\frac{1}{2}\nu-\frac{1}{2}s+\alpha+1)},$$

$$H_1^*(s) := (1-q^2)^{\nu/2-\mu/2+\alpha-\beta}\frac{\Gamma_{q^2}(\frac{1}{2}\nu-\frac{1}{2}s+\alpha+1)}{\Gamma_{q^2}(\frac{1}{2}\mu-\frac{1}{2}s+\beta+1)},$$

where $2\alpha-\nu < \Re(s) < 2+2\alpha+\nu$ and

$$H_2(s) := \mathcal{M}_q(u^{-2\beta}J_\mu(u;q^2))(s) = (1-q)(1-q^2)^{s-2\beta-1}\frac{\Gamma_{q^2}(\frac{1}{2}\mu+\frac{1}{2}s-\beta)}{\Gamma_{q^2}(\frac{1}{2}\mu-\frac{1}{2}s+\beta+1)},$$

$$H_2^*(s) := (1-q^2)^{\nu/2-\mu/2-\alpha+\beta}\frac{\Gamma_{q^2}(\frac{1}{2}\nu+\frac{1}{2}s-\alpha)}{\Gamma_{q^2}(\frac{1}{2}\mu+\frac{1}{2}s-\beta)},$$

where $\Re(s) > \max\{2\beta-\mu,\ 2\alpha-\nu\}$. It is worth noting that for

$$\max\{2\alpha-\nu, 2\beta-\mu\} < \Re(s) < 2+2\alpha+\nu,$$

we have

$$H_1(s)H_1^*(s) = H_2(s)H_2^*(s) = K(s), \tag{6-7}$$

where $K(s)$ is the function defined for $\Re(s) > 2\alpha-\nu$ by

$$K(s) := \mathcal{M}_q(k(u))(s), \quad k(u) := u^{\nu/2-\mu/2-\alpha-\beta}J_{\nu/2+\mu/2-\alpha+\beta}(u;q^2).$$

Theorem 6.6. *Let* α, β, γ, *and* ν *be real parameters satisfying*

$$\nu > 0 \quad \textit{and} \quad \max\{2\alpha-\nu, 2\beta-\mu\} < 2+2\alpha+\nu.$$

Let f and g be functions defined on A_q and B_q, respectively. Let f_1 and g_1 be the extensions of f and g defined on $\mathbb{R}_{q,+}$ by

$$\begin{aligned} &f_1 \equiv (\,\cdot\,)^{-2\alpha} f(\,\cdot\,) \text{ on } A_q, && f_1 \equiv 0 \text{ on } B_q,\\ &g_1 \equiv 0 \text{ on } A_q, && g_1 \equiv (\,\cdot\,)^{-2\beta} g(\,\cdot\,) \text{ on } B_q. \end{aligned}$$

Assume that the q-Mellin transforms of the functions ψ, ψw, f_1, and g_1 exist on Ω_1, $\Omega_1 \neq \phi$, where

$$\Omega_1 := (\alpha_{q,f_1}, \beta_{q,f_1}) \cap (\alpha_{q,g_1}, \beta_{q,g_1}) \cap (1-\beta_{q,\psi}, 1-\alpha_{q,\psi}) \cap (1-\beta_{q,w\psi}, 1-\alpha_{q,w\psi}).$$

Set

$$\eta := -\tfrac{1}{4}\nu + \tfrac{3}{4}\mu + \tfrac{1}{2}\alpha + \tfrac{3}{2}\beta + 1, \quad \lambda := \nu - \mu - 2\alpha - 2\beta - 2.$$

Then the system (6-6) *can be reduced to a Fredholm q-integral equation of the form*

$$\psi(t) = A(t) - t^{\mu/2-\nu/2+\alpha+\beta+1} \int_0^\infty u^{\nu/2-\mu/2-\alpha-\beta} \psi(u) w(u) L(t,u)\, d_q u,$$

where

$$\begin{aligned} A(t) = {} & (1-q^2)^{\nu/2-\mu/2+\alpha-\beta+1} \\ & \quad \times S_q^{\eta,\lambda}\big(H(1-\rho)\rho^{-\mu-2\beta-2} I_{q^2}^{-\nu/2+\mu/2-\alpha+\beta}[(\,\cdot\,)^{\nu/2} f(\sqrt{\cdot}\,)](\rho^2)\big) \\ & + (1-q^2)^{\nu/2-\mu/2+\alpha-\beta+1} \\ & \quad \times S_q^{\eta,\lambda}\big(H(\rho-q^{-1})\rho^{-2\beta} \mathcal{K}_{q^2}^{\nu/2-\alpha+\beta,-\nu/2+\mu/2+\alpha-\beta}[g(q^{-1}\sqrt{\cdot}\,)](\rho^2)\big); \end{aligned}$$

$H(x-a)$ is the Heaviside function defined to be 1 *if $x \geq 0$ and* 0 *otherwise*; *and*

$$\begin{aligned} L(t,u) = {} & \frac{(1-q)q^{\nu/2+\mu/2-\alpha+\beta}}{t^2-u^2} \\ & \times \big[t J_{\nu/2+\mu/2-\alpha+\beta+1}(t;q^2) J_{\nu/2+\mu/2-\alpha+\beta}(q^{-1}u;q^2) \\ & \qquad - u J_{\nu/2+\mu/2-\alpha+\beta+1}(u;q^2) J_{\nu/2+\mu/2-\alpha+\beta}(q^{-1}t;q^2)\big]. \end{aligned}$$

Proof. Assume that Ψ, Φ, F_1, and G_1 are the q-Mellin transform of the functions ψ, ψw, f_1, and g_1, respectively, on Ω_1. Let

$$\Omega_2 := \{s \in \mathbb{C} : \Re s \in \Omega_1 \text{ and } \Re s > \max\{2\alpha-\nu, 2\beta-\nu\}\}.$$

Applying Theorem 6.3 gives

$$H_1(s)[\Psi(1-s) + \Phi(1-s)] = F_1(s), \quad H_2(s)\Psi(1-s) = G_1(s). \tag{6-8}$$

Multiplying the first equation in (6-8) by $H_1^*(s)$ and the second by $H_2^*(s)$ yields

$$\begin{aligned} K(s)\Psi(1-s) &= H_1^*(s)[F_1(s) - H_1(s)\Phi(1-s)], \\ K(s)\Psi(1-s) &= H_2^*(s)G_1(s), \end{aligned} \tag{6-9}$$

for $\Re(s) \in \Omega := \Omega_2 \cap \{s \in \mathbb{C} : \max\{2\alpha-\nu, 2\beta-\mu\} < \Re(s) < 2+2\alpha+\nu\}$, where we used (6-7).

In the following we calculate the value of the q-integral

$$\int_0^\infty k(\rho t)\psi(t)\,d_qt \quad (\rho \in \mathbb{R}_{q,+}). \tag{6-10}$$

We distinguish two cases:

Case 1. $\rho \in A_q$. In this case, from (6-2) and (6-9), we obtain

$$\begin{aligned}\int_0^\infty & k(\rho t)\psi(t)\,d_qt \\ &= \frac{\log(q)}{2i\pi(1-q)}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} K(s)\Psi(1-s)\rho^{-s}\,ds, \\ &= \frac{\log(q)}{2i\pi(1-q)}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} H_1^*(s)[F_1(s)-H_1(s)\Phi(1-s)]\rho^{-s}\,ds.\end{aligned} \tag{6-11}$$

Substituting

$$\rho^{-s} = \frac{1-q}{1-q^{\nu-s+2\alpha}}\rho^{-\nu-2\alpha+1}D_{q,\rho}\rho^{\nu-s+2\alpha}$$

into the third line of (6-11), we obtain

$$\begin{aligned}\int_0^\infty & k(\rho t)\psi(t)\,d_qt \\ &= \frac{\log(q)\rho^{-\nu-2\alpha+1}}{2i\pi(1-q^2)^{-\nu/2+\mu/2-\alpha+\beta+1}}D_{q,\rho} \\ &\quad\times\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)}[F_1(s)-H_1(s)\Phi(1-s)]\rho^{\nu-s+2\alpha}\frac{\Gamma_{q^2}(\frac{1}{2}\nu-\frac{1}{2}s+\alpha)}{\Gamma_{q^2}(\frac{1}{2}\mu-\frac{1}{2}s+\beta+1)}\,ds \\ &= \frac{\log(q)\rho^{-\nu-2\alpha+1}}{2i\pi(1-q^2)^{-\nu/2+\mu/2-\alpha+\beta+1}}D_{q,\rho}\rho^{\nu+2\alpha} \\ &\qquad\times\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)}[F_1(s)-L(s)]\mathcal{H}_1^*(s)\rho^{-s}\,ds \\ &= \frac{(1-q)\rho^{-\nu-2\alpha+1}}{(1-q^2)^{-\nu/2+\mu/2-\alpha+\beta+1}}D_{q,\rho}\rho^{\nu+2\alpha}\int_0^\infty[f_1(t)-l(t)]h_1^*\Big(\frac{\rho}{t}\Big)\frac{1}{t}\,d_qt,\end{aligned} \tag{6-12}$$

where $\mathcal{H}_1^*(s) := \dfrac{\Gamma_{q^2}(\frac{1}{2}\nu-\frac{1}{2}s+\alpha)}{\Gamma_{q^2}(\frac{1}{2}\mu-\frac{1}{2}s+\beta+1)}$, $h_1^*(x)$ is given by Lemma 6.4, and

$$l(\rho) = \mathcal{M}_q^{-1}(H_1(s)\Phi(1-s)).$$

On the other hand, using the q-Mellin transform inversion formula and the q-Parseval relation in (6-2), we obtain

$$\begin{aligned} l(\rho) &= \frac{\log(q)}{2i\pi(1-q)} \int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} H_1(s)\Phi(1-s)\rho^{-s}\,ds \\ &= \rho^{-2\alpha}\int_0^\infty u^{-2\alpha}\psi(u)w(u)J_\nu(u\rho;q^2)\,d_qu. \end{aligned}$$

On simplifying (6-12), using (6-4), we obtain, for $\rho \in A_q$,

$$\begin{aligned} \int_0^\infty k(\rho t)\psi(t)\,d_qt &= \frac{(1-q^2)^{\nu/2-\mu/2+\alpha-\beta}\rho^{-\nu-2\alpha+1}}{\Gamma_{q^2}(-\frac{1}{2}\nu+\frac{1}{2}\mu-\alpha+\beta+1)} D_{q,\rho} \\ &\quad\times \int_0^\rho t^{\nu+2\alpha-1}[\rho^{-2\alpha}f(t)-l(t)](q^2t^2/\rho^2;q^2)_{-\nu/2+\mu/2-\alpha+\beta}\,d_qt. \end{aligned} \tag{6-13}$$

Hence,

$$\begin{aligned} &\int_0^\infty k(\rho t)\psi(t)\,d_qt \\ &= \frac{(1-q^2)^{\nu/2-\mu/2+\alpha-\beta}\rho^{-\nu-2\alpha+1}}{\Gamma_{q^2}(-\frac{1}{2}\nu+\frac{1}{2}\mu-\alpha+\beta+1)} D_{q,\rho}\biggl[\int_0^\rho t^{\nu-1}f(t)(q^2t^2/\rho^2;q^2)_{-\nu/2+\mu/2-\alpha+\beta}\,d_qt \\ &\quad -\int_0^\rho t^{\nu-1}(q^2t^2/\rho^2;q^2)_{-\nu/2+\mu/2-\alpha+\beta}\int_0^\infty u^{-2\alpha}\psi(u)w(u)J_\nu(ut;q^2)\,d_qu\,d_qt\biggr]. \end{aligned}$$

Since $\nu > 0$, we can apply Proposition 2.9 to obtain

$$\begin{aligned} &\int_0^\rho t^{\nu-1}(q^2t^2/\rho^2;q^2)_{-\nu/2+\mu/2-\alpha+\beta}J_\nu(ut;q^2)\,d_qt \\ &\qquad\qquad = \frac{\rho^{2\nu}u^\nu(1-q)(q^{\mu+\nu-2\alpha+2\beta+2};q^2)_\infty}{(1-q^{2\nu})(q^{\mu-\nu-2\alpha+2\beta+2};q^2)_\infty} \\ &\qquad\qquad\qquad \times {}_2\phi_2(0,q^{2\nu};q^{2\nu+2},q^{\nu+\mu-2\alpha+2\beta+2};q^2,q^2\rho^2u^2). \end{aligned}$$

Therefore,

$$\int_0^\infty k(\rho t)\psi(t)\,d_qt = m_1(\rho) = I_1 - I_2, \tag{6-14}$$

where

$$I_1 := \frac{(1-q^2)^{\nu/2-\mu/2+\alpha-\beta}\rho^{-\nu-2\alpha+1}}{\Gamma_{q^2}(-\frac{1}{2}\nu+\frac{1}{2}\mu-\alpha+\beta+1)} D_{q,\rho}\int_0^\rho t^{\nu-1}\biggl(\frac{q^2t^2}{\rho^2};q^2\biggr)_{-\nu/2+\mu/2-\alpha+\beta} f(t)\,d_qt$$

and

$$I_2 := \frac{(1-q)(q^{\nu+\mu-2\alpha+2\beta+2};q^2)_\infty}{(1-q^{2\nu})(q^2;q^2)_\infty}\rho^{1-\nu-2\alpha}$$
$$\times D_{q,\rho}\rho^{2\nu}\int_0^\infty u^{\nu-2\alpha}\psi(u)w(u)\,{}_2\phi_2(0,q^{2\nu};q^{2\nu+2},q^{\nu+\mu-2\alpha+2\beta+2};q^2,q^2\rho^2u^2)\,d_qu.$$

Using (2-14) and (2-12), we obtain

$$D_{q,\rho}\rho^{2\nu}{}_2\phi_2(0,q^{2\nu};q^{2\nu+2},q^{\nu+\mu-2\alpha+2\beta+2};q^2,q^2\rho^2u^2)$$
$$=\frac{(q^2;q^2)_\infty(1-q^{2\nu})\rho^{3\nu/2-\mu/2+\alpha-\beta-1}u^{-\nu/2-\mu/2+\alpha-\beta}}{(q^\nu+\mu-2\alpha+2\beta+2;q^2)_\infty(1-q)}J_{\nu/2+\mu/2-\alpha+\beta}(\rho u;q^2).$$

Thus,

$$I_2=\rho^{\nu/2-\mu/2-\alpha-\beta}\int_0^\infty u^{\nu/2-\mu/2-\alpha-\beta}\psi(u)w(u)J_{\nu/2+\mu/2-\alpha+\beta}(\rho u;q^2)\,d_qu.$$

Applying [Annaby and Mansour 2012, Lemma 1.12], one can verify that

$$\text{(6-15)}\quad D_{q,\rho}\int_0^\rho t^{\nu-1}\left(\frac{q^2t^2}{\rho^2};q^2\right)_{-\nu/2+\mu/2-\alpha+\beta}f(t)\,d_qt$$
$$=\int_0^\rho t^{\nu-1}D_{q,\rho}\left(\frac{q^2t^2}{\rho^2};q^2\right)_{-\nu/2+\mu/2-\alpha+\beta}f(t)\,d_qt.$$

Since $D_{q,\rho}(c/\rho^2;q^2)_\alpha=\dfrac{q^{-2}c}{\rho^3}\dfrac{1-q^{2\alpha}}{1-q}(c/\rho^2;q^2)_{\alpha-1}$ and

$$\int_0^\rho f(t)\,d_qt=\frac{1}{1+q}\int_0^{\rho^2}\frac{f(\sqrt{t})}{\sqrt{t}}\,d_qt,$$

we obtain

$$\text{(6-16)}\quad D_{q,\rho}\int_0^\rho t^{\nu-1}\left(\frac{q^2t^2}{\rho^2};q^2\right)_{-\nu/2+\mu/2-\alpha+\beta}f(\sqrt{t})\,d_qt$$
$$=\frac{1-q^{-\nu+\mu-2\alpha+2\beta}}{(1-q^2)\rho^3}\int_0^{\rho^2}t^{\nu/2}\left(\frac{q^2t}{\rho^2};q^2\right)_{-\nu/2+\mu/2-\alpha+\beta-1}f(\sqrt{t})\,d_{q^2}t.$$

Hence, from (3-12),

$$I_1=(1-q^2)^{\nu/2-\mu/2+\alpha-\beta}\rho^{-\mu-2\beta-2}I_{q^2}^{-\nu/2+\mu/2-\alpha+\beta}[(\cdot)^{\nu/2}f(\sqrt{\cdot})](\rho^2).$$

The q-integral in (6-14) is now

$$\text{(6-17)}\quad m_1(\rho)=(1-q^2)^{\nu/2-\mu/2+\alpha-\beta}\rho^{-\mu-2\beta-2}I_{q^2}^{-\nu/2+\mu/2-\alpha+\beta}[(\cdot)^{\nu/2}f(\sqrt{\cdot})](\rho^2)$$
$$-\rho^{\nu/2-\mu/2-\alpha-\beta}\int_0^\infty u^{\nu/2-\mu/2-\alpha-\beta}\psi(u)w(u)J_{\nu/2+\mu/2-\alpha+\beta}(\rho u;q^2)\,d_qu.$$

Case 2. $\rho \in B_q$. From (6-2) and (6-9), we obtain

$$\text{(6-18)}\quad \int_0^\infty k(\rho t)\psi(t)\,d_qt = \frac{\log(q)}{2i\pi(1-q)}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} K(s)\Psi(1-s)\rho^{-s}\,ds$$
$$= \frac{\log(q)}{2i\pi(1-q)}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} H_2^*(s)G_1(s)\rho^{-s}\,ds.$$

Substituting

$$\rho^{-s} = \rho^{\nu-2\alpha-1}\frac{1-q}{1-q^{-s-\nu+2\alpha+2}}D_{q,\rho}\rho^{-s-\nu+2\alpha+2}$$

in the second line of (6-18), we obtain

$$\text{(6-19)}\quad \int_0^\infty k(\rho t)\psi(t)\,d_qt = m_2(\rho)$$
$$= -\frac{\log(q)q^{\nu-2\alpha-2}\rho^{\nu-2\alpha-1}}{2i\pi(1-q^2)^{-\nu/2+\mu/2+\alpha-\beta+1}}$$
$$\times D_{q,\rho}\rho^{-\nu+2\alpha+2}\int_{c-i\pi/\log(q)}^{c+i\pi/\log(q)} G_1(s)\mathcal{H}_2^*(s)q^s\rho^{-s}\,ds$$
$$= -(1-q)(1-q^2)^{\nu/2-\mu/2-\alpha+\beta-1}q^{\nu-2\alpha-2}\rho^{\nu-2\alpha-1}$$
$$\times D_{q,\rho}\rho^{-\nu+2\alpha+2}\int_0^\infty g_1(t)h_2^*\Big(\frac{\rho}{qt}\Big)\frac{1}{t}\,d_qt,$$

where $\mathcal{H}_2^*(s) := \Gamma_{q^2}(\frac{1}{2}\nu+\frac{1}{2}s-\alpha-1)/\Gamma_{q^2}(\frac{1}{2}\mu+\frac{1}{2}s-\beta)$ and $h_2^*(\rho)$ is given by Lemma 6.5. On simplifying (6-19), using (6-5), we get

$$m_2(\rho) = -\frac{(1-q^2)^{\nu/2-\mu/2-\alpha+\beta}\rho^{\nu-2\alpha-1}}{\Gamma_{q^2}(-\frac{1}{2}\nu+\frac{1}{2}\mu+\alpha-\beta+1)}$$
$$\times D_{q,\rho}\int_\rho^\infty \frac{g(u)}{u^{\nu-2\alpha+2\beta-1}}\Big(\frac{\rho^2}{u^2};q^2\Big)_{-\nu/2+\mu/2+\alpha-\beta}\,d_qu.$$

In similar steps as in simplifying I_1 and by (3-1), we obtain

$$\text{(6-20)}\quad m_2(\rho)$$
$$= -(1-q^2)^{\nu/2-\mu/2-\alpha+\beta}\rho^{-2\beta}\mathcal{K}_{q^2}^{\nu/2-\alpha+\beta,-\nu/2+\mu/2+\alpha-\beta}[g(q^{-1}\sqrt{\cdot})](\rho^2).$$

If we set $m(\rho)$ to be

$$m(\rho) = \begin{cases} m_1(\rho) & \text{if } \rho \in A_q, \\ m_2(\rho) & \text{if } \rho \in B_q, \end{cases}$$

then combining (6-17) and (6-20) gives

$$\text{(6-21)}\quad \rho^{\nu/2-\mu/2-\alpha-\beta}\int_0^\infty t^{\nu/2-\mu/2-\alpha-\beta}J_{\nu/2+\mu/2-\alpha+\beta}(\rho t;q^2)\psi(t)\,d_qt = m(\rho).$$

Hence, from (2-20) and (6-21), we obtain

$$\text{(6-22)}\quad \psi(t)=t^{-\nu/2+\mu/2+\alpha+\beta+1}\\ \times\int_0^\infty \rho^{-\nu/2+\mu/2+\alpha+\beta+1}J_{\nu/2+\mu/2-\alpha+\beta}(\rho t;q^2)m(\rho)\,d_q\rho.$$

Substituting for $m(\rho)$ in (6-22), we obtain

$$\psi(t)=t^{-\nu/2+\mu/2+\alpha+\beta+1}\int_0^1 \rho^{-\nu/2+\mu/2+\alpha+\beta+1}J_{\nu/2+\mu/2-\alpha+\beta}(\rho t;q^2)m_1(\rho)\,d_q\rho\\ +t^{-\nu/2+\mu/2+\alpha+\beta+1}\int_1^\infty \rho^{-\nu/2+\mu/2+\alpha+\beta+1}J_{\nu/2+\mu/2-\alpha+\beta}(\rho t;q^2)m_2(\rho)\,d_q\rho.$$

Returning to (6-17) and (6-20), we obtain

$$\psi(t)=N_1+N_2-N_3,$$

where

$$N_1:=(1-q^2)^{\nu/2-\mu/2+\alpha-\beta}t^{-\nu/2+\mu/2+\alpha+\beta+1}\\ \times\int_0^1 \rho^{-\nu/2-\mu/2+\alpha-\beta-1}J_{\nu/2+\mu/2-\alpha+\beta}(\rho t;q^2)\\ \times I_{q^2}^{-\nu/2+\mu/2-\alpha+\beta}[(\,\cdot\,)^{\nu/2}f(\sqrt{\cdot}\,)](\rho^2)\,d_q\rho,$$

$$N_2:=-(1-q^2)^{\nu/2-\mu/2-\alpha+\beta}t^{-\nu/2+\mu/2+\alpha+\beta+1}\\ \times\int_1^\infty \rho^{-\nu/2+\mu/2+\alpha-\beta+1}J_{\nu/2+\mu/2-\alpha+\beta}(\rho t;q^2)\\ \times \mathcal{K}_{q^2}^{\nu/2-\alpha+\beta,-\nu/2+\mu/2+\alpha-\beta}[g(q^{-1}\sqrt{\cdot}\,)](\rho^2)\,d_q\rho,$$

$$N_3:=t^{-\nu/2+\mu/2+\alpha+\beta+1}\\ \times\int_0^1\biggl(\rho J_{\nu/2+\mu/2-\alpha+\beta}(\rho t;q^2)\int_0^\infty u^{\nu/2-\mu/2-\alpha-\beta}\psi(u)w(u)\\ \times J_{\nu/2+\mu/2-\alpha+\beta}(\rho u;q^2)\,d_qu\biggr)\,d_q\rho.$$

Assuming that the weight function w satisfies $\psi w\in L_q^1(\mathbb{R}_{q,+})$, then the double q-integrals defining N_3 is absolutely convergent, and therefore we can interchange the order of q-integration to obtain

$$\text{(6-23)}\quad \psi(t)=A(t)-t^{\mu/2-\nu/2+\alpha+\beta+1}\int_0^\infty u^{\nu/2-\mu/2-\alpha-\beta}\psi(u)w(u)L(t,u)\,d_qu,$$

where $L(t,u)$ comes from (2-19). Thus, the solution of the single integral equation in (6-23) gives us the value of the unknown function $\psi(t)$, which is the solution of the dual q-integral equation in (6-6), as well. □

In particular, if we replace α by $-\alpha$ in (6-6), and set $w = g \equiv 0$, $\beta = 0$, and $\nu = \mu$, we obtain the dual q-integral equations (4-2)–(4-3), and from Theorem 6.6, its solution is given by

$$\psi(t) = \frac{t^{1-\alpha}}{(1-q^2)^{\alpha}} \int_0^1 \rho^{1-\nu-\alpha} J_{\nu+\alpha}(\rho t; q^2) I_{q^2}^{\alpha}[(\cdot)^{\nu/2} f(\sqrt{\cdot})](\rho^2)\, d_q\rho. \tag{6-24}$$

This coincides with the result of Section 4.

Example 6.7. If $\nu = \frac{1}{2}$, $\mu = -\frac{1}{2}$, $\alpha = \beta = -\frac{1}{4}$, then the solution of the system

$$\frac{\sqrt{1-q^2}\rho^{-1/2}}{\Gamma_{q^2}(1/2)} \int_0^{\infty} \phi(u)[1+w(u)] \sin\Bigl(\frac{u\rho}{1-q}; q\Bigr)\, d_q u = f(\rho), \quad \rho \in A_q,$$

$$\frac{\sqrt{1-q^2}\rho^{-1/2}}{\Gamma_{q^2}(1/2)} \int_0^{\infty} \phi(u) \cos\Bigl(\frac{u\rho\sqrt{q}}{1-q}; q\Bigr)\, d_q u = g(\rho), \quad \rho \in B_q,$$

takes the form

$$\psi(t) = A(t) - \frac{1}{(1-q)} \int_0^{\infty} u\psi(u) w(u) L(t,u)\, d_q u,$$

where

$$L(t,u) = \frac{1-q}{t^2-u^2}[tJ_1(t;q^2)J_0(q^{-1}u;q^2) - uJ_1(u;q^2)J_0(q^{-1}t;q^2)]$$

and

$$\begin{aligned} A(t) = (1-q^2)^{1/2}\biggl[\int_0^1 \rho^{-1} J_0(\rho t; q^2) I_{q^2}^{-1/2}[(\cdot)^{1/4} f(\sqrt{\cdot})](\rho^2)\, d_q\rho \\ - \int_1^{\infty} J_0(\rho t; q^2) \mathcal{K}_{q^2}^{1/4,-1/2}[g(q^{-1}\sqrt{\cdot})](\rho^2)\, d_q\rho\biggr]. \end{aligned}$$

Acknowledgements

Zeinab Mansour would like to acknowledge her gratitude to professor Mourad Ismail for inviting her to visit him at the City University of Hong Kong, where a part of this work was carried out. This research is supported by the DSFP program of the King Saud University in Riyadh through grant DSFP/MATH 01.

References

[Agarwal 1969] R. P. Agarwal, "Certain fractional q-integrals and q-derivatives", *Proc. Cambridge Philos. Soc.* **66** (1969), 365–370. MR 40 #657 Zbl 0179.16901

[Al-Salam 1966] W. A. Al-Salam, "Some fractional q-integrals and q-derivatives", *Proc. Edinburgh Math. Soc.* (2) **15** (1966), 135–140. MR 36 #1932 Zbl 0171.10301

[Andrews et al. 1999] G. E. Andrews, R. Askey, and R. Roy, *Special functions*, Encyclopedia Math. Appl. **71**, Cambridge University Press, 1999. MR 2000g:33001 Zbl 0920.33001

[Annaby and Mansour 2012] M. H. Annaby and Z. S. Mansour, *q-fractional calculus and equations*, Lecture Notes in Math. **2056**, Springer, Heidelberg, 2012. MR 2963764 Zbl 1267.26001

[Busbridge 1938] I. W. Busbridge, "Dual integral equations", *Proc. London Math. Soc.* (2) **44** (1938), 115–129. MR 1576205 Zbl 0019.02801

[Butzer and Westphal 2000] P. L. Butzer and U. Westphal, "An introduction to fractional calculus", pp. 1–85 in *Applications of fractional calculus in physics*, edited by R. Hilfer, World Scientific, River Edge, NJ, 2000. MR 2003g:26007 Zbl 0987.26005

[Copson 1961] E. T. Copson, "On certain dual integral equations", *Proc. Glasgow Math. Assoc.* **5** (1961), 21–24. MR 33 #7803 Zbl 0158.12901

[Erdélyi 1951] A. Erdélyi, "On some functional transformations", *Univ. e Politecnico Torino. Rend. Sem. Mat.* **10** (1951), 217–234. MR 13,937c Zbl 0044.11001

[Erdélyi and Kober 1940] A. Erdélyi and H. Kober, "Some remarks on Hankel transforms", *Quart. J. Math.* **11** (1940), 212–221. MR 2,192a Zbl 0025.18601

[Erdélyi and Sneddon 1962] A. Erdélyi and I. N. Sneddon, "Fractional integration and dual integral equations", *Canad. J. Math.* **14** (1962), 685–693. MR 26 #564 Zbl 0108.29201

[Fitouhi et al. 2006] A. Fitouhi, N. Bettaibi, and K. Brahim, "The Mellin transform in quantum calculus", *Constr. Approx.* **23**:3 (2006), 305–323. MR 2007b:33031 Zbl 1111.33006

[Gasper and Rahman 2004] G. Gasper and M. Rahman, *Basic hypergeometric series*, 2nd ed., Encyclopedia Math. Appl. **96**, Cambridge University Press, 2004. MR 2006d:33028 Zbl 1129.33005

[Ismail 2005] M. E. H. Ismail, *Classical and quantum orthogonal polynomials in one variable*, Encyclopedia Math. Appl. **98**, Cambridge University Press, 2005. MR 2007f:33001 Zbl 1082.42016

[Jackson 1904] F. H. Jackson, "A generalization of the function $\Gamma(n)$ and x^n", *Proc. Roy. Soc.* (*London*) **74** (1904), 64–72.

[Jackson 1905] F. H. Jackson, "The basic gamma function and the elliptic functions", *Proc. Roy. Soc.* (*London*) **76**:508 (1905), 127–144. JFM 36.0513.03

[Jackson 1910] F. H. Jackson, "On q-definite integrals", *Quart. J. Pure Appl. Math.* **41** (1910), 193–203. JFM 41.0317.04

[Kesarwani 1967] R. N. Kesarwani, "Fractional integration and certain dual integral equations", *Math. Z.* **98** (1967), 83–88. MR 35 #2087 Zbl 0166.10202

[Koelink 1994] H. T. Koelink, "The quantum group of plane motions and the Hahn–Exton q-Bessel function", *Duke Math. J.* **76**:2 (1994), 483–508. MR 96a:33023 Zbl 0814.33011

[Koelink and Swarttouw 1994] H. T. Koelink and R. F. Swarttouw, "On the zeros of the Hahn–Exton q-Bessel function and associated q-Lommel polynomials", *J. Math. Anal. Appl.* **186**:3 (1994), 690–710. MR 95j:33050 Zbl 0811.33013

[Koornwinder and Swarttouw 1992] T. H. Koornwinder and R. F. Swarttouw, "On q-analogues of the Fourier and Hankel transforms", *Trans. Amer. Math. Soc.* **333**:1 (1992), 445–461. MR 92k:33013 Zbl 0759.33007

[Love 1963] E. R. Love, "Dual integral equations", *Canad. J. Math.* **15** (1963), 631–640. MR 27 #4029 Zbl 0117.32101

[Nasim 1986] C. Nasim, "On dual integral equations with Hankel kernel and an arbitrary weight function", *Int. J. Math. Math. Sci.* **9**:2 (1986), 293–300. MR 87j:45013 Zbl 0602.45004

[Noble 1955] B. Noble, "On some dual integral equations", *Quart. J. Math.* (2) **6** (1955), 81–87. MR 17,45d Zbl 0067.33902

[Noble 1963] B. Noble, "The solution of Bessel function dual integral equations by a multiplying-factor method", *Proc. Cambridge Philos. Soc.* **59** (1963), 351–362. MR 26 #2842 Zbl 0117.32201

[Peters 1961] A. S. Peters, *Certain dual integral equations and Sonine's integrals*, IMM-NYU **285**, New York University, Institute of Mathematical Sciences, New York, 1961. MR 42 #2268

[Samko et al. 1987] S. G. Samko, A. A. Kilbas, and O. I. Marichev, *Интегралы и производные дробного порядка и некоторые их приложения*, Nauka i Tekhnika, Minsk, 1987. Translated as *Fractional integrals and derivatives: theory and applications*, Vol. II, Gordon and Breach, Yverdon, 1993. MR 96d:26012 Zbl 0617.26004

[Sneddon 1960] I. N. Sneddon, "The elementary solution of dual integral equations", *Proc. Glasgow Math. Assoc.* **4**:3 (1960), 108–110. MR 22 #9825 Zbl 0118.31402

[Sneddon 1966] I. N. Sneddon, *Mixed boundary value problems in potential theory*, Wiley, New York, 1966. MR 35 #6853 Zbl 0139.28801

[Sneddon and Lowengrub 1969] I. N. Sneddon and M. Lowengrub, *Crack problems in the classical theory of elasticity*, Wiley, New York, 1969. MR 41 #2986 Zbl 0201.26702

[Swarttouw 1992] R. F. Swarttouw, *The Hahn–Exton q-Bessel function*, thesis, Technische Universiteit Delft, 1992, Available at http://tinyurl.com/Swarttouw-thesis-1992. MR 2714507

[Titchmarsh 1986] E. C. Titchmarsh, *Introduction to the theory of Fourier integrals*, 3rd ed., Chelsea, New York, 1986. MR 89c:42002 Zbl 0017.40404

[Tranter 1951] C. J. Tranter, "On some dual integral equations", *Quart. J. Math.* (2) **2** (1951), 60–66. MR 12,713a Zbl 0044.10402

[Williams 1961] W. E. Williams, "The solution of certain dual integral equations", *Proc. Edinburgh Math. Soc.* (2) **12** (1961), 213–216. MR 32 #8075 Zbl 0119.30804

Received February 10, 2014. Revised July 4, 2014.

OLA A. ASHOUR
DEPARTMENT OF MATHEMATICS, FACULTY OF SCIENCE
CAIRO UNIVERSITY
P. O. BOX 12613
GIZA 12613
EGYPT
ola@sci.cu.edu.eg

MOURAD E. H. ISMAIL
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CENTRAL FLORIDA
4000 CENTRAL FLORIDA BOULEVARD
P. O. BOX 161364
ORLANDO, FL 32816-1364
UNITED STATES
mourad.eh.ismail@gmail.com

DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KING SAUD UNIVERSITY
P. O. BOX 2455
RIYADH 11451
SAUDI ARABIA

ZEINAB S. MANSOUR
DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE
KING SAUD UNIVERSITY
P. O. BOX 2455
RIYADH 11451
SAUDI ARABIA
zsmansour@ksu.edu.sa

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.103

ON A CONJECTURE OF ERDŐS AND CERTAIN DIRICHLET SERIES

TAPAS CHATTERJEE AND M. RAM MURTY

Let $f:\mathbb{Z}/q\mathbb{Z}\to\mathbb{Z}$ be such that $f(a)=\pm 1$ for $1\le a<q$, and $f(q)=0$. Then Erdős conjectured that $\sum_{n\ge 1} f(n)/n\ne 0$. For q even, it is easy to show that the conjecture is true. The case $q\equiv 3 \pmod 4$ was solved by Murty and Saradha. In this paper, we show that this conjecture is true for 82% of the remaining integers $q\equiv 1 \pmod 4$.

1. Introduction

In a written communication with Livingston, Erdős made the following conjecture (see [Livingston 1965]): if f is a periodic arithmetic function with period q and

$$f(n)=\begin{cases}\pm 1 & \text{if } q\nmid n,\\ 0 & \text{otherwise,}\end{cases}$$

then

$$L(1,f)=\sum_{n=1}^{\infty}\frac{f(n)}{n}\ne 0$$

where the L-function $L(s,f)$ associated with f is defined by the series

$$L(s,f):=\sum_{n=1}^{\infty}\frac{f(n)}{n^s}. \tag{1}$$

In 1973, Baker, Birch and Wirsing, using Baker's theory of linear forms in logarithms, proved the conjecture for q prime [Baker et al. 1973, Theorem 1]. In 1982, Okada [1982] established the conjecture if $2\varphi(q)+1>q$. Hence, if q is a prime power or a product of two distinct odd primes, the conjecture is true. In 2002, R. Tijdeman [2002] proved the conjecture is true for periodic completely multiplicative functions f. Saradha and Tijdeman [2003] showed that if f is

Research of the first author was supported by a postdoctoral fellowship at Queen's University, and that of the second by an NSERC Discovery grant and a Simons fellowship.
MSC2010: primary 11M06, 11M41; secondary 11M20.
Keywords: Erdős conjecture, Okada's criterion, nonvanishing of Dirichlet series.

periodic and multiplicative with $|f(p^k)| < p-1$ for every prime divisor p of q and every positive integer k, then the conjecture is true.

It is easy to see that

$$L(1, f) = \sum_{n=1}^{\infty} \frac{f(n)}{n}$$

exists if and only if $\sum_{n=1}^{q} f(n) = 0$. If q is even and f takes values ±1 with $f(q) = 0$, then $\sum_{n=1}^{q} f(n) \neq 0$. Hence the conjecture holds for even q.

In 2007, Murty and Saradha [2007] proved that if q is odd and f is an odd integer-valued odd periodic function then the conclusion of the conjecture holds. In 2010, they proved that the Erdős conjecture is true if $q \equiv 3 \pmod 4$ [Murty and Saradha 2010, Theorem 7]. Thus the conjecture is open only in cases where $q \equiv 1 \pmod 4$. However, it seems that a novel idea will be needed to deal with these cases. In this paper, we adopt a new density-theoretic approach which is orthogonal to earlier methods. Here is the main consequence of our method:

Theorem 1.1. *Let* $S(X) = |\{q \equiv 1 \pmod 4, q \leq X \mid$ *Erdős conjecture is true for* $q\}|$. *Then*

$$\liminf_{X\to\infty} \frac{S(X)}{X/4} \geq 0.82.$$

In other words, the Erdős conjecture is true for at least 82% of the integers $q \equiv 1 \pmod 4$. Our method does not extend to show that the Erdős conjecture is true for 100% of the moduli $q \equiv 1 \pmod 4$. We examine this question briefly at the end of the paper. It seems to us that more ideas are needed to resolve the conjecture fully.

These questions have a long history beginning with Baker, Birch and Wirsing [Baker et al. 1973]. Their work was generalized by Gun, Murty and Rath [Gun et al. 2012] to the setting of algebraic number fields. The paper [Chatterjee and Murty 2014] gives new proofs of some of the background results of this area. We also refer the reader to [Tijdeman 2002] for an expanded survey of the early history.

2. Notations and preliminaries

From now onwards, we denote the field of rationals by $\mathbb{Q}$, the field of algebraic numbers by $\overline{\mathbb{Q}}$, Euler's totient function by φ and Euler's constant by γ. We say a function f is Erdősian modulo q if f is a periodic function with period q and

$$f(n) = \begin{cases} \pm 1 & \text{if } q \nmid n, \\ 0 & \text{otherwise.} \end{cases}$$

Also we will write $f(X) \lesssim g(X)$ to mean

$$\limsup_{X\to\infty} \frac{f(X)}{g(X)} \le 1.$$

Similarly, we write $f(x) \gtrsim g(x)$ to mean

$$\liminf_{X\to\infty} \frac{f(X)}{g(X)} \ge 1.$$

2A. *Okada's criterion.*

Proposition 2.1. *Let the q-th cyclotomic polynomial Φ_q be irreducible over the field $\mathbb{Q}(f(1), \dots, f(q))$. Let $M(q)$ be the set of positive integers which are composed of prime factors of q. For any integer r and prime p, let $v_p(r)$ be the exponent of p dividing r.*

Then $L(1, f) = 0$ if and only if the following conditions are satisfied:

$$\sum_{m\in M(q)} \frac{f(am)}{m} = 0 \qquad \text{for every } a \text{ with } 1 \le a < q \text{ and } (a, q) = 1, \text{ and}$$

$$\sum_{\substack{r=1\\(r,q)>1}}^{q} f(r)\epsilon(r, p) = 0 \quad \text{for every prime divisor } p \text{ of } q,$$

where

$$\epsilon(r, p) = \begin{cases} v_p(r) & \text{if } v_p(r) < v_p(q),\\ v_p(q) + \dfrac{1}{p-1} & \text{otherwise.}\end{cases}$$

This proposition is a modification, due to Saradha and Tijdeman [2003], of a result of Okada [1986]. Note that Okada deduced the sufficient condition $2\varphi(q) + 1 > q$ stated in the introduction from his original version of this criterion.

2B. *Wirsing's theorem.* The following proposition is due to Wirsing [1961].

Proposition 2.2. *Let f be a nonnegative multiplicative arithmetic function, satisfying*

$$|f(p)| \le G \text{ for all primes } p,$$

$$\sum_{p\le X} p^{-1} f(p) \log p \sim \tau \log X,$$

with some constants $G > 0$, $\tau > 0$ and

$$\sum_{p} \sum_{k\ge2} p^{-k} |f(p^k)| < \infty;$$

if $0 < \tau \le 1$, *then*, *in addition*, *the condition*

$$\sum_{p} \sum_{\substack{k\ge 2 \\ p^k \le X}} |f(p^k)| = O(X/\log X)$$

is assumed to hold. Then

$$\sum_{n\le X} f(n) = (1+o(1)) \frac{X}{\log X} \frac{e^{-\gamma\tau}}{\Gamma(\tau)} \prod_{p\le X} \biggl(1 + \frac{f(p)}{p} + \frac{f(p^2)}{p^2} + \cdots \biggr).$$

2C. ***Mertens' theorem.*** We also need a classical theorem of Mertens in a later section. We record the theorem here (see, for example, [Murty 2008, page 130]):

Proposition 2.3.
$$\lim_{X\to\infty} \log X \prod_{p\le X} \biggl(1 - \frac{1}{p}\biggr) = e^{-\gamma}.$$

3. Exceptions to the conjecture of Erdős

We say that the Erdős conjecture is false modulo q, if there is an Erdősian function f for which $L(1, f) = 0$. The following proposition plays a fundamental role in our approach.

Proposition 3.1. *If the Erdős conjecture is false modulo q with q odd*, *then*

$$1 \le \sum_{\substack{d\mid q \\ d\ge 3}} \frac{1}{\varphi(d)}.$$

Proof. By the hypothesis, there is an Erdősian function $f \pmod q$ for which, we have $L(1, f) = 0$. Applying Okada's criterion, we get

$$\sum_{b\in M(q)} \frac{f(b)}{b} = 0. \tag{2}$$

Let $d = (b, q)$, so that $b = db_1$ with $(b_1, q/d) = 1$. Then (2) can be written as

$$-f(1) = \sum_{\substack{d\mid q \\ d\ge 3}} \frac{1}{d} \sum_{\substack{b_1\in M(q) \\ (b_1, q/d)=1}} \frac{f(db_1)}{b_1}.$$

Taking absolute value of both sides, we get

$$1 \le \sum_{\substack{d\mid q \\ d\ge 3}} \frac{1}{d} \sum_{b_1\in M(d)} \frac{1}{b_1}. \tag{3}$$

Notice that the inner sum can be written as

$$\sum_{b_1 \in M(d)} \frac{1}{b_1} = \prod_{p|d} \left(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots\right) = \prod_{p|d} \left(1 - \frac{1}{p}\right)^{-1} = \frac{d}{\varphi(d)}.$$

Hence from (3), we get

$$1 \leq \sum_{\substack{d|q \\ d \geq 3}} \frac{1}{\varphi(d)}. \qquad \square$$

Corollary 3.2. *If q is a prime power or a product of two distinct odd primes, then the Erdős conjecture is true modulo q.*

Proof. This is a pleasant elementary exercise. □

Hence we have recovered the two basic cases of the conjecture which were given in the introduction, of course, also as a consequence of Okada's criterion.

Let $d(n)$ be the divisor function, that is, $d(n)$ is the number of divisors of n.

Corollary 3.3. *If the smallest prime factor of q is at least $d(q)$, then the Erdős conjecture is true for q.*

Proof. Let l be the smallest prime factor of q. From the above proposition, if the Erdős conjecture is false modulo q, then we have

$$\begin{aligned} 1 &\leq \sum_{\substack{d|q \\ d \geq 3}} \frac{1}{\varphi(d)} \\ &< \frac{1}{\varphi(l)} \sum_{\substack{d|q \\ d \geq 3}} 1 = \frac{d(q) - 2}{l - 1}, \end{aligned}$$

the strict inequality in the penultimate step coming from the fact that q has at least two prime divisors. Thus, $l < d(q)$. Hence if $l \geq d(q)$, then the Erdős conjecture is true modulo q. □

Note that, Corollary 3.3 was not known previously. It implies that the conjecture is true for any squarefree number q with k prime factors, provided the smallest prime factor of q is greater than 2^k. Proposition 3.1 opens the door for a new approach to the study of Erdős's conjecture. Let us consider the following:

$$S_1(X) = \left|\{q \equiv 1 \pmod 4, q \leq X \mid \text{Erdős conjecture is false modulo } q\}\right|.$$

Then, we have

$$\begin{aligned}
S_1(X) &\le \sum_{\substack{q\le X\\ q\equiv 1 \ (\mathrm{mod}\ 4)}} \sum_{\substack{d\mid q\\ d\ge 3}} \frac{1}{\varphi(d)} \le \sum_{\substack{3\le d\le X\\ d \text{ odd}}} \frac{1}{\varphi(d)} \sum_{\substack{q\le X\\ q\equiv 1 \ (\mathrm{mod}\ 4)\\ d\mid q}} 1\\
&\le \sum_{\substack{3\le d\le X\\ d \text{ odd}}} \frac{1}{\varphi(d)}\Big(\frac{X}{4d}+O(1)\Big) \le \sum_{\substack{3\le d\le X\\ d \text{ odd}}} \frac{1}{\varphi(d)}\frac{X}{4d} + O\Big(\sum_{3\le d\le X}\frac{1}{\varphi(d)}\Big)\\
&\le \sum_{\substack{3\le d\le X\\ d \text{ odd}}} \frac{1}{\varphi(d)}\frac{X}{4d} + O(\log X),
\end{aligned}$$

where we have used the well-known fact that (see, for example, [Murty 2008, page 67])

$$\sum_{d\le X}\frac{1}{\varphi(d)} = O(\log X).$$

Hence, we get

$$\begin{aligned}
S_1(X) &\lesssim \frac{X}{4}\sum_{\substack{3\le d\\ d \text{ odd}}} \frac{1}{d\varphi(d)}\\
&\lesssim \frac{X}{4}\Big(\prod_{p \text{ odd}}\Big(1+\frac{1}{p\varphi(p)}+\frac{1}{p^2\varphi(p^2)}+\cdots\Big)-1\Big)\\
&\lesssim \frac{X}{4}\Big(\prod_{p \text{ odd}}\Big(1+\frac{1}{p(p-1)}+\frac{1}{p^3(p-1)}+\cdots\Big)-1\Big)\\
&\lesssim \frac{X}{4}\Big(\prod_{p \text{ odd}}\Big(1+\frac{1}{p(p-1)}\Big(1+\frac{1}{p^2}+\frac{1}{p^4}+\cdots\Big)\Big)-1\Big)\\
&\lesssim \frac{X}{4}\Big(\prod_{p \text{ odd}}\Big(1+\frac{p}{(p-1)(p^2-1)}\Big)-1\Big).
\end{aligned}$$

The product is easily computed numerically and we have $S_1(X) \lesssim 0.33(X/4)$. The following is an immediate corollary.

Corollary 3.4. $|\{q \equiv 1 \pmod 4, q \le X \mid$ *Erdős conjecture is true for* $q\}| \gtrsim 0.67\frac{X}{4}$.

3A. ***Refinement using the second moment.*** By considering higher moments, we can improve the lower bound in the above corollary. We begin with the second moment. We include these estimates since they are of independent interest and self contained.

Proposition 3.5. $|\{q \equiv 1 \pmod 4, q \le X \mid$ *Erdős conjecture is true for* $q\}| \gtrsim 0.78\frac{X}{4}$.

Proof. Let us first consider the following inequality:

$$\begin{aligned}
S_1(X) &\le \sum_{\substack{q\le X\\ q\equiv 1 \ (\mathrm{mod}\, 4)}} \Bigg(\sum_{\substack{d|q\\ d\ge 3}} \frac{1}{\varphi(d)}\Bigg)^2 \\
&\le \sum_{\substack{q\le X\\ q\equiv 1 \ (\mathrm{mod}\, 4)}} \sum_{\substack{d_1|q, d_2|q\\ 3\le d_1, d_2<q}} \frac{1}{\varphi(d_1)\varphi(d_2)} \\
&\le \sum_{\substack{3\le d_1, d_2\le X\\ d_1, d_2 \text{ odd}}} \frac{1}{\varphi(d_1)\varphi(d_2)} \sum_{\substack{q\le X\\ q\equiv 1 \ (\mathrm{mod}\, 4)\\ d_1|q, d_2|q}} 1 \\
&\le \sum_{\substack{3\le d_1, d_2\le X\\ d_1, d_2 \text{ odd}}} \frac{1}{\varphi(d_1)\varphi(d_2)} \sum_{\substack{q\le X\\ q\equiv 1 \ (\mathrm{mod}\, 4)\\ [d_1, d_2]|q}} 1 \\
&\le \sum_{\substack{3\le d_1, d_2\le X\\ d_1, d_2 \text{ odd}}} \frac{1}{\varphi(d_1)\varphi(d_2)} \Bigg(\frac{X}{4[d_1, d_2]} + O(1)\Bigg).
\end{aligned}$$

Hence, we have

$$S_1(X) \le \frac{X}{4} \sum_{\substack{3\le d_1, d_2\le X\\ d_1, d_2 \text{ odd}}} \frac{1}{\varphi(d_1)\varphi(d_2)[d_1, d_2]} + O(\log^2 X).$$

By a simple numerical calculation, we deduce that

$$S_1(X) \lesssim 0.22\frac{X}{4}.$$

Hence the conjecture holds for at least 78% of the positive integers congruent to 1 (mod 4). □

Similarly one can compute higher fractional moments to get an optimal result. For any $r > 1$, we have

$$S_1(X) \le \sum_{\substack{q\le X\\ q\equiv 1 \ (\mathrm{mod}\, 4)}} \Bigg(\sum_{\substack{d|q\\ d\ge 3}} \frac{1}{\varphi(d)}\Bigg)^r.$$

We study this as a function of r. Using Maple we computed that the minimal value occurs at $r \sim 3.85$[1] and we get

$$S_1(X) \lesssim 0.18\frac{X}{4}.$$

[1] Code available at www.mast.queensu.ca/∼murty/maplecode.pdf.

Thus, we get $|\{q \equiv 1 \pmod 4, q \leq X \mid \text{Erdős conjecture is true for } q\}| \gtrsim 0.82\frac{X}{4}$, that is,

$$\liminf_{X\to\infty} \frac{S(X)}{X/4} \geq 0.82.$$

Hence, we have shown Theorem 1.1: the conjecture holds for at least 82% of the numbers congruent to 1 (mod 4).

3B. ***An alternative approach.*** In this subsection, we discuss an alternative approach to this problem. It leads to a slightly weaker result. However this method is of independent interest, so we record it here. We begin with a further refinement of Proposition 3.1 by considering fractional moments there. From Proposition 3.1, if the Erdős conjecture is false for odd q, then

$$1 \leq \sum_{\substack{d\mid q\\ d\geq 3}} \frac{1}{\varphi(d)}.$$

Adding 1 to both sides of the above inequality, we get

$$2 \leq \sum_{d\mid q} \frac{1}{\varphi(d)},$$

which can be rewritten as

$$1 \leq \frac{1}{2}\sum_{d\mid q} \frac{1}{\varphi(d)}.$$

Hence for any $\alpha > 0$, Proposition 3.1 can be rewritten as follows.

Proposition 3.6. *If Erdős conjecture is false for odd q, then*

$$1 \leq \frac{1}{2^\alpha}\biggl(\sum_{d\mid q} \frac{1}{\varphi(d)}\biggr)^\alpha.$$

As before, $S_1(X) = |\{q \equiv 1 \pmod 4, q \leq X \mid \text{Erdős conjecture is false for } q\}|$. Then from the above proposition, we get

$$S_1(X) \leq \frac{1}{2^\alpha} \sum_{\substack{q\leq X\\ q\equiv 1 \pmod 4}} \biggl(\sum_{d\mid q} \frac{1}{\varphi(d)}\biggr)^\alpha.$$

Let $f_\alpha(q) = \Bigl(\sum_{d\mid q} 1/\varphi(d)\Bigr)^\alpha$ and χ be the nontrivial Dirichlet character mod 4. Then the above inequality becomes

$$S_1(X) \leq \frac{1}{2^{\alpha+1}}\biggl(\sum_{\substack{q\leq X\\ q \text{ odd}}} f_\alpha(q) + \sum_{\substack{q\leq X\\ q \text{ odd}}} \chi(q) f_\alpha(q)\biggr). \tag{4}$$

Again, note that $f_\alpha(q)$ is a multiplicative arithmetic function. One can check that it also satisfies all the other hypotheses of Wirsing's theorem (Proposition 2.2) with $G = 2^\alpha$ and $\tau = 1$. So in light of Wirsing's theorem, we get

$$\sum_{\substack{q \le X \\ q \text{ odd}}} f_\alpha(q) \sim X \frac{e^{-\gamma}}{\log X} \prod_{\substack{p \le X \\ p \ne 2}} \left(1 + \frac{f_\alpha(p)}{p} + \frac{f_\alpha(p^2)}{p^2} + \cdots \right)$$

and

$$\sum_{\substack{q \le X \\ q \text{ odd}}} \chi(q) f_\alpha(q) \sim X \frac{e^{-\gamma}}{\log X} \prod_{\substack{p \le X \\ p \ne 2}} \left(1 + \frac{\chi(p) f_\alpha(p)}{p} + \frac{\chi(p^2) f_\alpha(p^2)}{p^2} + \cdots \right).$$

Again, from Mertens theorem we know that

$$\prod_{p \le X} (1 - 1/p) \sim \frac{e^{-\gamma}}{\log X}.$$

Hence we have

$$\begin{aligned} \sum_{\substack{q \le X \\ q \text{ odd}}} f_\alpha(q) &\sim \frac{X}{2} \prod_{\substack{p \le X \\ p \ne 2}} (1 - 1/p) \left(1 + \frac{f_\alpha(p)}{p} + \frac{f_\alpha(p^2)}{p^2} + \cdots \right) \\ &\sim \frac{X}{2} p_1 \text{ (say)} \end{aligned}$$

and

$$\begin{aligned} \sum_{\substack{q \le X \\ q \text{ odd}}} \chi(q) f_\alpha(q) &\sim \frac{X}{2} \prod_{\substack{p \le X \\ p \ne 2}} (1 - 1/p) \left(1 + \frac{\chi(p) f_\alpha(p)}{p} + \frac{\chi(p^2) f_\alpha(p^2)}{p^2} + \cdots \right) \\ &\sim \frac{X}{2} p_2 \text{ (say)}. \end{aligned}$$

Now using the above two inequalities, (4) becomes

$$S_1(X) \lesssim \frac{X}{2^{\alpha+2}} (p_1 + p_2).$$

Finally, using Maple[2] we find that the quantity on the right hand side is minimized at $\alpha \sim 8.11$ and we get

$$S_1(X) \lesssim 0.20 \frac{X}{4}.$$

[2] Code available at www.mast.queensu.ca/~murty/maplecode.pdf.

Hence, we get

$$\liminf_{X\to\infty} \frac{S(X)}{X/4} \geq 0.80.$$

Remarks. One cannot hope to obtain 100% by these methods. In fact, one can show that there is a positive density (albeit small) of q for which the inequality of Proposition 3.1 holds. Indeed, since

$$\sum_{d|q} \frac{1}{\varphi(d)} \geq \prod_{p|q}\left(1+\frac{1}{p-1}\right)$$

we can make the product (and hence the sum) arbitrarily large by ensuring that q is divisible by all the primes in an initial segment. We can even ensure that these primes are congruent to $1 \pmod 4$. We then take numbers which are divisible by this q and congruent to $1 \pmod 4$ and deduce that for all these numbers, the inequality in the proposition holds. Since the product on the right diverges slowly to infinity as we go through such numbers q, we obtain in this way a small density of numbers for which the inequality holds.

Acknowledgements

We thank Michael Roth for his help on using the Maple language as well as Sanoli Gun, Purusottam Rath and Ekata Saha for their comments on an earlier version of this paper. We also thank the referee for helpful comments that improved the quality of the paper.

References

[Baker et al. 1973] A. Baker, B. J. Birch, and E. A. Wirsing, "On a problem of Chowla", *J. Number Theory* **5** (1973), 224–236. MR 49 #4959 Zbl 0267.10065

[Chatterjee and Murty 2014] T. Chatterjee and M. R. Murty, "Non-vanishing of Dirichlet series with periodic coefficients", *J. Number Theory* **145** (2014), 1–21. MR 3253290

[Gun et al. 2012] S. Gun, M. R. Murty, and P. Rath, "Linear independence of Hurwitz zeta values and a theorem of Baker-Birch-Wirsing over number fields", *Acta Arith.* **155**:3 (2012), 297–309. MR 2983455

[Livingston 1965] A. E. Livingston, "The series $\sum_1^\infty f(n)/n$ for periodic f", *Canad. Math. Bull.* **8** (1965), 413–432. MR 32 #4104 Zbl 0129.02801

[Murty 2008] M. R. Murty, *Problems in analytic number theory*, Second ed., Graduate Texts in Mathematics **206**, Springer, New York, 2008. Readings in Mathematics. MR 2008j:11001 Zbl 1190.11001

[Murty and Saradha 2007] M. R. Murty and N. Saradha, "Transcendental values of the digamma function", *J. Number Theory* **125**:2 (2007), 298–318. MR 2008g:11123 Zbl 1222.11097

[Murty and Saradha 2010] M. R. Murty and N. Saradha, "Euler-Lehmer constants and a conjecture of Erdös", *J. Number Theory* **130**:12 (2010), 2671–2682. MR 2011h:11078

[Okada 1982] T. Okada, “On a certain infinite series for a periodic arithmetical function”, *Acta Arith.* **40**:2 (1982), 143–153. MR 83h:10099 Zbl 0402.10035

[Okada 1986] T. Okada, “Dirichlet series with periodic algebraic coefficients”, *J. London Math. Soc.* (2) **33**:1 (1986), 13–21. MR 87i:11087 Zbl 0589.10034

[Saradha and Tijdeman 2003] N. Saradha and R. Tijdeman, “On the transcendence of infinite sums of values of rational functions”, *J. London Math. Soc.* (2) **67**:3 (2003), 580–592. MR 2004b:11103 Zbl 1045.11051

[Tijdeman 2002] R. Tijdeman, “Some applications of Diophantine approximation”, pp. 261–284 in *Number theory for the millennium, III (Urbana, IL, 2000)*, A K Peters, Natick, MA, 2002. MR 2003j:11076 Zbl 1045.11022

[Wirsing 1961] E. Wirsing, “Das asymptotische Verhalten von Summen über multiplikative Funktionen”, *Math. Ann.* **143** (1961), 75–102. MR 24 #A1241 Zbl 0104.04201

Received February 20, 2014. Revised November 5, 2014.

TAPAS CHATTERJEE
DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY ROPAR
NANGAL ROAD
PUNJAB 140001
INDIA

tapasc@iitrpr.ac.in

M. RAM MURTY
DEPARTMENT OF MATHEMATICS & STATISTICS
QUEEN’S UNIVERSITY
KINGSTON ON K7L3N6
CANADA

murty@mast.queensu.ca

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.115

NORMAL FORMS FOR CR SINGULAR CODIMENSION-TWO LEVI-FLAT SUBMANIFOLDS

XIANGHONG GONG AND JIŘÍ LEBL

Real-analytic Levi-flat codimension-two CR singular submanifolds are a natural generalization to $\mathbb{C}^m$, $m > 2$, of Bishop surfaces in $\mathbb{C}^2$. Such submanifolds, for example, arise as zero sets of mixed-holomorphic equations with one variable antiholomorphic. We classify the codimension-two Levi-flat CR singular quadrics, and we notice that new types of submanifolds arise in dimension three or higher. In fact, the nondegenerate submanifolds, i.e., higher order perturbations of $z_m = \bar{z}_1 z_2 + \bar{z}_1^2$, have no analogue in dimension two. We prove that the Levi foliation extends through the singularity in the real-analytic nondegenerate case. Furthermore, we prove that the quadric is a (convergent) normal form for a natural large class of such submanifolds, and we compute its automorphism group. In general, we find a formal normal form in $\mathbb{C}^3$ in the nondegenerate case that shows infinitely many formal invariants.

Lebl was in part supported by NSF grant DMS 0900885 and DMS 1362337.
MSC2010: primary 32V40; secondary 53C12, 32S05.
Keywords: Normal form, Levi-flat, CR singular, codimension two, Bishop surface, mixed-holomorphic equations.

1. Introduction

Let $M \subset \mathbb{C}^{n+1}$ be a real submanifold. A fundamental question in CR geometry is to classify M at a point up to local biholomorphic transformations. One approach is to find a normal form for M.

A real-analytic hypersurface $M \subset \mathbb{C}^{n+1}$ is Levi-flat if the Levi form vanishes identically. Roughly speaking, a Levi-flat submanifold is a family of complex submanifolds. Intuitively, a Levi-flat submanifold is as close to a complex submanifold as possible. In the real-analytic smooth hypersurface case, it is well known that M can locally be transformed into the real hyperplane given by

$$\operatorname{Im} z_1 = 0. \tag{1}$$

We therefore focus on higher codimension case, in particular on codimension two. A codimension-two submanifold is again given by a single equation, but in this case a complex valued equation. A new phenomenon that appears in codimension two is that M may no longer be a CR submanifold. Let $T_p^c M \subset T_p M$ be the largest subspace with $J T_p^c M = T_p^c M$, where J is the complex structure on $\mathbb{C}^{n+1}$. A submanifold is CR if $\dim T_p^c M$ is constant.

Real submanifolds of dimension $n+1$ in $\mathbb{C}^{n+1}$ with a nondegenerate complex tangent point has been studied extensively after the fundamental work of E. Bishop [1965]. In $\mathbb{C}^2$, he studied the submanifolds

$$w = z\bar{z} + \gamma(z^2 + \bar{z}^2) + O(3), \tag{2}$$

where $\gamma \in [0, \infty]$ is called the Bishop invariant, with $\gamma = \infty$ interpreted as $w = z^2 + \bar{z}^2 + O(3)$. One of Bishop's motivations was to study the hull of holomorphy of the real submanifolds by attaching analytic discs. His work on the family of attached analytic discs has been refined by Kenig and Webster [1982; 1984], Huang and Krantz [1995], and Huang [1998]. The normal form theory for real submanifolds for Bishop surfaces or submanifolds was established by Moser and Webster [1983]; see also Moser [1985], Gong [1994a; 1994b; 2004], Huang and Yin [2009a], and Coffman [2010]. We mention that the Moser–Webster normal form does not deal with the case of vanishing Bishop invariant.

The formal normal form and its application to holomorphic classification for surfaces with vanishing Bishop invariant were achieved by Huang and Yin [2009a] by a completely different method. Real submanifolds with complex tangents have been studied in other situations. See, for example, [Lebl et al. 2014], where CR singular submanifolds that are images of CR manifolds were studied. Normal forms for the quadratic part of general codimension-two CR singular submanifolds in $\mathbb{C}^3$ was completely solved by Coffman [2009]. Huang and Yin [2009b] studied the normal form for codimension-two CR singular submanifolds of the form $w =$

$|z|^2 + O(3)$. Dolbeault, Tomassini and Zaitsev [2005; 2011] and Huang and Yin [2012] studied CR singular submanifolds of codimension-two that are boundaries of Levi-flat hypersurfaces. Burcea [2013] constructed the formal normal form for codimension-two CR singular submanifolds approximating a sphere. Coffman [2006] found an algebraic normal form for nondegenerate CR singular manifolds in high codimension and one-dimensional complex tangent.

To motivate our work, we observe that in Bishop's work, the real submanifolds are Levi-flat away from their CR singular sets. Our purpose is to understand such submanifolds in higher dimensional case with codimension being exactly two. Notice that the latter is the smallest codimension for CR singularity to be present in (smooth) submanifolds. Regarding CR singular Levi-flat real codimension-two submanifolds on $\mathbb{C}^{n+1}$ as a natural generalization of Bishop surfaces to $\mathbb{C}^{n+1}$, we wish to find their normal forms. For singular Levi-flat hypersurfaces and related work on foliations with singularity, see [Bedford 1977; Burns and Gong 1999; Brunella 2007; Cerveau and Lins Neto 2011; Lebl 2013; Fernández-Pérez 2013].

Our techniques revolve around the study of the Levi map (the generalization of the Levi form to higher codimension submanifolds) of codimension-two submanifolds. Extending the CR structure through the singular point via Nash blowup and then extending the Levi map to this blowup has been studied previously by Garrity [2000].

A CR submanifold is *Levi-flat* if the Levi map vanishes identically. Locally, all CR real-analytic Levi-flat submanifolds of real codimension two can be, after holomorphic change of coordinates, written as

$$\operatorname{Im} z_1 = 0, \quad \operatorname{Im} z_2 = 0. \tag{3}$$

If a submanifold M is CR singular, denote by M_{CR} the set of points where M is CR. We say M is Levi-flat if M_{CR} is Levi-flat in the usual sense. A Levi-flat CR singular submanifold has no local biholomorphic invariants at the CR points, just as in the case of Bishop surfaces.

A real, real-analytic codimension-two submanifold that is CR singular at the origin can be written in coordinates $(z, w) \in \mathbb{C}^n \times \mathbb{C} = \mathbb{C}^{n+1}$ as

$$w = \rho(z, \bar{z}) \tag{4}$$

for ρ that is $O(2)$. We will be concerned with submanifolds where the quadratic part in ρ is nonzero in any holomorphic coordinates. We say that such submanifolds have a *nondegenerate complex tangent*. For example, the Bishop surfaces in $\mathbb{C}^2$ are precisely the CR singular submanifolds with nondegenerate complex tangent.

First, let us classify the quadratic parts of CR singular Levi-flats, and in the process completely classify the CR singular Levi-flat quadrics, that is, those where ρ is a quadratic.

Theorem 1.1. *Suppose that $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is a germ of a real-analytic real codimension-two submanifold, CR singular at the origin, written in coordinates $(z,w) \in \mathbb{C}^n \times \mathbb{C}$ as*

$$w = A(z,\bar{z}) + B(\bar{z},\bar{z}) + O(3), \tag{5}$$

for quadratic A and B, where $A+B \not\equiv 0$ (nondegenerate complex tangent). Suppose that M is Levi-flat (that is, M_{CR} is Levi-flat).

(i) *If M is a quadric, then M is locally biholomorphically equivalent to one and exactly one of the following:*

$$\begin{aligned}
&(\mathrm{A}.1) && w = \bar{z}_1^2,\\
&(\mathrm{A}.2) && w = \bar{z}_1^2 + \bar{z}_2^2,\\
& && \vdots\\
&(\mathrm{A}.n) && w = \bar{z}_1^2 + \bar{z}_2^2 + \cdots + \bar{z}_n^2,\\
&(\mathrm{B}.\gamma) && w = |z_1|^2 + \gamma \bar{z}_1^2, \quad \gamma \geq 0,\\
&(\mathrm{C}.0) && w = \bar{z}_1 z_2,\\
&(\mathrm{C}.1) && w = \bar{z}_1 z_2 + \bar{z}_1^2.
\end{aligned}$$

(ii) *If M is real-analytic, then the quadric*

$$w = A(z,\bar{z}) + B(z,\bar{z}) \tag{6}$$

is Levi-flat, and can be put via a biholomorphic transformation into exactly one of the forms above.

By part (ii), the quadratic part in (5) is an invariant of M at a point. We say the *type* of M at the origin is A.x, B.γ, or C.x depending on the type of the quadratic form. Following Bishop, we call types B.γ and A.1 Bishop-like, and we could think of $\gamma = \infty$ as A.1.

By type being *stable* we mean that the type does not change at all complex tangents in a neighborhood of the origin under any small (or higher order) perturbations that stay within the class of Levi-flat CR singular submanifolds. As a consequence of Theorem 1.1 and because the rank is lower semicontinuous, we get that the only types that are stable are A.n and C.1, although A.n are degenerate because the form $A(z,\bar{z})$ is identically zero. See also Proposition 15.1.

The quadrics A.k for $k \geq 2$ do not possess a nonsingular foliation extending the Levi foliation of M_{CR} through the origin. In fact, there is a singular complex subvariety of dimension one through the origin contained in M. See Section 6.

In the sequel, when we wish to refer to the quadric of certain type we will use the notation $M_{C.1}$ to denote the quadric of type C.1.

The quadratic form $A(z, \bar{z})$ carries the "Levi map" of the submanifold. Type C.1 is the unique quadric that is stable and has nonzero A. Having nonzero A is also stable in a neighborhood of the origin under any small (or higher order) perturbations. Therefore, we say a type is *nondegenerate* if it is C.1 and we focus mostly on such submanifolds. First, we show that submanifolds of type C.x possess a nonsingular real-analytic foliation that extends the Levi foliation, due to the form $A(z, \bar{z})$:

Theorem 1.2. *Suppose that $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is a real-analytic Levi-flat CR singular submanifold of type C.1 or C.0, that is, M is given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + O(3) \quad \textit{or} \quad w = \bar{z}_1 z_2 + O(3). \tag{7}$$

Then there exists a nonsingular real-analytic foliation defined on M that extends the Levi foliation on M_{CR}; and consequently, there exists a CR real-analytic mapping $F\colon U \subset \mathbb{R}^2 \times \mathbb{C}^{n-1} \to \mathbb{C}^{n+1}$ such that F is a diffeomorphism onto $F(U) = M \cap U'$, for some neighborhood U' of zero.

Here the CR structure on $\mathbb{R}^2 \times \mathbb{C}^{n-1}$ is induced from $\mathbb{C}^2 \times \mathbb{C}^{n-1}$. As a corollary of this theorem we obtain in Section 8 using the results of [Lebl et al. 2014] that the CR singular set of any type C.1 submanifold is a Levi-flat submanifold of dimension $2n-2$ and CR dimension $n-2$.

The Levi foliation on a type C.x submanifold cannot extend to a whole neighborhood of M as a nonsingular holomorphic foliation. If it did, we could flatten the foliation and M would be a Cartesian product, in particular Bishop-like. Thus, the study of normal form theory for the special case when the foliation extends to a neighborhood is reduced to the case of Bishop surfaces, which have been studied extensively.

A codimension-two submanifold in $\mathbb{C}^m$ can arise from

$$f(\bar{z}', z'') = 0 \tag{8}$$

for a suitable holomorphic function f in m variables. The zero set admits two holomorphic foliations. We are interested in the case where one of foliations has leaves of maximum dimension $m-2$, while the other has leaves of minimum dimension zero. Therefore, we will assume that $z' = z_1$ and $z'' = (z_2, \ldots, z_m)$. Functions holomorphic in some variables and anti-holomorphic in other variables, such as (8), are often called *mixed-holomorphic* or mixed-analytic, and come up often in complex geometry, the simplest example being the standard inner product. An interesting feature of the mixed-holomorphic setting is that the equation can be complexified into $\mathbb{C}^m$, so the sets share some of the properties of complex varieties. However, they have a different automorphism group if we wish to classify them under biholomorphic transformations. Such mixed-analytic sets are automatically real codimension two, are Levi-flat or complex, and may have CR singularities. We study their normal form in Section 9. See also Theorem 1.3 below.

When a type C.1 CR singular submanifold has a defining equation that does not depend on $\bar{z}_2, \dots, \bar{z}_n$ we prove that it is automatically Levi-flat, and it is equivalent to $M_{C.1}$.

Theorem 1.3. *Let $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, be a real-analytic submanifold given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(z_1, \bar{z}_1, z_2, z_3, \dots, z_n), \tag{9}$$

where r is $O(3)$. Then M is Levi-flat and at the origin and M is locally biholomorphically equivalent to the quadric $M_{C.1}$ submanifold

$$w = \bar{z}_1 z_2 + \bar{z}_1^2. \tag{10}$$

The theorem is also true formally; given a formal submanifold of the form (9), it is formally equivalent to $M_{C.1}$.

A key idea in the proof of the convergence of the normalizing transformation is that the form $B(\bar{z}, \bar{z}) = \bar{z}_1^2$ induces a natural mixed-holomorphic involution on quadric $M_{C.1}$. This involution also plays a key role in computing the automorphism group of the quadric in Theorem 12.4.

Finally, we also compute the automorphism group for the quadric $M_{C.1}$; see Theorem 12.4. In particular we show that the automorphism group is infinite-dimensional.

Not every type C.1 Levi-flat submanifold is biholomorphically equivalent to the C.1 quadric. We will find a formal normal form for type C.1 Levi-flat submanifolds in $\mathbb{C}^3$ that shows infinitely many formal invariants. Let us give a simplified statement. For details, see Theorem 14.3.

Theorem 1.4. *Let M be a real-analytic Levi-flat type C.1 submanifold in $\mathbb{C}^3$. There exists a formal biholomorphic map transforming M into the image of*

$$\hat{\varphi}(z, \bar{z}, \xi) = (z + A(z, \xi, w) w \eta, \xi, w) \tag{11}$$

with $\eta = \bar{z} + \frac{1}{2}\xi$ and $w = \bar{z}\xi + \bar{z}^2$. Here $A = 0$, or A satisfies certain normalizing conditions.

When $A \neq 0$, the formal automorphism group preserving the normal form is finite or one-dimensional.

We do not know if the formal normal form above can be achieved by convergent transformations, even if $A = 0$.

2. Invariants of codimension-two CR singular submanifolds

Before we impose the Levi-flat condition, let us find some invariants of codimension-two CR singular submanifolds in $\mathbb{C}^{n+1}$ with CR singularity at zero. Such a submanifold, locally near the origin, can be put into the form

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}) + O(3), \tag{12}$$

where $(z, w) \in \mathbb{C}^n \times \mathbb{C}$ and A and B are quadratic forms. We think of A and B as matrices and z a column vector and write the forms as z^*Az and $z^*B\bar{z}$ respectively. The matrix B is not unique. Hence we make B symmetric to make the choice of the matrix B canonical. The following proposition is not difficult and well known. Since the details are important and will be used later, let us prove:

Proposition 2.1. *A biholomorphic transformation of* (12) *taking the origin to itself and preserving the form of* (12) *takes the matrices* (A, B) *to*

$$(\lambda T^*AT, \lambda T^*B\overline{T}), \tag{13}$$

for $T \in \mathrm{GL}_n(\mathbb{C})$ *and* $\lambda \in \mathbb{C}^*$. *If* $(F_1, \dots, F_n, G) = (F, G)$ *is the transformation, then the linear part of* G *is* $\lambda^{-1}w$ *and the linear part of* F *restricted to* z *is* Tz.

Let us emphasize that A is an arbitrary complex matrix and B is symmetric, but not necessarily Hermitian.

Proof. Let $(F_1, \dots, F_n, G) = (F, G)$ be a change of coordinates taking

$$w = \widetilde{A}(z, \bar{z}) + \widetilde{B}(\bar{z}, \bar{z}) + O(3) = \rho(z, \bar{z}) \tag{14}$$

to

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}) + O(3). \tag{15}$$

Then

$$\begin{aligned} &G(z, \rho(z, \bar{z})) \\ &\quad = A(F(z, \rho(z, \bar{z})), \bar{F}(\bar{z}, \bar{\rho}(\bar{z}, z))) + B(\bar{F}(\bar{z}, \bar{\rho}(\bar{z}, z)), \bar{F}(\bar{z}, \bar{\rho}(\bar{z}, z))) + O(3) \end{aligned} \tag{16}$$

is true for all z. The right hand side has no linear terms, so the linear terms in G do not depend on z. That is, $G = \lambda^{-1}w + O(2)$, where λ is a nonzero scalar and the negative power is for convenience.

Let $T = [T_1, T_2]$ denote the matrix representing the linear terms of F. Here T_1 is an $n \times n$ matrix and T_2 is $n \times 1$. Since the linear terms in G do not depend on any z_j, T_1 is nonsingular. Then the quadratic terms in (16) are

$$\lambda^{-1}(\widetilde{A}(z, \bar{z}) + \widetilde{B}(\bar{z}, \bar{z})) = z^*T_1^*AT_1z + z^*T_1^*B\overline{T}_1\bar{z}. \tag{17}$$

In other words as matrices,

$$\widetilde{A} = \lambda T_1^*AT_1 \quad \text{and} \quad \widetilde{B} = \lambda T_1^*B\overline{T}_1. \tag{18}$$

□

We will need to at times reduce to the three-dimensional case, and so we need:

Lemma 2.2. *Let* $M \subset \mathbb{C}^{n+1}$, $n \geq 3$, *be a real-analytic Levi-flat CR singular submanifold of the form*

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}) + O(3), \tag{19}$$

where A and B are quadratic. Let L be a nonsingular $(n-2)\times n$ matrix. If $A+B$ is not zero on the set $\{Lz=0\}$, then the submanifold

$$M_L = M \cap \{Lz = 0\} \tag{20}$$

is a Levi-flat CR singular submanifold.

Proof. Clearly, if M_L is not contained in the CR singularity of M, then M_L is a Levi-flat CR singular submanifold. $M_{L'}$ is not contained in the CR singularity of M for a dense open subset of $(n-2)\times n$ matrices L'. If M_L is a subset of the CR singularity of M, pick a CR point p of M_L then pick a sequence L_n approaching L such that M_{L_n} are not contained in the CR singularity of M. As $A+B$ is not zero on the set $\{Lz=0\}$, then M_L is not a complex submanifold, and therefore a CR singular submanifold. As the Levi form of M_{L_n} vanishes at all CR points of M_{L_n}, the Levi form of M_L vanishes at p, so M_L is Levi-flat. □

3. Levi-flat quadrics

Let us first focus on Levi-flat quadrics. We will prove later that the quadratic part of a Levi-flat submanifold is Levi-flat. Let M be defined in $(z,w)\in\mathbb{C}^n\times\mathbb{C}$ by

$$w = A(z,\bar z) + B(\bar z,\bar z). \tag{21}$$

Being Levi-flat has several equivalent formulations. The main idea is that the $T^{(1,0)}M \times T^{(0,1)}M$ vector fields are completely integrable at CR points and we obtain a foliation of M at CR points by complex submanifolds of complex dimension $n-1$. An equivalent notion is that the Levi map is identically zero; see [Baouendi et al. 1999]. The Levi map for a CR submanifold defined by two real equations $\rho_1=\rho_2=0$ (for ρ_1 and ρ_2 with linearly independent differentials) is the pair of Hermitian forms

$$i\partial\bar\partial\rho_1 \quad \text{and} \quad i\partial\bar\partial\rho_2, \tag{22}$$

applied to $T^{(1,0)}M$ vectors. The full quadratic forms $i\partial\bar\partial\rho_1$ and $i\partial\bar\partial\rho_2$ of course depend on the defining equations themselves and are therefore extrinsic information. It is important to note that for the Levi map we restrict it to $T^{(1,0)}M$ vectors. We can define these two forms $i\partial\bar\partial\rho_1$ and $i\partial\bar\partial\rho_2$ even at a CR singular point $p\in M$.

These forms are the complex Hessian matrices of the defining equations. For our quadric M, they are the real and imaginary parts of the $(n+1)\times(n+1)$ complex matrix

$$\widetilde{A} = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix}, \tag{23}$$

where the variables are ordered as $(z_1,\dots,z_n,w)$.

For M to be Levi-flat, the quadratic form defined by $\widetilde{A}$ has to be zero when restricted to the $(n-1)$-dimensional space spanned by $T_p^{(1,0)}M$ for every $p \in M_{\mathrm{CR}}$. In other words, for every $p \in M_{\mathrm{CR}}$,

$$v^*\widetilde{A}v = 0 \quad \text{for all } v \in T_p^{(1,0)}M. \tag{24}$$

The space $T_p^{(1,0)}M$ is of dimension $n-1$, and furthermore, the vector $\partial/\partial w$ is not in $T_p^{(1,0)}M$. Therefore, $z^*Az = 0$ for $z \in \mathbb{C}^n$ in a subspace of dimension $n-1$.

Before we proceed, let us note the following general fact about CR singular Levi-flat submanifolds:

Lemma 3.1. *Suppose that $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is a Levi-flat connected real-analytic real codimension-two submanifold, CR singular at the origin. Then there exists a germ of a complex analytic variety of complex dimension $n-1$ through the origin, contained in M.*

Proof. Through each point of M_{CR} there exists a germ of a complex variety of complex dimension $n-1$ contained in M. The set of CR points is dense in M. Take a sequence p_k of CR points converging to the origin and take complex varieties of dimension $n-1$, $W_k \subset M$ with $p_k \in W_k$. A theorem of Fornæss (see [Kohn 1979, Theorem 6.23] for a proof using the methods of Diederich and Fornæss [1978]) implies that there exists a variety through $W \subset M$ with $0 \in W$ and of complex dimension at least $n-1$. □

Let us first concentrate on $n = 2$, in which case $T^{(1,0)}M$ is one-dimensional at CR points. Write

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}. \tag{25}$$

Note that B is symmetric. A short computation shows that the vector field can be written as

$$\alpha \frac{\partial}{\partial w} + \beta_1 \frac{\partial}{\partial z_1} + \beta_2 \frac{\partial}{\partial z_2} = \alpha \frac{\partial}{\partial w} + \beta \frac{\partial}{\partial z}, \tag{26}$$

where

$$\begin{aligned} \beta_1 &= \bar{a}_{21}\bar{z}_1 + \bar{a}_{22}\bar{z}_2 + 2\bar{b}_{12}z_1 + 2\bar{b}_{22}z_2, \\ \beta_2 &= -\bar{a}_{11}\bar{z}_1 - \bar{a}_{12}\bar{z}_2 - 2\bar{b}_{11}z_1 - 2\bar{b}_{12}z_2, \\ \alpha &= a_{11}\bar{z}_1\beta_1 + a_{21}\bar{z}_2\beta_1 + a_{12}\bar{z}_1\beta_2 + a_{22}\bar{z}_2\beta_2. \end{aligned} \tag{27}$$

M_{CR} is dense in M, since the CR singular set is defined by $\beta_1 = \beta_2 = 0$. Thus for M to be Levi-flat we need to check that the following product is identically zero:

$$\begin{bmatrix} \beta^* & \bar{\alpha} \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \beta^* A \beta. \tag{28}$$

If A is the zero matrix, then M is automatically Levi-flat. We diagonalize B via T into a diagonal matrix with ones and zeros on the diagonal. We obtain (recall $n = 2$) the submanifolds

$$w = \bar{z}_1^2 \quad \text{or} \quad w = \bar{z}_1^2 + \bar{z}_2^2. \tag{29}$$

The first submanifold is of the form $M \times \mathbb{C}$, where $M \subset \mathbb{C}^2$ is a Bishop surface.

Let us from now on suppose that $A \neq 0$.

As M is Levi-flat, then through each CR point $p = (z_p, w_p) \in M_{\text{CR}}$ we have a complex submanifold of dimension one in M. It is well known that this submanifold is contained in the Segre variety (see also Section 4)

$$w = A(z, \bar{z}_p) + B(\bar{z}_p, \bar{z}_p), \quad \bar{w}_p = \bar{A}(\bar{z}_p, z) + \bar{B}(z, z). \tag{30}$$

By Lemma 3.1 we obtain a complex variety $V \subset M$ of dimension one through the origin. Suppose without loss of generality that V is irreducible. V has to be contained in the Segre variety at the origin, in particular $w = 0$ on V. Therefore, to simplify notation, let us consider V to be subvariety of $\{w = 0\}$. Denote by $\overline{V}$ the complex conjugate of V. Then as V is irreducible, $V \times \overline{V}$ is also irreducible (the smooth part of V is connected and so is the smooth part of $V \times \overline{V}$; see [Whitney 1972]). Hence, by complexifying, we have $A(z, \bar{\xi}) + B(\bar{\xi}, \bar{\xi}) = 0$ for all $z \in V$ and $\xi \in V$.

If $B \neq 0$, then setting $z = 0$, we have $B(\bar{\xi}, \bar{\xi}) = 0$ on V. As B is homogeneous and V is irreducible, V is a one-dimensional complex line. If $B = 0$, then $A(z, \bar{\zeta}) = 0$ for $z, \zeta \in V$ as mentioned above. We consider two cases. Suppose first that every $\sum_{j=1}^{2} a_{ij} \bar{\zeta}_j$ is identically zero for all $\zeta \in V$ and $i = 1, 2$. Then V is contained in some complex line $\sum_{j=1}^{2} \bar{a}_{ij} \zeta_j = 0$. Suppose now that $A(z, \bar{\zeta}_*)$ is not identically zero for some $\zeta_* \in V$. Then V is contained in the complex line $A(z, \bar{\zeta}_*) = 0$. This shows that V is a complex line.

Thus as $A(z, \bar{z}) + B(\bar{z}, \bar{z})$ is zero on a one-dimensional linear subspace, we make this subspace $\{z_1 = 0\}$ and so each monomial in $A(z, \bar{z}) + B(\bar{z}, \bar{z})$ is divisible by either z_1 or $\bar{z}_1$. Therefore, A and B are matrices of the form

$$\begin{bmatrix} * & * \\ * & 0 \end{bmatrix}, \tag{31}$$

that is, $a_{22} = 0$ and $b_{22} = 0$.

To normalize the pair (A, B), we apply arbitrary invertible transformations $(T, \lambda) \in \mathrm{GL}_n(\mathbb{C}) \times \mathbb{C}^*$ as

$$(A, B) \mapsto (\lambda T^* A T, \lambda T^* B \overline{T}). \tag{32}$$

Recall that we are assuming that $A \neq 0$. If $a_{21} = 0$ or $a_{12} = 0$, then A has rank one, and via a transformation T of the form

$$z_1' = z_1, \quad z_2' = z_2 + c z_1 \qquad \text{or} \qquad z_2' = z_1, \quad z_1' = z_2 + c z_1 \tag{33}$$

and rescaling by nonzero λ, the matrix A can be put in the form

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{34}$$

The transformation T and λ must also be applied to B and this could possibly make $b_{22} \neq 0$. However, we will show that we actually have $b_{22} = 0$. Thus $B = 0$ on $z_1 = 0$ still holds true.

Let us first focus on

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}. \tag{35}$$

We apply the $T^{(1,0)}$ vector field computed above. Only a_{11} is nonzero in A. Therefore $\beta^* A \beta$, which must be identically zero, is

$$\begin{aligned} 0 = \beta^* A \beta = \bar{\beta}_1 \beta_1 &= \overline{(2\bar{b}_{12} z_1 + 2\bar{b}_{22} z_2)}(2\bar{b}_{12} z_1 + 2\bar{b}_{22} z_2) \\ &= 4(|b_{12}|^2 z_1 \bar{z}_1 + |b_{22}|^2 z_2 \bar{z}_2 + b_{12}\bar{b}_{22} \bar{z}_1 z_2 + \bar{b}_{12} b_{22} z_1 \bar{z}_2). \end{aligned} \tag{36}$$

This polynomial must be identically zero and hence all coefficients must be identically zero. So $b_{12} = 0$ and $b_{22} = 0$. In other words, only b_{11} in B can be nonzero, in which case we make it nonnegative via a diagonal T to obtain the quadric

$$w = |z_1|^2 + \gamma \bar{z}_1^2, \quad \gamma \geq 0. \tag{37}$$

Next let us focus on

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{38}$$

As above, we compute $\beta^* A \beta$:

$$\begin{aligned} 0 = \beta^* A \beta = \bar{\beta}_1 \beta_2 &= \overline{(2\bar{b}_{12} z_1 + 2\bar{b}_{22} z_2)}(-\bar{z}_2 - 2\bar{b}_{11} z_1 - 2\bar{b}_{12} z_2) \\ = -2b_{12}\bar{z}_1\bar{z}_2 &- 2b_{22}\bar{b}_{11} z_1 \bar{z}_2 - 4\bar{b}_{11} b_{12} z_1 \bar{z}_1 - 4 b_{12}\bar{b}_{12}\bar{z}_1 z_2 \\ &\qquad - 2b_{22}\bar{z}_2^2 - 4b_{22}\bar{b}_{12} z_2 \bar{z}_2. \end{aligned} \tag{39}$$

Again, as this polynomial must be identically zero, all coefficients must be zero. Hence $b_{12} = 0$ and $b_{22} = 0$. Again only b_{11} is left possibly nonzero.

Suppose that $b_{11} \neq 0$. Then let s be such that $b_{11}\bar{s}^2 = 1$, and let $\bar{t} = 1/\bar{s}$. The matrix $T = \left[\begin{smallmatrix} s & 0 \\ 0 & t \end{smallmatrix}\right]$ is such that $T^* A T = A$ and $T^* B \overline{T} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$. If $b_{11} = 0$, we have

$B = 0$. Therefore we have obtained two distinct possibilities for B, and thus the two submanifolds

$$w = \bar{z}_1 z_2 \quad \text{or} \quad w = \bar{z}_1 z_2 + \bar{z}_1^2. \tag{40}$$

We emphasize that after A is normalized by a transformation of the form (33), only one coordinate change is needed to normalize b_{11} and this coordinate change preserves A. Both are required in a reduction proof for higher dimensions.

We have handled the rank-one case. Next we focus on the rank-two case, that is, $a_{21} \neq 0$ and $a_{12} \neq 0$ (recall $a_{22} = 0$). We normalize (rescale) A to have $a_{12} = 1$ and take

$$A = \begin{bmatrix} a_{11} & 1 \\ a_{21} & 0 \end{bmatrix}. \tag{41}$$

Again, let us compute $\beta^* A \beta$. In the computation for the rank-two case, recall that we have not done any normalization other than rescaling, so we can safely still assume that $b_{22} = 0$,

$$\begin{aligned} 0 = \beta^* A \beta &= a_{11} \bar{\beta}_1 \beta_1 + \bar{\beta}_1 \beta_2 + a_{21} \beta_1 \bar{\beta}_2 \\ &= a_{11} \overline{(\bar{a}_{21} \bar{z}_1 + 2\bar{b}_{12} z_1)} (\bar{a}_{21} \bar{z}_1 + 2\bar{b}_{12} z_1) \\ &\quad + \overline{(\bar{a}_{21} \bar{z}_1 + 2\bar{b}_{12} z_1)} (-\bar{a}_{11} \bar{z}_1 - \bar{z}_2 - 2\bar{b}_{11} z_1 - 2\bar{b}_{12} z_2) \\ &\quad + a_{21} \overline{(-\bar{a}_{11} \bar{z}_1 - \bar{z}_2 - 2\bar{b}_{11} z_1 - 2\bar{b}_{12} z_2)} (\bar{a}_{21} \bar{z}_1 + 2\bar{b}_{12} z_1) \\ &= (-4|b_{12}|^2 - |a_{21}|^2) \bar{z}_1 z_2 + (\text{other terms}). \end{aligned} \tag{42}$$

All coefficients must be zero. So $a_{21} = 0$, and A would not be rank two.

Let us now focus on $n > 2$. First suppose that $A = 0$. Then as before, M is automatically Levi-flat and by diagonalizing B we obtain the n distinct submanifolds

$$\begin{aligned} w &= \bar{z}_1^2, \\ w &= \bar{z}_1^2 + \bar{z}_2^2, \\ &\;\vdots \\ w &= \bar{z}_1^2 + \bar{z}_2^2 + \cdots + \bar{z}_n^2. \end{aligned} \tag{43}$$

Thus suppose from now on that $A \neq 0$. As before, we have an irreducible $(n-1)$-dimensional variety $V \subset M$ through the origin, such that $w = 0$ and $A(z, \bar{z}) + B(\bar{z}, \bar{z}) = 0$ on V.

We wish to show that $A(z, \bar{z}) + B(\bar{z}, \bar{z}) = 0$ on an $(n-1)$-dimensional linear subspace. For any $\xi \in V$, we obtain $A(z, \bar{\xi}) + B(\bar{\xi}, \bar{\xi}) = 0$ for all $z \in V$. If V is contained in the kernel of the matrix A^*, then V is a linear subspace of dimension

$n-1$. So suppose that $\bar{\xi}$ is not in the kernel of the matrix A^t. Then for a fixed $\bar{\xi}$, we obtain a linear equation $A(z,\bar{\xi}) + B(\bar{\xi},\bar{\xi}) = 0$ for $z \in V$.

Therefore, as $A(z,\bar{z}) + B(\bar{z},\bar{z})$ needs to be zero on an $(n-1)$-dimensional subspace, we can just make this $\{z_1 = 0\}$ and so each monomial is divisible by either z_1 or $\bar{z}_1$. Therefore, A and B is of the form

$$\begin{bmatrix} * & * & \cdots & * \\ * & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & 0 & \cdots & 0 \end{bmatrix}, \tag{44}$$

that is, only first column and first row are nonzero. We normalize A via

$$(A,B) \mapsto (\lambda T^* A T, \lambda T^* B \overline{T}), \tag{45}$$

as before. We use column operations on all but the first column to make all but the first two columns have nonzero elements. Similarly we can do row operations on all but the first two rows to make all but first three rows nonzero. That is, A has the form

$$\begin{bmatrix} * & * & 0 & \cdots & 0 \\ * & 0 & 0 & \cdots & 0 \\ * & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{46}$$

By Lemma 2.2, setting $z_3 = \cdots = z_n = 0$, we obtain a Levi-flat submanifold where the matrix corresponding to A is the principal 2×2 submatrix of A. This submatrix cannot be of rank two and hence either $a_{12} = 0$ or $a_{21} = 0$. If $a_{21} = 0$ and $a_{12} \neq 0$, then setting $z_2 = z_3$ and $z_4 = \cdots = z_n = 0$, we again must have a rank-one matrix and therefore $a_{31} = 0$.

Therefore, if $a_{12} \neq 0$ then all but a_{11} and a_{12} are zero. If $a_{12} = 0$, then via a further linear map not involving z_1 we can ensure that $a_{31} = 0$. In particular, A is of rank one and can only be nonzero in the principal 2×2 submatrix. At this point B is still of the form (44).

Via a linear change of coordinates in the first two variables, the principal 2×2 submatrix of A can be normalized into one of the two possible forms

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{47}$$

Recall that $A = 0$ was already handled.

Via the two-dimensional computation we obtain that $b_{22} = b_{12} = b_{21} = 0$. We use a linear map in z_1 and z_2 to normalize the principal 2×2 matrix of B, so that the submanifold restricted to (z_1, z_2, w) is in (B.γ), (C.0) or (C.1).

Finally we need to show that all entries of B other than b_{11} are zero. As we have done a linear change of coordinates in z_1 and z_2, B may not be in the form (44), but we know $b_{jk} = 0$ as long as $j > 2$ and $k > 2$.

Now fix $k = 3, \dots, n$. Restrict to the submanifold given by $z_1 = \lambda z_2$ for $\lambda = 1$ or $\lambda = -1$, and $z_j = 0$ for all $j = 3, \dots, n$ except for $j = k$. In the variables (z_2, z_k, w), we obtain a Levi-flat submanifold where the matrix corresponding to A is $\left[\begin{smallmatrix} \lambda & 0 \\ 0 & 0 \end{smallmatrix}\right]$. The matrix corresponding to B is

$$\begin{bmatrix} b_{11} & b_{1k} + \lambda b_{2k} \\ b_{1k} + \lambda b_{2k} & 0 \end{bmatrix}. \tag{48}$$

Via the two-dimensional calculation we have $b_{1k} + \lambda b_{2k} = 0$. As this is true for $\lambda = 1$ and $\lambda = -1$, we get that $b_{1k} = b_{2k} = 0$.

We have proved the following classification result. It is not difficult to see that the submanifolds in the list are biholomorphically inequivalent by Proposition 2.1. The ranks of A and B are invariants. It is obvious that the A matrix of B.γ and C.x submanifolds are inequivalent. Therefore, it is only necessary to directly check that B.γ are inequivalent for different $\gamma \geq 0$, which is easy.

Lemma 3.2. *If M defined in $(z, w) \in \mathbb{C}^n \times \mathbb{C}$, $n \geq 1$, by*

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}) \tag{49}$$

is Levi-flat, then M is biholomorphic to one and exactly one of the expressions (A.1)–(C.1) *of Theorem 1.1.*

The normalizing transformation used above is linear.

Lemma 3.3. *If M, defined by*

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}) + O(3), \tag{50}$$

is Levi-flat at all points where M is CR, then the quadric

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}) \tag{51}$$

is also Levi-flat.

Proof. Write M as

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}) + r(z, \bar{z}), \tag{52}$$

where r is $O(3)$.

Let A be the matrix giving the quadratic form $A(z, \bar{z})$ as before. The Levi map is given by taking the $n \times n$ matrix

$$L = L(p) = A + \left[\frac{\partial^2 r}{\partial z_j \partial \bar{z}_k} \right]_{j,k} \tag{53}$$

and applying it to vectors in $\pi(T^{(1,0)}M)$, where π is the projection onto the $\{w = 0\}$ plane. That is, we parametrize M by the $\{w = 0\}$ plane, and work there as before.

Let

$$\begin{aligned} a_j &= -\bar{A}_{z_j} - \bar{B}_{z_j} - \bar{r}_{z_j}, \\ b &= \bar{A}_{z_1} + \bar{B}_{z_1} + \bar{r}_{z_1}, \\ c &= a_j(A_{z_1} + B_{z_1} + r_{z_1}) + b(A_{z_j} + B_{z_j} + r_{z_j}). \end{aligned} \tag{54}$$

Then for $j = 2, \dots, n$, we write the $T^{(1,0)}$ vector fields as

$$X_j = a_j \frac{\partial}{\partial z_1} + b \frac{\partial}{\partial z_j} + c \frac{\partial}{\partial w}. \tag{55}$$

Hence

$$a_j \frac{\partial}{\partial z_1} + b \frac{\partial}{\partial z_j} \tag{56}$$

are the vector fields in $\pi(T^{(1,0)}M)$.

Notice that a_j, b, and c vanish at the origin. Furthermore, if we take the linear terms of a_j, b, and the quadratic terms in c, that is,

$$\begin{aligned} \tilde{a}_j &= -\bar{A}_{z_j} - \bar{B}_{z_j}, \\ \tilde{b} &= \bar{A}_{z_1} + \bar{B}_{z_1}, \\ \tilde{c} &= \tilde{a}_j(A_{z_1} + B_{z_1}) + \tilde{b}(A_{z_j} + B_{z_j}), \end{aligned} \tag{57}$$

then away from the CR singular set of the quadric,

$$\widetilde{X}_j = \tilde{a}_j \frac{\partial}{\partial z_1} + \tilde{b} \frac{\partial}{\partial z_j} + \tilde{c} \frac{\partial}{\partial w} \tag{58}$$

span the $T^{(1,0)}$ vector fields on the quadric $w = A(z, \bar{z}) + B(\bar{z}, \bar{z})$.

Since M is Levi-flat, then we have that

$$\pi_*(X_j)^* L \pi_*(X_j) = 0. \tag{59}$$

The terms linear in z and $\bar{z}$ respectively in the expression $\pi_*(X_j)^* L \pi_*(X_j)$ are

$$\pi_*(\widetilde{X}_j)^* A \pi_*(\widetilde{X}_j). \tag{60}$$

As this is identically zero, the quadric $w = A(z, \bar{z}) + B(\bar{z}, \bar{z})$ is Levi-flat. □

4. Quadratic Levi-flat submanifolds and their Segre varieties

A very useful invariant in CR geometry is the Segre variety. Suppose that a real-analytic variety $X \subset \mathbb{C}^N$ is defined by

$$\rho(z, \bar{z}) = 0, \tag{61}$$

where ρ is a real-analytic real vector-valued with $p \in X$. Suppose that ρ converges on some polydisc Δ centered at p. We complexify and treat z and $\bar{z}$ as independent variables, and the power series of ρ at $(p, \bar{p})$ converges on $\Delta \times \Delta$. The Segre variety at p is then defined as the variety

$$Q_p = \{z \in \Delta : \rho(z, \bar{p}) = 0\}. \tag{62}$$

Of course the variety depends on the defining equation itself and the polydisc Δ. For ρ, it is useful to take the defining equation or equations that generate the ideal of the complexified X in $\mathbb{C}^N \times \mathbb{C}^N$ at p. If ρ is polynomial we take $\Delta = \mathbb{C}^N$.

It is well known that any irreducible complex variety that lies in X and goes through the point p also lies in Q_p. In the case of Levi-flat submanifolds, we generally get equality as germs. For example, for the CR Levi-flat submanifold M given by

$$\operatorname{Im} z_1 = 0, \quad \operatorname{Im} z_2 = 0, \tag{63}$$

the Segre variety Q_0 through the origin is precisely $\{z_1 = z_2 = 0\}$, which happens to be the unique complex variety in M through the origin.

Let us take the Levi-flat quadric

$$w = A(z, \bar{z}) + B(\bar{z}, \bar{z}). \tag{64}$$

As we want to take the generating equations in the complexified space we also need the conjugate

$$\bar{w} = \bar{A}(\bar{z}, z) + \bar{B}(z, z). \tag{65}$$

The Segre variety is then given by

$$w = 0, \quad \bar{B}(z, z) = 0. \tag{66}$$

Through any CR singular point of a real-analytic Levi-flat M, there is a complex variety of dimension $n-1$ that is the limit of the leaves of the Levi foliation of M_{CR}, via Lemma 3.1. Let us take all possible such limits, and call their union Q'_p. Notice that there could be other complex varieties in M through p of dimension $n-1$. Note that $Q'_p \subset Q_p$.

Let us write down and classify the Segre varieties for all the quadric Levi-flat submanifolds in $\mathbb{C}^{n+1}$:

Type	Segre variety Q_0	Q_0 singular?	$\dim_{\mathbb{C}} Q_0$	$Q_0 \subset M$?	Q'_0
A.1	$w=0,\ z_1^2=0$	no	$n-1$	yes	Q_0
A.k	$w=0$ $z_1^2+\cdots+z_k^2=0$	yes	$n-1$	yes	Q_0
B.0	$w=0$	no	n	no	$w=0$ $z_1=0$
B.γ, $\gamma>0$	$w=0,\ z_1^2=0$	no	$n-1$	yes	Q_0
C.0	$w=0$	no	n	no	$w=0$ $z_1=0$
C.1	$w=0,\ z_1^2=0$	no	$n-1$	yes	Q_0

The submanifold C.0 also contains the complex variety $\{w=0, z_2=0\}$, but this variety is transversal to the leaves of the foliation, and so cannot be in Q'_0

Notice that in the cases A.k for all k, B.γ for $\gamma>0$, and C.1, the variety Q_0 actually gives the complex variety Q'_0 contained in M through the origin. In these cases, the variety is nonsingular only in the set theoretic sense. Scheme-theoretically the variety is always at least a double line or double hyperplane in general.

5. The CR singularity of Levi-flat quadrics

Proposition 5.1. *Let $M \subset \mathbb{C}^{n+1}$ be given by*

$$w=\rho(z,\bar{z}), \tag{67}$$

where ρ is $O(2)$, and M is not a complex submanifold. Then the set S of CR singularities of M is given by

$$S=\{(z,w) : \bar{\partial}\rho=0, w=\rho(z,\bar{z})\}. \tag{68}$$

Proof. This is well-known; we give a proof for convenience. In codimension two, a real submanifold is either CR singular, complex, or generic. A submanifold is generic if $\bar{\partial}$ of all the defining equations are pointwise linearly independent (see [Baouendi et al. 1999]). As M is not complex, to find the set of CR singularities, we find the set of points where M is not generic. We need both defining equations for M,

$$w=\rho(z,\bar{z}) \quad \text{and} \quad \bar{w}=\rho(z,\bar{z}). \tag{69}$$

As the second equation always produces a $d\bar{w}$ while the first does not, the only way that the two can be linearly dependent is for the $\bar{\partial}$ of the first equation to be zero. In other words, $\bar{\partial}\rho=0$. □

Let us compute and classify the CR singular sets for the CR singular Levi-flat quadrics:

Type	CR singularity S	$\dim_{\mathbb{R}} S$	CR structure of S
A.k	$z_1 = \cdots = z_k = 0,\ w = 0$	$2n-2k$	complex
B.0	$z_1 = 0,\ w = 0$	$2n-2$	complex
B.$\frac{1}{2}$	$z_1 + \bar{z}_1 = 0,\ w = 0$	$2n-1$	Levi-flat
B.γ, $\gamma > 0$, $\gamma \neq \frac{1}{2}$	$z_1 = 0,\ w = 0$	$2n-2$	complex
C.0	$z_2 = 0,\ w = 0$	$2n-2$	complex
C.1	$z_2 + 2\bar{z}_1 = 0,\ w = -\frac{1}{4}z_2^2$	$2n-2$	Levi-flat

By Levi-flat we mean that S is a Levi-flat CR submanifold in $\{w = 0\}$. There is a conjecture that a real subvariety that is Levi-flat at CR points has a stratification by Levi-flat CR submanifolds. This computation gives further evidence of this conjecture.

6. Levi foliations and images of generic Levi-flats

A CR Levi-flat submanifold $M \subset \mathbb{C}^n$ of codimension two has a certain canonical foliation defined on it with complex analytic leaves of real codimension two in M. The submanifold M is locally equivalent to $\mathbb{R}^2 \times \mathbb{C}^{n-2}$, defined by

$$\operatorname{Im} z_1 = 0, \quad \operatorname{Im} z_2 = 0. \tag{70}$$

The leaves of the foliation are the submanifolds given by fixing z_1 and z_2 at a real constant. By foliation we always mean the standard nonsingular foliation as locally comes up in the implicit function theorem. This foliation on M is called the *Levi foliation*. It is obvious that the Levi foliation on M extends to a neighborhood of M as a nonsingular holomorphic foliation. The same is not true in general for CR singular submanifolds. We say that a smooth holomorphic foliation $\mathcal{L}$ defined in a neighborhood of M is an *extension* of the Levi foliation of M_{CR}, if $\mathcal{L}$ and the Levi foliation have the same germs of leaves at each CR point of M. We also say that a smooth real-analytic foliation $\widetilde{\mathcal{L}}$ on M is an extension of the Levi foliation on M_{CR} if $\widetilde{\mathcal{L}}$ and the Levi foliation have the same germs of leaves at each CR point of M. In our situation (real-analytic), M_{CR} is a dense and open subset of M. This implies that the leaves of $\mathcal{L}$ and $\widetilde{\mathcal{L}}$ through a CR singular point are complex analytic submanifolds contained in M. The latter could lead to an obvious obstruction to extension. First let us see what happens if the foliation of M_{CR} is the restriction of a nonsingular holomorphic foliation of a whole neighborhood of M.

The Bishop-like quadrics, that is, A.1 and B.γ in $\mathbb{C}^{n+1}$, have a Levi foliation that extends as a holomorphic foliation to all of $\mathbb{C}^{n+1}$. That is because these submanifolds are of the form

$$N \times \mathbb{C}^{n-1}. \tag{71}$$

For submanifolds of the form (71) we can find normal forms using the well-developed theory of Bishop surfaces in $\mathbb{C}^2$.

Proposition 6.1. *Suppose $M \subset \mathbb{C}^{n+1}$ is a real-analytic Levi-flat CR singular submanifold where the Levi foliation on M_{CR} extends near $p \in M$ to a nonsingular holomorphic foliation of a neighborhood of p in $\mathbb{C}^{n+1}$. Then at p, M is locally biholomorphically equivalent to a submanifold of the form*

$$N \times \mathbb{C}^{n-1}, \tag{72}$$

where $N \subset \mathbb{C}^2$ is a CR singular submanifold of real dimension two. Therefore if M has a nondegenerate complex tangent, then it is Bishop-like, i.e., of type A.1 or B.γ.

Furthermore, two submanifolds of the form (72) *are locally biholomorphically* (*resp. formally*) *equivalent if and only if the corresponding Ns are locally biholomorphically* (*resp. formally*) *equivalent in $\mathbb{C}^2$.*

Proof. We flatten the holomorphic foliation near p so that in some polydisc Δ, the leaves of the foliation are given by $\{q\} \times \mathbb{C}^{n-1} \cap \Delta$ for $q \in \mathbb{C}^2$. Let us suppose that M is closed in Δ. At any CR point of M, the leaf of the Levi foliation agrees with that of the holomorphic foliation and therefore the leaf that lies in M agrees with a leaf of the form $\{q\} \times \mathbb{C}^{n-1}$ as a germ and so $\{q\} \times \mathbb{C}^{n-1} \cap \Delta \subset M$. As M_{CR} is dense in M, then M is a union of sets of the form $\{q\} \times \mathbb{C}^{n-1} \cap \Delta$ and the first part follows.

It is classical that every Bishop surface (two-dimensional real submanifold of $\mathbb{C}^2$ with a nondegenerate complex tangent) is equivalent to a submanifold whose quadratic part is of the form A.1 or B.γ.

Finally, the proof that two submanifolds of the form (72) are equivalent if and only if the Ns are equivalent is straightforward. □

Not every Bishop-like submanifold is a cross product as above. In fact the Bishop invariant may well change from point to point. See Section 15. In such cases the foliation does not extend to a nonsingular holomorphic foliation of a neighborhood.

Let us now focus on extending the Levi foliation to M, and not to a neighborhood of M. Let us prove a useful proposition about recognizing certain CR singular Levi-flats from the form of the defining equation. That is, if the r in the equation does not depend on $\bar{z}_2$ through $\bar{z}_n$.

Proposition 6.2. *Suppose that near the origin $M \subset \mathbb{C}^{n+1}$ is given by*

$$w = r(z_1, \bar{z}_1, z_2, z_3, \dots, z_n), \tag{73}$$

where r is $O(2)$ and $\partial r / \partial \bar{z}_1 \not\equiv 0$. Then M is a CR singular Levi-flat submanifold and the Levi foliation of M_{CR} extends through the origin to a real-analytic foliation on M. Furthermore, there exists a real-analytic CR mapping $F \colon U \subset \mathbb{R}^2 \times \mathbb{C}^{n-1} \to \mathbb{C}^{n+1}$, $F(0) = 0$, which is a diffeomorphism onto its image $F(U) \subset M$.

Near zero, M is the image of a CR mapping that is a diffeomorphism onto its image of the standard CR Levi-flat. The proposition also holds in two dimensions ($n = 1$), although in this case it is somewhat trivial.

Proof. As in [Lebl et al. 2014], let us define the mapping F by

$$(x, y, \xi) \mapsto (x + iy,\ \xi,\ r(x + iy, x - iy, \xi)), \tag{74}$$

where $\xi = (\xi_2, \dots, \xi_n) \in \mathbb{C}^{n-1}$. Near points where M is CR, this mapping is a CR diffeomorphism and hence M must be Levi-flat. Furthermore, since F is a diffeomorphism, it takes the Levi foliation on $\mathbb{R}^2 \times \mathbb{C}^{n-1}$ to a foliation on M near zero. □

Lemma 6.3. *Let $M \subset \mathbb{C}^{n+1}$ be a CR singular real-analytic Levi-flat submanifold of codimension two through the origin.*

Then M is a CR singular Levi-flat submanifold whose Levi foliation of M_{CR} extends through the origin to a nonsingular real-analytic foliation on M if and only if there exists a real-analytic CR mapping $F\colon U \subset \mathbb{R}^2 \times \mathbb{C}^{n-1} \to \mathbb{C}^{n+1}$, $F(0) = 0$, which is a diffeomorphism onto its image $F(U) \subset M$.

Proof. One direction is easy and was used above. For the other direction, suppose that we have a foliation extending the Levi foliation through the origin. Let us consider M_{CR} an abstract CR manifold. That is a manifold M_{CR} together with the bundle $T^{(0,1)}M_{\mathrm{CR}} \subset \mathbb{C} \otimes TM_{\mathrm{CR}}$. The extended foliation on M gives a real-analytic subbundle $\mathcal{W} \subset TM$. Since we are extending the Levi foliation, when $p \in M_{\mathrm{CR}}$, then $\mathcal{W}_p = T_p^c M$, where $T_p^c M = J(T_p^c M)$ is the complex tangent space and J is the complex structure on $\mathbb{C}^{n+1}$. Since M_{CR} is dense in M, then $J\mathcal{W} = \mathcal{W}$ on M.

Define the real-analytic subbundle $\mathcal{V} \subset \mathbb{C} \otimes TM$ as

$$\mathcal{V}_p = \{X + iJ(X) : X \in \mathcal{W}_p\}. \tag{75}$$

At CR points $\mathcal{V}_p = T_p^{(0,1)}M$ (see, for example, [Baouendi et al. 1999, p. 8]). Then we can find vector fields $X^1, \dots, X^{n-1}$ in $\mathcal{W}$ such that

$$X^1, J(X^1), X^2, J(X^2), \dots, X^{n-1}, J(X^{n-1}) \tag{76}$$

is a basis of $\mathcal{W}$ near the origin. Then the basis for $\mathcal{V}$ is given by

$$X^1 + iJ(X^1), X^2 + iJ(X^2), \dots, X^{n-1} + iJ(X^{n-1}). \tag{77}$$

As the subbundle is integrable, we obtain that $(M, \mathcal{V})$ gives an abstract CR manifold, which at CR points agrees with M_{CR}. This manifold is Levi-flat as it is Levi-flat on a dense open set. As it is real-analytic, it is embeddable; and hence there exists a real-analytic CR diffeomorphism from a neighborhood of $\mathbb{R}^2 \times \mathbb{C}^{n-1}$ to a neighborhood of zero in M (as an abstract CR manifold). This is our mapping F. □

The quadrics A.k, $k \geq 2$, defined by

$$w = \bar{z}_1^2 + \cdots + \bar{z}_k^2, \tag{78}$$

contain the singular variety defined by $w = 0$, $z_1^2 + \cdots + z_k^2 = 0$, and hence the Levi foliation cannot extend to a nonsingular foliation of the submanifold. The quadric A.1 does admit a holomorphic foliation, but other type A.1 submanifolds do not in general. For example, the submanifold

$$w = \bar{z}_1^2 + \bar{z}_2^3 \tag{79}$$

is of type A.1 and the unique complex variety through the origin is $0 = z_1^2 + z_2^3$, which is singular. Therefore the foliation cannot extend to M.

7. Extending the Levi foliation of type C.x submanifolds

Let us prove Theorem 1.2, that is, let us start with a type C.0 or C.1 submanifold and show that the Levi foliation must extend real-analytically to all of M. Equivalently, we show that the real-analytic bundle $T^{(1,0)}M_{\mathrm{CR}}$ extends to a real-analytic subbundle of $\mathbb{C} \otimes TM$. Taking real parts we obtain an involutive subbundle of TM extending $T^c M_{\mathrm{CR}} = \mathrm{Re}(T^{(1,0)}M_{\mathrm{CR}})$.

Proof of Theorem 1.2. Let M be the submanifold given by

$$w = \bar{z}_1 z_2 + \epsilon \bar{z}_1^2 + r(z, \bar{z}), \tag{80}$$

where $\epsilon = 0, 1$. Let us treat the z variables as the parameters on M. Let π be the projection onto the $\{w = 0\}$ plane, which is tangent to M at zero as a real $2n$-dimensional hyperplane. We will look at all the vector fields on this plane $\{w = 0\}$. All vectors in $\pi(T^{(1,0)}M)$ can be written in terms of $\partial/\partial z_j$ for $j = 1, \ldots, n$.

The Levi map is given by taking the $n \times n$ matrix

$$L = L(p) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} + \left[\frac{\partial^2 r}{\partial z_j \partial \bar{z}_k} \right]_{j,k} (p) \tag{81}$$

to vectors $v \in \pi(T^{(1,0)}M)$ (π is the projection) as $v^* L v$. The excessive term in L vanishes at zero.

Notice that for $p \in M_{\mathrm{CR}}$, $\pi(T_p^{(1,0)}M)$ is $(n-1)$-dimensional. As M is Levi-flat, then $v^* L v$ vanishes for $v \in \pi(T_p^{(1,0)}M)$. Write the vector $v = (v_1, \ldots, v_n)^t$. The zero set of the function

$$(z, v) \in \mathbb{C}^n \times \mathbb{C}^n \overset{\varphi}{\mapsto} v^* L(z, \bar{z}) v \tag{82}$$

is a variety V of real codimension two at the origin of $\mathbb{C}^n \times \mathbb{C}^n$ because of the form of L. That is, at $z = 0$, the only vectors v such that $v^*Lv = 0$ are those where $v_1 = 0$ or $v_2 = 0$. So the codimension is at least two. And we know that v^*Lv vanishes for vectors in $\pi(T_p^{(1,0)}M)$ for $p \in M$ near zero, which is real codimension two at each z corresponding to a CR point. Therefore, $V \cap (\pi(M_{\mathrm{CR}}) \times \mathbb{C}^n)$ has a connected component that is equal to a connected component of the real-analytic subbundle $\pi(T^{(1,0)}M_{\mathrm{CR}})$. We will verify that the latter is connected.

We show below that this subbundle extends past the CR singularity. The key point is to show that the restriction of $\pi(T^{(1,0)}(M_{\mathrm{CR}}))$ extends to a smooth real-analytic submanifold of $T^{(1,0)}\mathbb{C}^n$. Write

$$\varphi(z, v) = v_1\bar{v}_2 + \sum a_{jk}(z)v_j\bar{v}_k, \tag{83}$$

where $a_{jk}(0) = 0$.

By Proposition 5.1, $\pi(M \setminus M_{\mathrm{CR}})$ is contained in

$$z_2 + 2\epsilon\bar{z}_1 + r_{\bar{z}_1} = 0. \tag{84}$$

Thus M_{CR} is connected. Assume that $v \cdot (\partial/\partial z) \in T_p^{(1,0)}M$ at a CR point p. Then

$$(z_2 + 2\epsilon\bar{z}_1 + r_{\bar{z}_1})\bar{v}_1 + \sum_{j>1} r_{\bar{z}_j}\bar{v}_j = 0. \tag{85}$$

When p is in the open set $U_\delta \subset \pi(M_{\mathrm{CR}})$ defined by $|z_2 + 2\epsilon\bar{z}_1| > \frac{1}{2}|z|$ and $0 < |z| < \delta$, v is contained in

$$V_C \colon |v_1| \le |v|/C. \tag{86}$$

When δ is sufficiently small, $\varphi(z, v) = 0$ admits a unique solution

$$v_1 = f(z, v_3, \dots, v_n), \quad v_2 = 1 \tag{87}$$

by imposing $v \in V_C$. Note that f is given by convergent power series. For $|z| < \delta$, define

$$w_j = (w_{j1}(z), \dots, w_{jn}(z)) \in V_C, \quad j = 2, \dots, n \tag{88}$$

such that $\varphi(z, w_j(z)) = 0$ and

$$w_{j2} = 1, \quad w_{jk} = \delta_{jk}, \quad j \ge 2,\ k > 2. \tag{89}$$

To see why we can do so, fix $p \in U_\delta$. First we can find a vector w_2 in $E_p = \pi(T_p^{(1,0)}M_{\mathrm{CR}})$ such that $v_2 = 1$. Otherwise, $E_p \subset V_C$ cannot have dimension $n-1$. Let E_p' be the vector subspace of E_p with $v_2 = 0$. Then E_p' has rank $n-2$ and remains in the cone V_C. Then E_p' has an element w_2 with v_2 component being one. Repeating this, we find $w_2, \dots, w_n$ in E_p such that the v_j component of w_i is zero for $2 < j < i$. Using linear combinations, we find a unique basis $\{w_2, \dots, w_n\}$ of E_p that satisfies condition (89).

Assume that C is sufficiently large. By the above uniqueness assertion on $\varphi(z, v) = 0$, we conclude that when $p \in U_\delta$, $\{w_2(p), \dots, w_n(p)\}$ is a basis of $\pi(T_p^{(1,0)} M_{\mathrm{CR}})$. Also it is real analytic at $p = 0$. Define

$$\omega_j(z) = w_j(z) \cdot \frac{\partial}{\partial z}, \quad |z| < \delta. \tag{90}$$

We lift the functions ω_j via π to a subbundle of $\mathbb{C} \otimes TM$; let us call them $\tilde{\omega}_j$. Then consider the vector fields $w_j^* = 2 \operatorname{Re} \tilde{\omega}_j = \tilde{\omega}_j + \overline{\tilde{\omega}_j}$ and $w_{n+j}^* = \operatorname{Im} \tilde{\omega}_j$ for $j = 2, \dots, n$. Above CR points over U_δ, $\tilde{w}_j$ is in $TM_{\mathrm{CR}} \otimes \mathbb{C}$ and so tangent to M. We thus obtain a $(2n-2)$-dimensional real-analytic subbundle of TM that agrees with the real-analytic real subbundle of TM_{CR} induced by the Levi foliation above U_δ. Since M_{CR} and the subbundles are real analytic and M_{CR} is connected, they agree over M_{CR}.

The real-analytic distribution spanned by $\{\omega_i^*\}$ has constant rank $2n-2$ everywhere and is involutive on an open subset of M_{CR} and hence everywhere. □

8. CR singular set of type C.x submanifolds

Let $M \subset \mathbb{C}^{n+1}$ be a codimension-two Levi-flat CR singular submanifold that is an image of $\mathbb{R}^2 \times \mathbb{C}^{n-1}$ via a real-analytic CR map, and let $S \subset M$ be the CR singular set of M. In [Lebl et al. 2014] it was proved that near a generic point of S, exactly one of the following is true:

(i) S is a Levi-flat submanifold of dimension $2n-2$ and CR dimension $n-2$.

(ii) S is a complex submanifold of complex dimension $n-1$ (real dimension $2n-2$).

(iii) S is a Levi-flat submanifold of dimension $2n-1$ and CR dimension $n-1$.

We only have the above classification for a generic point of S, and S need not be a CR submanifold everywhere. See [Lebl et al. 2014] for examples.

If M is a Levi-flat CR singular submanifold and the Levi foliation of M_{CR} extends to M, then by Lemma 6.3 at a generic point S has to be of one of the above types. A corollary of Theorem 1.2 is:

Corollary 8.1. *Suppose that $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is a real-analytic Levi-flat CR singular type C.1 or type C.0 submanifold. Let $S \subset M$ denote the CR singular set. Then near the origin S is a submanifold of dimension $2n-2$, and at a generic point, S is either CR Levi-flat of dimension $2n-2$ (CR dimension $n-2$) or a complex submanifold of complex dimension $n-1$.*

Furthermore, if M is of type C.1, then at the origin S is a CR Levi-flat submanifold of dimension $2n-2$ (CR dimension $n-2$).

Proof. Let M be given by

$$w = \bar{z}_1 z_2 + \epsilon \bar{z}_1^2 + r(z, \bar{z}), \tag{91}$$

where r is $O(3)$ and $\epsilon = 0$ or $\epsilon = 1$.

By Proposition 5.1 the CR singular set is exactly where

$$z_2 + \epsilon 2\bar{z}_1 + r_{\bar{z}_1}(z, \bar{z}) = 0 \qquad \text{and} \qquad r_{\bar{z}_j}(z, \bar{z}) = 0 \quad \text{for all } j = 2, \ldots, n. \tag{92}$$

By considering the real and imaginary parts of the first equation and applying the implicit function theorem, the set $\widetilde{S} = \{z : z_2 + \epsilon 2\bar{z}_1 + r_{\bar{z}_1}(z, \bar{z}) = 0\}$ is a real submanifold of real dimension $2n - 2$ (real codimension two in M). Now $S \subset \widetilde{S}$, but as we saw above that S is of dimension at least $2n-2$. Therefore $S = \widetilde{S}$ near the origin. The conclusion of the first part then follows from the classification above.

The stronger conclusion for type C.1 submanifolds follows by noticing that when $\epsilon = 1$, the submanifold

$$z_2 + 2\bar{z}_1 + r_{\bar{z}_1}(z, \bar{z}) = 0 \tag{93}$$

is CR and not complex at the origin. □

9. Mixed-holomorphic submanifolds

Let us study sets in $\mathbb{C}^m$ defined by

$$f(\bar{z}_1, z_2, \ldots, z_m) = 0, \tag{94}$$

for a single holomorphic function f of m variables.

Such sets have much in common with complex varieties, since they are in fact complex varieties when $\bar{z}_1$ is treated as a complex variable. The distinction is that the automorphism group is different since we are interested in automorphisms that are holomorphic, not mixed-holomorphic.

Proposition 9.1. *If $M \subset \mathbb{C}^m$ is a submanifold with a defining equation of the form* (94), *where f is a holomorphic function that is not identically zero, then M is a real codimension-two set and M is either a complex submanifold or a Levi-flat submanifold, possibly CR singular. Furthermore, if M is CR singular at $p \in M$, and has a nondegenerate complex tangent at p, then M has type A.k, C.0, or C.1 at p.*

Proof. Since the zero set of f is a complex variety in the $(\bar{z}_1, z_2, \ldots, z_m)$ space, we get automatically that it is real codimension two. We also have that as it is a submanifold, then it can be written as a graph of one variable over the rest.

Let $m = n + 1$ for convenience and suppose that $M \subset \mathbb{C}^{n+1}$ is a submanifold through the origin. By factorization for germs of holomorphic functions and by the smoothness assumption on M we may assume that $df(0) \neq 0$. Call the variables $(z_1, \dots, z_n, w)$ and write M as a graph. One possibility is that we write M as

$$\bar{w} = \rho(z_1, \dots, z_n), \tag{95}$$

where $\rho(0) = 0$ and ρ has no linear terms. M is complex if $\rho \equiv 0$. Otherwise M is CR singular and we rewrite it as

$$w = \bar{\rho}(\bar{z}_1, \dots, \bar{z}_n). \tag{96}$$

We notice that the matrix representing the Levi-map must be identically zero, so we must get Levi-flat. If there are any quadratic terms, we obtain a type A.k submanifold.

Alternatively, M can be written as

$$w = \rho(\bar{z}_1, z_2, \dots, z_n), \tag{97}$$

with $\rho(0) = 0$. If ρ does not depend on $\bar{z}_1$ then M is complex. Assume that ρ depends on $\bar{z}_1$. If ρ has linear terms in $\bar{z}_1$, then M is CR. Otherwise it is a CR singular submanifold, and near a nonCR singular point it is a generic codimension-two submanifold. The CR singular set of M is defined by $\partial\rho/\partial\bar{z}_1 = 0$.

Suppose that M is CR singular. That M is Levi-flat follows from Proposition 6.2. We can therefore normalize the quadratic term, after linear terms in $z_2, \dots, z_n$ are absorbed into w.

If not all quadratic terms are zero, we must have an A.k, C.0, or C.1 type submanifold. □

Let us now study normal forms for such sets in $\mathbb{C}^2$ and $\mathbb{C}^m$, $m \geq 3$. First in two variables we can completely answer the question. This result is surely well known and classical.

Proposition 9.2. *If $M \subset \mathbb{C}^2$ is a submanifold with a defining equation of the form* (94), *then it is locally biholomorphically equivalent to a submanifold in coordinates $(z, w) \in \mathbb{C}^2$ of the form*

$$w = \bar{z}^d \tag{98}$$

for $d = 0, 1, 2, 3, \dots$, where d is a local biholomorphic invariant of M. If $d = 0$, M is complex; *if $d = 1$, it is a CR totally-real submanifold*; *and if $d \geq 2$, then M is CR singular.*

Proof. Write the submanifold as a graph of one variable over the other. Without loss of generality and after possibly taking a conjugate of the equation, we have

$$w = f(\bar{z}) \tag{99}$$

for some holomorphic function f. Assume $f(0) = 0$. If f is identically zero, then $d = 0$ and we are finished. If f is not identically zero, we apply a holomorphic change of coordinates in z, and the rest follows easily. □

In three or more variables, if $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is a submanifold through the origin, then when the quadratic part is nonzero we have seen above that it can be a type A.k, C.0, or C.1 submanifold. If the submanifold is the nondegenerate type C.1 submanifold, then we will show in the next section that M is biholomorphically equivalent to the quadric $M_{C.1}$.

Before we move to type C.1, let us quickly consider the mixed-holomorphic submanifolds of type A.n. The submanifolds of type A.n in $\mathbb{C}^{n+1}$ can in some sense be considered nondegenerate when talking about mixed-holomorphic submanifolds.

Proposition 9.3. *If $M \subset \mathbb{C}^{n+1}$ is a submanifold of type A.n at the origin of the form*

$$w = \bar{z}_1^2 + \cdots + \bar{z}_n^2 + r(\bar{z}), \tag{100}$$

where $r \in O(3)$. Then M is locally near the origin biholomorphically equivalent to the A.n quadric

$$w = \bar{z}_1^2 + \cdots + \bar{z}_n^2. \tag{101}$$

Proof. The complex Morse lemma (see, e.g., [Ebeling 2007, Proposition 3.15]) states that there is a local change of coordinates near the origin in just the z variables such that

$$z_1^2 + \cdots + z_n^2 + \bar{r}(z) \tag{102}$$

is equivalent to $z_1^2 + \cdots + z_n^2$. □

It is not difficult to see that the normal form for mixed-holomorphic submanifolds in $\mathbb{C}^{n+1}$ of type A.k, $k < n$, is equivalent to a local normal form for a holomorphic function in n variables. Therefore, for example, the submanifold $w = \bar{z}_1^2 + \bar{z}_2^3$ is of type A.1 and is not equivalent to any quadric.

10. Formal normal form for certain type C.1 submanifolds I

In this section we prove the formal normal form in Theorem 1.3. That is, we prove that if $M \subset \mathbb{C}^{n+1}$ is defined by

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(z_1, \bar{z}_1, z_2, z_3, \ldots, z_n), \tag{103}$$

where r is $O(3)$, then M is Levi-flat and formally equivalent to

$$w = \bar{z}_1 z_2 + \bar{z}_1^2. \tag{104}$$

That M is Levi-flat follows from Proposition 6.2.

Lemma 10.1. *If $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(z_1, \bar{z}_1, z_2, z_3, \dots, z_n), \tag{105}$$

where r is an $O(3)$ formal power series, then M is formally equivalent to $M_{C.1}$ given by

$$w = \bar{z}_1 z_2 + \bar{z}_1^2. \tag{106}$$

In fact, the normalizing transformation can be of the form

$$(z, w) = (z_1, \dots, z_n, w) \mapsto (z_1,\ f(z, w),\ z_3, \dots, z_n,\ g(z, w)), \tag{107}$$

where f and g are formal power series.

Proof. Suppose that the normalization was done to degree $d - 1$, then suppose that

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r_1(z_1, \bar{z}_1, z_2, \dots, z_n) + r_2(z_1, \bar{z}_1, z_2, \dots, z_n), \tag{108}$$

where r_1 is degree-d homogeneous and r_2 is $O(d+1)$. Write

$$r_1(z_1, \bar{z}_1, z_2, \dots, z_n) = \sum_{j=0}^{k} \sum_{|\alpha|+j=d} c_{j,\alpha} \bar{z}_1^j z^\alpha, \tag{109}$$

where k is the highest power of $\bar{z}_1$ in r_1, and α is a multiindex.

If k is even, then we use the transformation that replaces w with

$$w + \sum_{|\alpha|+k=d} c_{j,\alpha} w^{k/2} z^\alpha. \tag{110}$$

Let us look at the degree-d terms in

$$(\bar{z}_1 z_2 + \bar{z}_1^2) + \sum_{|\alpha|+k=d} c_{j,\alpha} (\bar{z}_1 z_2 + \bar{z}_1^2)^{k/2} z^\alpha = \bar{z}_1 z_2 + \bar{z}_1^2 + r_1(z_1, \bar{z}_1, z_2, \dots, z_n). \tag{111}$$

We need not include r_2 as the terms have degree $d+1$ or higher. After canceling out the new terms on the left, we notice that the formal transformation removed all the terms in r_1 with a power $\bar{z}_1^k$ and replaced them with terms that have a smaller power of $\bar{z}_1$.

Next suppose that k is odd. We use the transformation that replaces z_2 with

$$z_2 - \sum_{|\alpha|+k=d} c_{j,\alpha} w^{(k-1)/2} z^\alpha. \tag{112}$$

Let us look at the degree-d terms in

$$\begin{aligned} \bar{z}_1 z_2 + \bar{z}_1^2 \\ = \bar{z}_1 \Bigl(z_2 - \sum_{|\alpha|+k=d} c_{j,\alpha} w^{(k-1)/2} z^{\alpha} \Bigr) + \bar{z}_1^2 + r_1(z_1, \bar{z}_1, z_2, \dots, z_n). \end{aligned} \tag{113}$$

Again we need not include r_2 as the terms have degree $d+1$ or higher, and we need not add the new terms to z_2 in the argument list for r_1 since all those terms would be of higher degree. Again we notice that the formal transformation removed all the terms in r_1 with a power $\bar{z}_1^k$ and replaced them with terms that have a smaller power of $\bar{z}_1$.

The procedure above does not change the form of the submanifold, but it lowers the degree of $\bar{z}_1$ by one. Since we can assume that all terms in r_1 depend on $\bar{z}_1$, we are finished with degree-d terms after k iterations of the above procedure. □

11. Convergence of normalization for certain type C.1 submanifolds

A key point in the computation below is the following natural involution for the quadric $M_{C.1}$. Notice that the map

$$(z_1, z_2, \dots, z_n, w) \mapsto (-\bar{z}_2 - z_1,\ z_2, \dots, z_n,\ w) \tag{114}$$

takes $M_{C.1}$ to itself. The involution simply replaces the $\bar{z}_1$ in the equation with $-z_2 - \bar{z}_1$. The way this involution is defined is by noticing that the equation $w = \bar{z}_1 z_2 + \bar{z}_1^2$ has generically two solutions for $\bar{z}_1$ keeping z_2 and w fixed. In the same way we could define an involution on all type C.1 submanifolds of the form $w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(\bar{z}_1, z_2, \dots, z_n)$, although we will not require this construction.

We prove convergence via:

Lemma 11.1. *Let $m_1, \dots, m_N$ be positive integers. Suppose $T(z)$ is a formal power series in $z \in \mathbb{C}^N$. Suppose $T(t^{m_1} v_1, \dots, t^{m_N} v_N)$ is a convergent power series in $t \in \mathbb{C}$ for all $v \in \mathbb{C}^N$. Then T is convergent.*

The proof is a standard application of the Baire category theorem and the Cauchy inequality. See [Baouendi et al. 1999, p. 153, Theorem 5.5.30], where all m_j are one. For $m_j > 1$ we first change variables by setting $v_j = w_j^{m_j}$ and apply the lemma with $m_j = 1$.

The following lemma finishes the proof of Theorem 1.3. By absorbing any holomorphic terms into w, we assume that $r(z_1, 0, z_2, \dots, z_n) \equiv 0$. In Lemma 10.1 we have also constructed a formal transformation that only changes the z_2 and w coordinates, so it is enough to prove convergence in this case. Key points of this proof are that the right hand side of the defining equation for $M_{C.1}$ is homogeneous, and that we have a natural involution on $M_{C.1}$.

Lemma 11.2. *If $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(z_1, \bar{z}_1, z_2, z_3, \dots, z_n), \tag{115}$$

where r is $O(3)$ and convergent, and $r(z_1, 0, z_2, \dots, z_n) \equiv 0$. Suppose that two formal power series $f(z,w)$ and $g(z,w)$ satisfy

$$\begin{aligned} &g(z, \bar{z}_1 z_2 + \bar{z}_1^2) \\ &\qquad = \bar{z}_1 f(z, \bar{z}_1 z_2 + \bar{z}_1^2) + \bar{z}_1^2 + r(z_1, \bar{z}_1, f(z, \bar{z}_1 z_2 + \bar{z}_1^2), z_3, \dots, z_n). \end{aligned} \tag{116}$$

Then f and g are convergent.

Proof. Equation (116) is true formally, treating z_1 and $\bar{z}_1$ as independent variables. Notice that (116) has one equation for two unknown functions.

We now use the involution on $M_{C.1}$ to create a system that we can solve uniquely. We replace $\bar{z}_1$ with $-z_2 - \bar{z}_1$. We leave z_1 untouched (treating as an independent variable). We obtain an identity in formal power series:

$$\begin{aligned} g(z, \bar{z}_1 z_2 + \bar{z}_1^2) = (-z_2 - \bar{z}_1) f(z, \bar{z}_1 z_2 + \bar{z}_1^2) + (-z_2 - \bar{z}_1)^2 \\ + r(z_1, (-z_2 - \bar{z}_1), f(z, \bar{z}_1 z_2 + \bar{z}_1^2), z_3, \dots, z_n). \end{aligned} \tag{117}$$

The formal series $\xi = f(z, \bar{z}_1 z_2 + \bar{z}_1^2)$ and $\omega = g(z, \bar{z}_1 z_2 + \bar{z}_1^2)$ are solutions of the system

$$\omega = \bar{z}_1 \xi + \bar{z}_1^2 + r(z_1, \bar{z}_1, \xi, z_3, \dots, z_n), \tag{118}$$

$$\omega = (-z_2 - \bar{z}_1)\xi + (-z_2 - \bar{z}_1)^2 + r(z_1, (-z_2 - \bar{z}_1), \xi, z_3, \dots, z_n). \tag{119}$$

We next replace z_j with tz_j and $\bar{z}_1$ with $t\bar{z}_1$ for $t \in \mathbb{C}$. Because $\bar{z}_1 z_2 + \bar{z}_1^2$ is homogeneous of degree two, we obtain that the formal series in t, given by $\xi(t) = f(tz, t^2(\bar{z}_1 z_2 + \bar{z}_1^2))$, $\omega(t) = g(tz, t^2(\bar{z}_1 z_2 + \bar{z}_1^2))$ for every $(z_1, \bar{z}_1, z_2, \dots, z_n) \in \mathbb{C}^{n+1}$, are solutions of the system

$$\omega = t\bar{z}_1 \xi + t^2 \bar{z}_1^2 + r(tz_1, t\bar{z}_1, \xi, tz_3, \dots, tz_n), \tag{120}$$

$$\omega = t(-z_2 - \bar{z}_1)\xi + t^2(-z_2 - \bar{z}_1)^2 + r(tz_1, t(-z_2 - \bar{z}_1), \xi, tz_3, \dots, tz_n). \tag{121}$$

We eliminate ω to obtain an equation for ξ,

$$\begin{aligned} &t(2\bar{z}_1 + z_2)(\xi - tz_2) \\ &\qquad = r(tz_1, t(-z_2 - \bar{z}_1), \xi, tz_3, \dots, tz_n) - r(tz_1, t\bar{z}_1, \xi, tz_3, \dots, tz_n). \end{aligned} \tag{122}$$

We now treat ξ as a variable and we have a holomorphic (convergent) equation. The right hand side must be divisible by $t(2\bar{z}_1 + z_2)$: It is divisible by t since r is divisible by $\bar{z}_1$. It is also divisible by $2\bar{z}_1 + z_2$ as setting $z_2 = -2\bar{z}_1$ makes the right hand side vanish. Therefore,

$$\xi - tz_2 = \frac{r(tz_1, t(-z_2 - \bar{z}_1), \xi, tz_3, \dots, tz_n) - r(tz_1, t\bar{z}_1, \xi, tz_3, \dots, tz_n)}{t(2\bar{z}_1 + z_2)}, \tag{123}$$

where the right hand side is a holomorphic function (that is, a convergent power series) in $z_1, \bar{z}_1, z_2, \dots, z_n, t, \xi$. For any fixed $z_1, \bar{z}_1, z_2, \dots, z_n$, we solve for ξ in terms of t via the implicit function theorem, and we obtain that ξ is a holomorphic function of t. The power series of ξ is given by $\xi(t) = f(tz, t^2(\bar{z}_1 z_2 + \bar{z}_1^2))$.

Let $v \in \mathbb{C}^{n+1}$ be any nonzero vector. Via a proper choice of $z_1, \bar{z}_1, z_2, \dots, z_n$ (still treating $\bar{z}_1$ and z_1 as independent variables) we write $v = (z, \bar{z}_1 z_2 + \bar{z}_1^2)$. We apply the above argument to $\xi(t) = f(tv_1, \dots, tv_n, t^2 v_{n+1})$, and $\xi(t)$ converges as a series in t. As we get convergence for every $v \in \mathbb{C}^{n+1}$, we obtain that f converges by Lemma 11.1. Then via (120) we obtain that $g(tv_1, \dots, tv_n, t^2 v_{n+1})$ converges as a series in t for all v, and hence g converges. □

12. Automorphism group of the C.1 quadric

With the normal form achieved in previous sections, let us study the automorphism group of the C.1 quadric in this section. We will again use the mixed-holomorphic involution that is obtained from the quadric.

We study the local automorphism group at the origin. That is the set of germs at the origin of biholomorphic transformations taking M to M and fixing the origin.

First we look at the linear parts of automorphisms. We already know that the linear term of the last component only depends on w. For $M_{C.1}$ we can say more about the first two components.

Proposition 12.1. *Let $(F, G) = (F_1, \dots, F_n, G)$ be a formal invertible or biholomorphic automorphism of $M_{C.1} \subset \mathbb{C}^{n+1}$, that is the submanifold of the form*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2. \tag{124}$$

Then $F_1(z, w) = az_1 + \alpha w + O(2)$, $F_2(z, w) = \bar{a} z_2 + \beta w + O(2)$, and $G(z, w) = \bar{a}^2 w + O(2)$, where $a \neq 0$.

Proof. Let $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ be such that

$$F_1(z, w) = a \cdot z + \alpha w + O(2) \quad \text{and} \quad F_2(z, w) = b \cdot z + \beta w + O(2).$$

Then from Proposition 2.1 we have

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} = \lambda \begin{bmatrix} a^* & b^* & \cdots \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ \vdots \end{bmatrix}. \tag{125}$$

Therefore $\lambda \bar{a}_1 b_2 = 1$, and $\bar{a}_j b_k = 0$ for all $(j,k) \neq (1,2)$. Similarly,

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} = \lambda \begin{bmatrix} a^* & b^* & \cdots \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \bar{a} \\ \bar{b} \\ \vdots \end{bmatrix}. \tag{126}$$

Therefore $\lambda \bar{a}_1^2 = 1$, and $\bar{a}_j \bar{a}_k = 0$ for all $(j,k) \neq (1,1)$. Putting these two together we obtain that $a_j = 0$ for all $j \neq 1$, and as $a_1 \neq 0$ we get $b_j = 0$ for all $j \neq 2$. As λ is the reciprocal of the coefficient of w in G, we are finished. □

Lemma 12.2. *Let $M_{C.1} \subset \mathbb{C}^3$ be given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2. \tag{127}$$

Suppose that a local biholomorphism (resp. formal automorphism) (F_1, F_2, G) transforms $M_{C.1}$ into $M_{C.1}$. Then F_1 depends only on z_1, and F_2 and G depend only on z_2 and w.

Proof. Let us define a $(1,0)$ tangent vector field on M by

$$Z = \frac{\partial}{\partial z_2} + \bar{z}_1 \frac{\partial}{\partial w}. \tag{128}$$

Write $F = (F_1, F_2, G)$. F must take Z into a multiple of itself when restricted to $M_{C.1}$. That is, on $M_{C.1}$ we have

$$\frac{\partial F_1}{\partial z_2} + \bar{z}_1 \frac{\partial F_1}{\partial w} = 0, \tag{129}$$

$$\frac{\partial F_2}{\partial z_2} + \bar{z}_1 \frac{\partial F_2}{\partial w} = \lambda, \tag{130}$$

$$\frac{\partial G}{\partial z_2} + \bar{z}_1 \frac{\partial G}{\partial w} = \lambda \bar{F}_1(\bar{z}, \bar{w}), \tag{131}$$

for some function λ. Let us take the first equation and plug in the defining equation for M_1,

$$\frac{\partial F_1}{\partial z_2}(z_1, z_2, \bar{z}_1 z_2 + \bar{z}_1^2) + \bar{z}_1 \frac{\partial F_1}{\partial w}(z_1, z_2, \bar{z}_1 z_2 + \bar{z}_1^2) = 0. \tag{132}$$

This is true for all $z \in \mathbb{C}^2$, and so we may treat z_1 and $\bar{z}_1$ as independent variables.

We have an involution on $M_{C.1}$ that takes $\bar{z}_1$ to $-z_2 - \bar{z}_1$. Therefore we also have

$$\frac{\partial F_1}{\partial z_2}(z_1, z_2, \bar{z}_1 z_2 + \bar{z}_1^2) + (-z_2 - \bar{z}_1) \frac{\partial F_1}{\partial w}(z_1, z_2, \bar{z}_1 z_2 + \bar{z}_1^2) = 0. \tag{133}$$

This means that $\partial F_1 / \partial w$ and therefore $\partial F_1 / \partial z_2$ must be identically zero. That is, F_1 only depends on z_1.

We have that the following must hold for all z:

$$G(z_1,z_2,\bar{z}_1z_2+\bar{z}_1^2)=\bar{F}_1(\bar{z}_1)F_2(z_1,z_2,\bar{z}_1z_2+\bar{z}_1^2)+(\bar{F}_1(\bar{z}_1))^2. \tag{134}$$

Again we treat z_1 and $\bar{z}_1$ as independent variables. Differentiate with respect to z_1:

$$\frac{\partial G}{\partial z_1}(z_1,z_2,\bar{z}_1z_2+\bar{z}_1^2)=\bar{F}_1(\bar{z}_1)\frac{\partial F_2}{\partial z_1}(z_1,z_2,\bar{z}_1z_2+\bar{z}_1^2). \tag{135}$$

We plug in the involution again to obtain

$$\frac{\partial G}{\partial z_1}(z_1,z_2,\bar{z}_1z_2+\bar{z}_1^2)=\bar{F}_1(-z_2-\bar{z}_1)\frac{\partial F_2}{\partial z_1}(z_1,z_2,\bar{z}_1z_2+\bar{z}_1^2). \tag{136}$$

Therefore as F_1 is not identically zero, then as before both $\partial F_2/\partial z_1$ and $\partial G/\partial z_1$ must be identically zero. □

Lemma 12.3. *Take $M_{C.1}\subset\mathbb{C}^3$ given by*

$$w=\bar{z}_1z_2+\bar{z}_1^2, \tag{137}$$

and let (F_1,F_2,G) be a local automorphism at the origin. Then F_1 uniquely determines F_2 and G. Furthermore, given any invertible function of one variable F_1 with $F_1(0)=0$, there exist unique F_2 and G that complete an automorphism and they are determined by

$$\begin{aligned}F_2(z_2,\bar{z}_1z_2+\bar{z}_1^2)&=\bar{F}_1(\bar{z}_1)+\bar{F}_1(-\bar{z}_1-z_2),\\ G(z_2,\bar{z}_1z_2+\bar{z}_1^2)&=-\bar{F}_1(\bar{z}_1)\bar{F}_1(-\bar{z}_1-z_2).\end{aligned} \tag{138}$$

We should note that the lemma also works formally. Given any formal F_1, there exist unique formal F_2 and G satisfying the above property.

Proof. By Lemma 12.2, F_1 depends only on z_1 and F_2 and G depend only on z_2 and w. We write the automorphism as a composition of the two mappings $(F_1(z_1),z_2,w)$ and $(z_1,F_2(z_2,w),G(z_2,w))$.

We plug the transformation into the defining equation for $M_{C.1}$ and obtain

$$G(z_2,\bar{z}_1z_2+\bar{z}_1^2)=\bar{F}_1(\bar{z}_1)F_2(z_2,\bar{z}_1z_2+\bar{z}_1^2)+(\bar{F}_1(\bar{z}_1))^2. \tag{139}$$

We use the involution $(z_1,z_2)\mapsto(-\bar{z}_1-z_2,z_2)$ which preserves $M_{C.1}$ and obtain a second equation

$$G(z_2,\bar{z}_1z_2+\bar{z}_1^2)=\bar{F}_1(-\bar{z}_1-z_2)F_2(z_2,\bar{z}_1z_2+\bar{z}_1^2)+(\bar{F}_1(-\bar{z}_1-z_2))^2. \tag{140}$$

We eliminate G and solve for F_2:

$$\begin{aligned}F_2(z_2,\bar{z}_1z_2+\bar{z}_1^2)&=\frac{(\bar{F}_1(-\bar{z}_1-z_2))^2-(\bar{F}_1(\bar{z}_1))^2}{\bar{F}_1(\bar{z}_1)-\bar{F}_1(-\bar{z}_1-z_2)}\\ &=\bar{F}_1(\bar{z}_1)+\bar{F}_1(-\bar{z}_1-z_2).\end{aligned} \tag{141}$$

Next we note that trivially, F_2 is unique if it exists: its difference vanishes on $M_{C.1}$.

If we suppose that F_1 is convergent, then just as before, substituting z_2 with tz_2 and $\bar{z}_1$ with $t\bar{z}_1$, we are restricting to curves $(tz_2, t^2 w)$ for all (z_2, w). The series is convergent in t for every fixed z_2 and w. Therefore if F_2 exists and F_1 is convergent, then F_2 is convergent by Lemma 11.1.

Now we need to show the existence of the formal solution F_2. Notice that the right hand side of (141) is invariant under the involution. It suffices to show that any power series in $\bar{z}_1, z_2$ that is invariant under the involution is a formal power series in z_2 and $\bar{z}_1 z_2 + \bar{z}_1^2$. Let us treat $\xi = \bar{z}_1$ as an independent variable. The original involution becomes a holomorphic involution in ξ, z_2:

$$\tau : \xi \to -\xi - z_2, \quad z_2 \to z_2. \tag{142}$$

By a theorem of Noether we obtain a set of generators for the ring of invariants by applying the averaging operation $R(f) = \frac{1}{2}(f + f \circ \tau)$ to all monomials in ξ and z_2 of degree two or lower. By direct calculation it is not difficult to see that ξ, $\xi z_2 + \xi^2$ generate the ring of invariants. Therefore any invariant power series in z_2, ξ is a power series in ξ, $\xi z_2 + \xi^2$. This shows the existence of F_2. The existence of G follows in the same way.

The equation for $G(z_2, \bar{z}_1 z_2 + \bar{z}_1^2) = -\bar{F}_1(\bar{z}_1)\bar{F}_1(-\bar{z}_1 - z_2)$ is obtained by plugging in the equation for F_2. Its existence, uniqueness, and convergence in the case where F_1 converges, follow exactly the same as for F_2. □

Theorem 12.4. *If $M \subset \mathbb{C}^{n+1}$, $n \geq 2$ is given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2, \tag{143}$$

and $(F_1, F_2, \dots, F_n, G)$ is a local automorphism at the origin, then F_1 depends only on z_1, F_2 and G depend only on z_2 and w, and F_1 completely determines F_2 and G via (138). *The mapping $(z_1, z_2, F_3, \dots, F_n)$ has rank n at the origin.*

Furthermore, given any invertible function F_1 of one variable with $F_1(0) =$ 0, and arbitrary holomorphic functions $F_3, \dots, F_n$ such that $F_j(0) = 0$, and $(z_1, z_2, F_3, \dots, F_n)$ has rank n at the origin, then there exist unique F_2 and G so that $(F_1, \dots, F_n, G)$ is an automorphism.

Proof. Let $(F_1, \dots, F_n, G)$ be an automorphism. Then we have

$$\begin{aligned} &G(z_1, \dots, z_n, w) \\ &\quad = \bar{F}_1(\bar{z}_1, \dots, \bar{z}_n, \bar{w}) F_2(z_1, \dots, z_n, w) + (\bar{F}_1(\bar{z}_1, \dots, \bar{z}_n, \bar{w}))^2. \end{aligned} \tag{144}$$

Proposition 12.1 says that the linear terms in G only depend on w, the linear terms of F_1 depend only on z_1 and w, and the linear terms of F_2 only depend on z_2 and w.

Let us embed $M_{C.1} \subset \mathbb{C}^3$ into M via setting $z_3 = \alpha_3 z_2, \dots, z_n = \alpha_n z_2$, for arbitrary $\alpha_3, \dots, \alpha_n$. Then we obtain

$$\begin{aligned}(145)\quad & G(z_1, z_2, \alpha_3 z_2, \dots, \alpha_n z_2, w)\\ &= \bar{F}_1(\bar{z}_1, \bar{z}_2, \bar{\alpha}_3 \bar{z}_2, \dots, \bar{\alpha}_n \bar{z}_2, \bar{w}) F_2(z_1, z_2, \alpha_3 z_2, \dots, \alpha_n z_2, w)\\ &\qquad + (\bar{F}_1(\bar{z}_1, \bar{z}_2, \bar{\alpha}_3 \bar{z}_2, \dots, \bar{\alpha}_n \bar{z}_2, \bar{w}))^2.\end{aligned}$$

By noting what the linear terms are, we notice that the above is the equation for an automorphism of $M_{C.1}$. Therefore by Lemma 12.2 we have

$$(146)\qquad \frac{\partial F_1}{\partial w} = 0, \quad \frac{\partial F_2}{\partial z_1} = 0, \quad \text{and} \quad \frac{\partial G}{\partial z_1} = 0,$$

as that is true for all $\alpha_3, \dots, \alpha_n$. Plugging in the defining equation for $M_{C.1}$, we obtain an equation that holds for all z and we can treat z and $\bar{z}$ independently. We plug in $z = 0$ to obtain

$$\begin{aligned}(147)\quad 0 &= \bar{F}_1(\bar{z}_1, \bar{z}_2, \bar{\alpha}_3 \bar{z}_2, \dots, \bar{\alpha}_n \bar{z}_2, 0) F_2(0, \dots, 0, \bar{z}_1^2)\\ &\qquad + (\bar{F}_1(\bar{z}_1, \bar{z}_2, \bar{\alpha}_3 \bar{z}_2, \dots, \bar{\alpha}_n \bar{z}_2, 0))^2.\end{aligned}$$

Differentiating with respect to $\bar{\alpha}_j$ we obtain $\partial F_1 / \partial z_j = 0$, for $j = 3, \dots, n$. We set $\bar{\alpha}_j = 0$ in the equation, differentiate with respect to $\bar{z}_2$ and obtain that $\partial F_1 / \partial z_2 = 0$. In other words, F_1 is a function of z_1 only. We rewrite (145) by writing F_1 as a function of z_1 only and F_2 and G as functions of $z_2, \dots, z_n, w$, and we plug in $w = \bar{z}_1 z_2 + \bar{z}_1^2$ to obtain

$$\begin{aligned}(148)\quad & G(z_2, \alpha_3 z_2, \dots, \alpha_n z_2, \bar{z}_1 z_2 + \bar{z}_1^2)\\ &= \bar{F}_1(\bar{z}_1) F_2(z_2, \alpha_3 z_2, \dots, \alpha_n z_2, \bar{z}_1 z_2 + \bar{z}_1^2) + (\bar{F}_1(\bar{z}_1))^2.\end{aligned}$$

By Lemma 12.3, we know that F_1 uniquely determines $F_2(z_2, \alpha_3 z_2, \dots, \alpha_n z_2, w)$ and $G(z_2, \alpha_3 z_2, \dots, \alpha_n z_2, w)$. These two functions therefore do not depend on $\alpha_3, \dots, \alpha_n$, and in turn F_2 and G do not depend on $z_3, \dots, z_n$ as claimed. Furthermore F_1 does uniquely determine F_2 and G.

Finally since the mapping is a biholomorphism, and from what we know about the linear parts of F_1, F_2, and G, it is clear that $(z_1, z_2, F_3, \dots, F_n)$ has rank n.

The other direction follows by applying Lemma 12.3. We start with F_1, determine F_2 and G as in three dimensions. Then add $F_3, \dots, F_n$ and the rank condition guarantees an automorphism. □

13. Normal form for certain type C.1 submanifolds II

The goal of this section is to find the normal form for Levi-flat submanifolds $M \subset \mathbb{C}^{n+1}$ given by

$$(149)\qquad w = \bar{z}_1 z_2 + \bar{z}_1^2 + \operatorname{Re} f(z),$$

for a holomorphic $f(z)$ of order $O(3)$.

Since $f(z)$ can be absorbed into w via a holomorphic transformation, the goal is really to prove:

Theorem 13.1. *Let $M \subset \mathbb{C}^{n+1}$ be a real-analytic Levi-flat given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(\bar{z}), \tag{150}$$

where r is $O(3)$. Then M can be put into the $M_{C.1}$ normal form

$$w = \bar{z}_1 z_2 + \bar{z}_1^2, \tag{151}$$

by a convergent normalizing transformation.

Furthermore, if r is a polynomial and the coefficient of $\bar{z}_1^3$ in r is zero, then there exists an invertible polynomial mapping taking $M_{C.1}$ to M.

In Theorem 1.3, we have already shown that a submanifold of the form

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(\bar{z}_1) \tag{152}$$

is necessarily Levi-flat and has the normal form $M_{C.1}$. The first part of Theorem 13.1 will follow once we prove:

Lemma 13.2. *If $M \subset \mathbb{C}^{n+1}$ is given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(\bar{z}), \tag{153}$$

where r is $O(3)$ and M is Levi-flat, then r depends only on $\bar{z}_1$.

Proof. First let us assume that $n = 2$. For $p \in M_{\mathrm{CR}}$, $T_p^{(1,0)} M$ is one dimensional. The Levi map is the matrix

$$L = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{154}$$

applied to the $T^{(1,0)} M$ vectors. As M is Levi-flat, then the Levi map has to vanish. The only vectors v for which $v^* L v = 0$, are the ones without $\partial/\partial z_1$ component or $\partial/\partial z_2$ component, that is, vectors of the form

$$a \frac{\partial}{\partial z_1} + b \frac{\partial}{\partial w} \quad \text{or} \quad a \frac{\partial}{\partial z_2} + b \frac{\partial}{\partial w}. \tag{155}$$

We apply these vectors to the defining equation and its conjugate and we obtain in the first case the equations

$$b = 0, \quad a \left(\bar{z}_2 + 2 z_1 + \frac{\partial \bar{r}}{\partial z_1} \right) = 0. \tag{156}$$

This cannot be satisfied identically on M since this is supposed to be true for all z, but a cannot be identically zero and the second factor in the second equation has only one nonholomorphic term, which is $\bar{z}_2$.

Let us try the second form and we obtain the equations

$$b = a\bar{z}_1, \quad a\left(\frac{\partial \bar{r}}{\partial z_2}\right) = 0. \tag{157}$$

Again a cannot be identically zero, and hence the second factor of the second equation $\partial\bar{r}/\partial z_2$ must be identically zero, which is possible only if r depends only on $\bar{z}_1$.

Finally, it is possible to pick $b = \bar{z}_1$ and $a = 1$, to obtain a $T^{(1,0)}$ vector field

$$\frac{\partial}{\partial z_2} + \bar{z}_1\frac{\partial}{\partial w}, \tag{158}$$

and therefore these submanifolds are necessarily Levi-flat.

Next suppose that $n > 2$. Notice that replacing z_k with $\lambda_k\xi$ for $k \geq 2$ and then fixing λ_k for $k \geq 2$, we get

$$w = \bar{z}_1\lambda_2\xi + \bar{z}_1^2 + r(\bar{z}_1, \bar{\lambda}_2\bar{\xi}, \dots, \bar{\lambda}_n\bar{\xi}). \tag{159}$$

By Lemma 2.2, we obtain a Levi-flat submanifold in $(z_1, \xi, w) \in \mathbb{C}^3$, and hence can apply the above reasoning to obtain that $r(\bar{z}_1, \bar{\lambda}_2\bar{\xi}, \dots, \bar{\lambda}_n\bar{\xi})$ does not depend on $\bar{\xi}$. As this was true for any λ_k, we have that r can only depend on $\bar{z}_1$. □

It is left to prove the claim about the polynomial normalizing transformation:

Lemma 13.3. *Suppose that $M \subset \mathbb{C}^{n+1}$ is given by*

$$w = \bar{z}_1 z_2 + \bar{z}_1^2 + r(\bar{z}_1), \tag{160}$$

where r is a polynomial that vanishes to the fourth order. Then there exists an invertible polynomial mapping taking $M_{C.1}$ to M.

Proof. We will take a transformation of the form

$$(z_1, z_2, w) \mapsto (z_1, z_2 + f(z_2, w), w + g(z_2, w)). \tag{161}$$

We are therefore trying to find polynomials f and g that satisfy

$$\bar{z}_1 z_2 + \bar{z}_1^2 + g(z_2, \bar{z}_1 z_2 + \bar{z}_1^2) = \bar{z}_1(z_2 + f(z_2, \bar{z}_1 z_2 + \bar{z}_1^2)) + \bar{z}_1^2 + r(\bar{z}_1). \tag{162}$$

If we simplify, we obtain

$$g(z_2, \bar{z}_1 z_2 + \bar{z}_1^2) - \bar{z}_1 f(z_2, \bar{z}_1 z_2 + \bar{z}_1^2) = r(\bar{z}_1). \tag{163}$$

Consider the involution $S\colon (\bar{z}_1, z_2) \to (-\bar{z}_1 - z_2, z_2)$. Its invariant polynomials $u(\bar{z}_1, z_2)$ are precisely the polynomials in $z_2, z_2\bar{z}_1 + \bar{z}_1^2$. The polynomial $r(\bar{z}_1)$ can be uniquely written as

$$r^+(z_2, \bar{z}_1 z_2 + \bar{z}_1^2) + \bigl(\bar{z}_1 + \tfrac{1}{2} z_2\bigr) r^-(z_2, \bar{z}_1 z_2 + \bar{z}_1^2) \tag{164}$$

in two polynomials $r^{\pm}$. Taking $f = -r^-$ and $g = r^+ + \frac{1}{2} z_2 r^-$, we find the desired solutions. □

14. Normal form for general type C.1 submanifolds

In this section we show that generically a Levi-flat type C.1 submanifold is not formally equivalent to the quadric $M_{C.1}$ submanifold. In fact, we find a formal normal form that shows infinitely many invariants. There are obviously infinitely many invariants if we do not impose the Levi-flat condition. The trick therefore is, how to impose the Levi-flat condition and still obtain a formal normal form.

Let $M \subset \mathbb{C}^3$ be a real-analytic Levi-flat type C.1 submanifold through the origin. We know that M is an image of $\mathbb{R}^2 \times \mathbb{C}$ under a real-analytic CR map that is a diffeomorphism onto its target; see Theorem 1.2. After a linear change of coordinates we assume that the mapping is

$$\begin{aligned}(x, y, \xi) \in \mathbb{R}^2 \times \mathbb{C} \mapsto \bigl(x + iy + a(x, y, \xi),\ \xi + b(x, y, \xi),&\\ (x - iy)\xi + (x - iy)^2 + r(x, y, \xi)\bigr),&\end{aligned} \tag{165}$$

where a, b are $O(2)$ and r is $O(3)$. As the mapping is a CR mapping and a local diffeomorphism, then given any such a, b, and r, the image is necessarily Levi-flat at CR points. Therefore the set of all these mappings gives us all type C.1 Levi-flat submanifolds.

We precompose with an automorphism of $\mathbb{R}^2 \times \mathbb{C}$ to make $b = 0$. We cannot similarly remove a as any automorphism must have real-valued first two components (the new x and the new y), and hence those components can only depend on x and y but not on ξ. So if a depends on ξ, we cannot remove it by precomposing.

Next we notice that we can treat M as an abstract CR manifold. Suppose we have two equivalent submanifolds M_1 and M_2, with F being the biholomorphic map taking M_1 to M_2. If M_j is the image of a map φ_j, then note that φ_2^{-1} is CR on $(M_2)_{\mathrm{CR}}$. Therefore, $G = \varphi_2^{-1} \circ F \circ \varphi_1$ is CR on $(F \circ \varphi_1)^{-1}((M_2)_{\mathrm{CR}})$, which is dense in a neighborhood of the origin of $\mathbb{R}^2 \times \mathbb{C}$ (the CR singularity of M_2 is a thin set, and we pull it back by two real-analytic diffeomorphisms). A real-analytic diffeomorphism that is CR on a dense set is a CR mapping. The same argument works for the inverse of G, and therefore we have a CR diffeomorphism of $\mathbb{R}^2 \times \mathbb{C}$. We conclude:

Proposition 14.1. *If $M_j \subset \mathbb{C}^3$, $j = 1, 2$ are given by the maps φ_j*

$$(x, y, \xi) \in \mathbb{R}^2 \times \mathbb{C} \overset{\varphi_j}{\mapsto} (x + iy + a_j(x, y, \xi),\ \xi + b_j(x, y, \xi), \\ (x - iy)\xi + (x - iy)^2 + r_j(x, y, \xi)), \tag{166}$$

and M_1 and M_2 are locally biholomorphically (resp. formally) equivalent at zero, then there exist local biholomorphisms (resp. formal equivalences) F and G at zero, with $F(M_1) = M_2$, $G(\mathbb{R}^2 \times \mathbb{C}) = \mathbb{R}^2 \times \mathbb{C}$ as germs (resp. formally) and

$$\varphi_2 = F \circ \varphi_1 \circ G. \tag{167}$$

In other words, the proposition states that if we find a normal form for the mapping, we find a normal form for the submanifolds. Let us prove that the proposition also works formally.

Proof. We have to prove that G restricted to $\mathbb{R}^2 \times \mathbb{C}$ is CR, that is, $\partial G/\partial \bar{\xi} = 0$. Let us consider

$$\varphi_2 \circ G = F \circ \varphi_1. \tag{168}$$

The right hand side does not depend on $\bar{\xi}$ and thus the left hand side does not either. Write $G = (G^1, G^2, G^3)$. Let us write $b = b_2$ and $r = r_2$ for simplicity. Taking derivative of $\varphi_2 \circ G$ with respect to $\bar{\xi}$ we get

$$\begin{aligned} & G^1_{\bar{\xi}} + iG^2_{\bar{\xi}} + a_x(G)G^1_{\bar{\xi}} + a_y(G)G^2_{\bar{\xi}} + a_\xi(G)G^3_{\bar{\xi}} = 0, \\ & G^3_{\bar{\xi}} + b_x(G)G^1_{\bar{\xi}} + b_y(G)G^2_{\bar{\xi}} + b_\xi(G)G^3_{\bar{\xi}} = 0, \\ & (G^1_{\bar{\xi}} - iG^2_{\bar{\xi}})G^3 + (G^1 - iG^2)G^3_{\bar{\xi}} + 2(G^1 - iG^2)(G^1_{\bar{\xi}} - iG^2_{\bar{\xi}}) \\ & \qquad + r_x(G)G^1_{\bar{\xi}} + r_y(G)G^2_{\bar{\xi}} + r_\xi(G)G^3_{\bar{\xi}} = 0. \end{aligned} \tag{169}$$

Suppose that the homogeneous parts of $G^j_{\bar{\xi}}$ are zero for all degrees up to $d - 1$. If we look at the degree-d homogeneous parts of the first two equations above we immediately note that it must be that $G^1_{\bar{\xi}} + iG^2_{\bar{\xi}} = 0$ and $G^3_{\bar{\xi}} = 0$ in degree d. We then look at the degree-($d + 1$) part of the third equation. Recall that $[\cdot]_d$ is the degree-d part of an expression. We get

$$\left[G^1_{\bar{\xi}} - iG^2_{\bar{\xi}}\right]_d [G^3 + 2G^1 - i2G^2]_1 = 0. \tag{170}$$

As G is an automorphism we cannot have the linear terms be linearly dependent and hence $G^1_{\bar{\xi}} = G^2_{\bar{\xi}} = 0$ in degree d. We finish by induction on d. □

Using the proposition we can restate the result of Theorem 1.3 by parametrization:

Corollary 14.2. *A real-analytic Levi-flat type C.*1 *submanifold $M \subset \mathbb{C}^3$ is biholomorphically equivalent to the quadric $M_{C.1}$ if and only if the mapping giving M is* *equivalent to a mapping of the form*

$$(x,y,\xi)\in\mathbb{R}^2\times\mathbb{C}\mapsto\big(x+iy,\ \xi,\ (x-iy)\xi+(x-iy)^2+r(x,y,\xi)\big). \tag{171}$$

That is, M is equivalent to $M_{C.1}$ if and only if we can get rid of the $a(x,y,\xi)$ via pre and post composing with automorphisms. The proof of the corollary follows as a submanifold that is realized by this map must be of the form $w=\bar z_1z_2+\bar z_1^2+\rho(z_1,\bar z_1,z_2)$ and we apply Theorem 1.3.

We have seen that the involution τ on M, in particular when M is the quadric, is useful to compute the automorphism group and to construct Levi-flat submanifolds of type $C.1$. We will also need to deal with power series in $z,\bar z,\xi$. Thus we extend τ, which is originally defined on $\mathbb{C}^2$, as follows:

$$\sigma(z,\bar z,\xi)=(z,-\bar z-\xi,\xi). \tag{172}$$

Here $z,\bar z,\xi$ are treated as independent variables. Note that $z,\xi,w=\bar z\xi+\bar z^2$ are invariant by σ, while $\eta=\bar z+\frac12\xi$ is skew invariant by σ. A power series in $z,\bar z,\xi$ that is invariant by σ is precisely a power series in z,ξ,w. In general, a power series u in $z,\bar z,\xi$ admits a unique decomposition

$$u(z,\bar z,\xi)=u^+(z,\xi,w)+\eta u^-(z,\xi,w). \tag{173}$$

First we introduce degree for power series $u(z,\bar z,\xi)$ and weights for power series $v(z,\xi,w)$. As usual we assign degree $i+j+k$ to the monomial $z^i\bar z^j\xi^k$. We assign weight $i+j+2k$ to the monomial $z^i\xi^jw^k$. For simplicity, we will call them weight in both situations. Let us also set

$$\begin{aligned}[u]_d(z,\bar z,\xi)&=\sum_{i+j+k=d}u_{ijk}z^i\bar z^j\xi^k,\\ [v]_d(z,\xi,w)&=\sum_{i+j+2k=d}v_{ijk}z^i\xi^jw^k.\end{aligned} \tag{174}$$

Set $[u]_i^j=[u]_i+\cdots+[u]_j$ and $[v]_i^j=[v]_i+\cdots+[v]_j$ for $i\le j$.

Theorem 14.3. *Let M be a real-analytic Levi-flat type $C.1$ submanifold in $\mathbb{C}^3$. There exists a formal biholomorphic map transforming M into the image of*

$$\hat\varphi(z,\bar z,\xi)=(z+A(z,\xi,w)w\eta,\xi,w) \tag{175}$$

with $\eta=\bar z+\frac12\xi$ and $w=\bar z\xi+\bar z^2$. Suppose further that $A\not\equiv0$. Fix i_*,j_*,k_* such that j_* is the largest integer satisfying $A_{i_*j_*k_*}\neq0$ and $i_*+j_*+2k_*=s$. Then we can achieve

$$A_{i_*(j_*+n)k_*}=0,\quad n=1,2,\dots. \tag{176}$$

Furthermore, the power series A is uniquely determined up to the transformation

$$A(z,\xi,w)\to \bar{c}^3A(cz,\bar{c}\xi,\bar{c}^2w),\quad c\in\mathbb{C}\setminus\{0\}. \tag{177}$$

In the above normal form with $A\not\equiv 0$, the group of formal biholomorphisms that preserve the normal form consists of dilations

$$(z,\xi,w)\to(\nu z,\bar{\nu}\xi,\bar{\nu}^2w), \tag{178}$$

satisfying $\bar{\nu}^3A(\nu z,\bar{\nu}\xi,\bar{\nu}^2w)=A(z,\xi,w)$.

Proof. It will be convenient to write the CR diffeomorphism G of $\mathbb{R}^2\times\mathbb{C}$ as (G_1,G_2) where G_1 is complex-valued and depends on $z,\bar{z}$, while G_2 depends on $z,\bar{z},\xi$. Let M be the image of a mapping φ defined by

$$(z,\bar{z},\xi)\overset{\varphi}{\mapsto}\bigl(z+a(z,\bar{z},\xi),\ \xi,\ \bar{z}\xi+\bar{z}^2+r(z,\bar{z},\xi)\bigr) \tag{179}$$

with $a=O(2), r=O(3)$. We want to find a formal biholomorphic map F of $\mathbb{C}^3$ and a formal CR diffeomorphism G of $\mathbb{R}^2\times\mathbb{C}$ such that

$$F\widehat{\varphi}G^{-1}=\varphi \tag{180}$$

with $\widehat{\varphi}$ in the normal form.

To simplify the computation, we will first achieve a preliminary normal form where $r=0$ and the function a is skew-invariant by σ. For the preliminary normal form we will only apply F and G that are tangent to the identity. We will then use the general F and G to obtain the final normal form.

Let us assume that F and G are tangent to the identity. Let $M=F(\widehat{\varphi}(\mathbb{R}^2\times\mathbb{C}))$ where $\widehat{\varphi}$ is determined by $\widehat{a}$, $\widehat{r}$. We write

$$F=I+(f_1,f_2,f_3),\quad G=I+(g_1,g_2). \tag{181}$$

The ξ components in $\varphi G=F\widehat{\varphi}$ give us

$$g_2(z,\bar{z},\xi)=f_2\bigl(z+\widehat{a}(z,\bar{z},\xi),\xi,\bar{z}\xi+\bar{z}^2+\widehat{r}(z,\bar{z},\xi)\bigr). \tag{182}$$

Thus, we are allowed to define g_2 by the above identity for any choice of $f_2=O(2)$. Eliminating g_2 in other components of $\varphi G=F\widehat{\varphi}$, we obtain

$$f_1\circ\widehat{\varphi}-g_1=a\circ G-\widehat{a}, \tag{183}$$

$$f_3\circ\widehat{\varphi}-\bar{z}f_2\circ\widehat{\varphi}=r\circ G-\widehat{r}+2\eta\tilde{g}_1+\tilde{g}_1f_2\circ\widehat{\varphi}+\tilde{g}_1^2, \tag{184}$$

where $\tilde{g}_1(z,\bar{z})=\bar{g}_1(\bar{z},z)$ and

$$(a\circ G)(z,\bar{z},\xi):=a(G_1(z,\bar{z}),\bar{G}_1(\bar{z},z),G_2(z,\bar{z},\xi)). \tag{185}$$

Each power series $r(z,\bar z,\xi)$ admits a unique decomposition

$$r(z,\bar z,\xi)=r^+(z,\xi,w)+\eta r^-(z,\xi,w), \tag{186}$$

where both $r^\pm$ are invariant by σ. Note that $r(z,\bar z,\xi)$ is a power series in z, ξ, and w, if and only if it is invariant by σ, i.e., if $r^-=0$. We write

$$r^+=wt(k)\quad\text{or}\quad wt(r^+)\geq k, \tag{187}$$

if $r^+_{abc}=0$ for $a+b+2c<k$. Define $r^-=wt(k)$ analogously and write $\eta r^-=wt(k)$ if $r^-=wt\,(k-1)$. We write $r=wt(k)$ if $(r^+,\eta r^-)=wt(k)$. Note that

$$r=O(k)\Rightarrow r=wt(k),\quad wt(rs)\geq wt(r)+wt(s). \tag{188}$$

The power series in z and $\bar z$ play a special role in describing normal forms. Let us define $T^\pm$ via

$$u(z,\bar z)=(T^+u)(z,\xi,w)+(T^-u)(z,\xi,w)\eta. \tag{189}$$

Let S_k^+ (resp. S_k^-) be spanned by monomials in $z,\bar z,\xi$ which have weight k and are invariant (resp. skew-invariant) by σ. Then the range of ηT^- in S_k^- is a linear subspace R_k. We decompose

$$S_k^-=R_k\oplus(S_k^-\ominus R_k). \tag{190}$$

The decomposition is of course not unique. We will take

$$S_k^-\ominus R_k=\bigoplus_{\substack{a+b+2c=k-1\\ c>0}}\mathbb{C}z^a\xi^bw^c\eta. \tag{191}$$

Here we have used $\eta=\bar z+\frac12\xi$, $\eta^2=w+\frac14\xi^2$, and

$$T^+u(z,\xi,w)=\sum_{i,j\geq0}\sum_{0\leq\alpha\leq j/2}u_{ij}\binom{j}{2\alpha}z^i\left(w+\tfrac14\xi^2\right)^\alpha\left(-\tfrac12\xi\right)^{j-2\alpha}, \tag{192}$$

$$\begin{aligned}T^-u(z,\xi,w)&\\ &=\sum_{\substack{i\geq0\\ j>0}}\sum_{0\leq\alpha<j/2}u_{ij}\binom{j}{2\alpha+1}z^i\left(w+\tfrac14\xi^2\right)^\alpha\left(-\tfrac12\xi\right)^{j-2\alpha-1}.\end{aligned} \tag{193}$$

In particular, we have

$$T^-u(z,\xi,0)=\sum_{\substack{i\geq0\\ j>0}}(-1)^{j-1}u_{ij}z^i\xi^{j-1}. \tag{194}$$

This shows that

$$T^- u(z,\xi,0) = \frac{1}{-\xi}(u(z,-\xi) - u(z,0)). \tag{195}$$

We are ready to show that under the condition that $g_1(z,\bar z)$ has no pure holomorphic terms, there exists a unique (F,G) which is tangent to the identity such that $\hat r = 0$,

$$\hat a \in \mathcal{N} := \bigoplus \mathcal{N}_k, \quad \mathcal{N}_k := S_k^- \ominus R_k. \tag{196}$$

We start with terms of weight two in (183) and (184) to get

$$[f_1]_2 - [g_1]_2 = [a]_2 - \eta[\hat a^-]_1, \tag{197}$$

$$[f_3]_2 = 0. \tag{198}$$

Note that $f_j^- = 0$. The first identity implies that

$$[f_1]_2 - [T^+ g_1]_2 = [a^+]_2, \quad [T^- g_1]_1 = [\hat a^-]_1 - [a^-]_1. \tag{199}$$

The first equation is solvable with kernel defined by

$$[f_1]_k - [T^+ g_1]_k = 0, \tag{200}$$

for $k = 2$. This shows that $[g_1]_2$ is still arbitrary and we use it to achieve

$$\eta[\hat a^-]_1 \in S_2^- \ominus R_2 = \{0\}. \tag{201}$$

Then the kernel space is defined by (200) and

$$[g_1(z,\bar z) - g_1(z,0)]_k = 0 \tag{202}$$

with $k = 2$. In particular, under the restriction

$$[g_1(z,0)]_k = 0, \tag{203}$$

for $k = 2$, we have achieved $\hat a^- \in \mathcal{N}_2$ by unique $[f_1]_2, [g_1]_2, [f_2]_1, [f_3]_2$. By induction, we verify that if (203) holds for all k, we determine uniquely $[f_1]_k$ and $[g_1]_k$ by normalizing $[\hat a]_k \in \mathcal{N}_k$. We then determine $[f_2]_k$ and $[f_3]_{k+1}$ uniquely to normalize $[\hat r]_{k+1} = 0$. For details, let us find formulae for the solutions. We rewrite (183) as

$$T^- g_1 = -(a \circ G - \hat a - f_1 \circ \hat\varphi)^-, \tag{204}$$

$$(f_1 \circ \hat\varphi)^+ = (a \circ G - \hat a)^+ + T^+ g_1. \tag{205}$$

Using (194), we can solve

$$(-1)^{j-1} g_{1,ij} = -((a \circ G)^-)_{i(j-1)0}, \quad j \geq 1,\ i + j = k. \tag{206}$$

Then we have

$$(\widehat{a}^-)_{ij0}=0, \qquad i+j=k-1; \tag{207}$$

$$(\widehat{a}^-)_{ijm}=((a\circ G-f_1\circ\widehat{\varphi}+g_1)^-)_{ijm}, \quad m\geq 1,\ i+j+m=k-1. \tag{208}$$

Note that $-[g_1]_k(z,-\bar{z})=\bar{z}[(a\circ G-\widehat{a})^-]_{k-1}(z,\bar{z},0)$. We obtain

$$[g_1]_k(z,\bar{z})=\bar{z}[(a\circ G-\widehat{a})^-]_{k-1}(z,-\bar{z},0). \tag{209}$$

Having determined $[g_1]_k$, we take

$$[f_1]_k=[(a\circ G-\widehat{a}+g_1)^+]_k. \tag{210}$$

We then solve (184) by taking

$$[f_2]_k=[E^-]_k, \quad [f_3]_{k+1}=\big[(E-\tfrac{1}{2}\xi f_2)^+\big]_{k+1}, \tag{211}$$

$$E:=r\circ G-\widehat{r}+2\eta\tilde{g}_1+\tilde{g}_1 f_2\circ\widehat{\varphi}+\tilde{g}_1^2. \tag{212}$$

We have achieved the preliminary normalization.

Assume now that

$$\varphi(z,\bar{z},\xi)=(z+a^-(z,\xi,w)\eta,\xi,w), \tag{213}$$

$$\widehat{\varphi}(z,\bar{z},\xi)=(z+\widehat{a}^-(z,\xi,w)\eta,\xi,w) \tag{214}$$

are in the preliminary normal form, i.e.,

$$w\mid a^-(z,\xi,w), \quad w\mid\widehat{a}^-(z,\xi,w). \tag{215}$$

Let us assume that

$$a^-(z,\xi,w)=wt(s), \quad [a^-]_s\not\equiv 0, \quad \widehat{a}^-(z,\xi,w)=wt(s). \tag{216}$$

We assume that $\varphi G=F\widehat{\varphi}$ with

$$F(z,\xi,w)=I+(f_1,f_2,f_3), \tag{217}$$

$$G(z,\bar{z},\xi)=(z+g_1(z,\bar{z}),\xi+g_2(z,\bar{z},\xi)). \tag{218}$$

Here f_i,g_j start with terms of weight and order at least two. In particular, we have

$$f_i=wt(N), \quad g_i=wt(N), \qquad i=1,2; \tag{219}$$

$$f_3=wt(N'), \qquad N'\geq N\geq 2. \tag{220}$$

Set $(P,Q,R):=\varphi G$. Using $N\geq 2$, $s\geq 2$, and Taylor's theorem, we obtain

$$\begin{aligned} P = z + g_1(z,\bar z) + a^-(z,\xi,w)\eta + a^-(z,\xi,w)\big(\bar g_1(\bar z,z) + \tfrac{1}{2} g_2(z,\bar z,\xi)\big) \\ + \eta\nabla a^-(z,\xi,w)\big(g_1(z,\bar z), g_2(z,\bar z,\xi), (\xi+2\bar z)\bar g_1(\bar z,z) + \bar z g_2(z,\bar z,\xi)\big) \\ + wt(s+N+1), \end{aligned} \tag{221}$$

$$Q = \xi + g_2(z,\bar z,\xi), \tag{222}$$

$$R = w + (2\bar z + \xi)\bar g_1(\bar z,z) + \bar z g_2(z,\bar z,\xi) + wt(2N). \tag{223}$$

We also have $(P,Q,R) = F\widehat{\varphi}$. Thus

$$\begin{aligned} P = z + \widehat{a}^-(z,\xi,w)\eta + f_1(z,\xi,w) + \partial_z f_1(z,\xi,w)\widehat{a}^-(z,\xi,w)\eta \\ + wt(N+s+1), \end{aligned} \tag{224}$$

$$Q = \xi + f_2(z,\xi,w) + \partial_z f_2(z,\xi,w)\widehat{a}^-(z,\xi,w)\eta + wt(N+s+1), \tag{225}$$

$$R = w + f_3(z,\xi,w) + \partial_z f_3(z,\xi,w)\widehat{a}^-(z,\xi,w)\eta + wt(N'+s+1). \tag{226}$$

We will use the above six identities for P, Q, R in two ways. First we use their lower order terms to get

$$f_1(z,\xi,w) = g_1(z,\bar z) + (a^-(z,\xi,w) - \widehat{a}^-(z,\xi,w))\eta + wt(N+s), \tag{227}$$

$$f_2(z,\xi,w) = g_2(z,\bar z,\xi) + wt(N+s), \tag{228}$$

$$f_3(z,\xi,w) = (2\bar z + \xi)\bar g_1(\bar z,z) + \bar z g_2(z,\bar z,\xi) + wt(2N) + wt(N'+s). \tag{229}$$

Hence, we can take $N' = N+1$. By (227) and the preliminary normalization, we first know that

$$\widehat{a} = a + wt(N+s-1), \tag{230}$$

$$f_1(z,\xi,w) = b(z) + wt(N+s), \quad g_1(z,\bar z) = b(z) + wt(N+s). \tag{231}$$

We compose (229) by σ and then take the difference of the two equations to get

$$f_2(z,\xi,w) = -\bar b(\bar z) - \bar b(-\bar z - \xi) + wt(2N-1) + wt(N+s), \tag{232}$$

$$f_3(z,\xi,w) = -\bar z\bar b(-\bar z - \xi) + (\bar z + \xi)\bar b(\bar z) + wt(2N) + wt(N+s+1). \tag{233}$$

Here we have used $N' = N+1$. Let $b(z) = b_N z^N + wt(N+1)$. Therefore, we have

$$g_2(z,\bar z,\xi) = -\bar b_N\big(\bar z^N + (-\bar z - \xi)^N\big) + wt(N+1), \tag{234}$$

$$\bar g_1(\bar z,z) + \tfrac{1}{2} g_2(z,\bar z,\xi) = \eta\bar b_N \sum \bar z^i(-\bar z - \xi)^{N-1-i} + wt(N+1), \tag{235}$$

$$(2\bar z+\xi)\bar g_1(\bar z,z) + \bar z g_2(z,\bar z,\xi) = \bar b_N(\bar z^{N-1} + (-\bar z - \xi)^{N-1})w + wt(N+2). \tag{236}$$

Next, we use the two formulae for P and (231) to get the identity in higher weight,

$$\widehat{a}^- = a^- + g_1^- + Lb_N + wt(N+s), \quad f_1 - g_1^+ = wt(N+s+1). \tag{237}$$

Here we have used $f_1^- = 0$ and

$$\begin{aligned}(238)\quad & Lb_N(z,\xi,w) := \\ & -Nb_N z^{N-1}[a^-]_s(z,\xi,w) - [a^-]_s(z,\xi,w)\bar{b}_N \sum_i \bar{z}^i(-\bar{z}-\xi)^{N-1-i} \\ & \quad + \nabla[a^-]_s \cdot \big(b_N z^N, -\bar{b}_N(\bar{z}^N + (-\bar{z}-\xi)^N), \bar{b}_N w(\bar{z}^{N-1} + (-\bar{z}-\xi)^{N-1})\big).\end{aligned}$$

Recall that $w \mid a^-$ and $w \mid \widehat{a}^-$. We also have that $w \mid Lb_N(z,\xi,w)$ and Lb_N is homogenous in weighted variables and of weight $N+s-1$. This shows that $[g_1^-(z,\xi,0)]_{N+s-1} = 0$. By (194), we get

$$(239)\qquad [g_1(z,\bar{z})]_{N+s} = [g_1(z,0)]_{N+s}, \quad [\widehat{a}^-]_{s+N-1} = [a^-]_{s+N-1} + Lb_N.$$

Let us make some observations. First, Lb_N depends only on b_N and it does not depend on coefficients of $b(z)$ of degree larger than N. We observe that the first identity says that all coefficients of $[g_1]_{N+s}$ must be zero, except that the coefficient $g_{1,(N+s)0}$ is arbitrary. On the other hand Lb_N, which has weight $N+s-1$, depends only on $g_{1,N0}$, while $N+s-1 > N$. Let us assume for the moment that we have $Lb_N \neq 0$ for all $b_N \neq 0$. We will then choose a suitable complement subspace $\mathcal{N}^*_{N+s-1}$ in the space of weighted homogenous polynomials in z, ξ, w of weight $N+s-1$ for Lb_N. Then $\widehat{a}^- \in w\sum_{N>1}\mathcal{N}^*_{N+s-1}$ will be the required normal form. The normal form will be obtained by the following procedures: Assume that φ is not formally equivalent to the quadratic mapping in the preliminary normalization. We first achieve the preliminary normal form by a mapping $F^0 = I + (f_1^0, f_2^0, f_3^0)$ and $G^0 = I + (g_1^0, g_2^0)$ which are tangent to the identity. We can make F^0 and G^0 to be unique by requiring $f_1^1(z,0) = 0$. Then a is normalized such that $\widehat{a} = \widehat{a}^-\eta$ with $[\widehat{a}^-]_s$ being nonzero homogenous part of the lowest weight. We may assume that $[a]_{s+1} = [\widehat{a}]_{s+1}$. Inductively, we choose $f^1_{1,N00}$ $(N = 2, 3, \dots)$ to achieve $[\widehat{a}^-]_{N+s-1} \in w\mathcal{N}^*_{N+s-1}$. In this step for a given N, we determine mappings $F^1 = I + (f_1^1, f_2^1, f_3^1)$ and $G^1 = I + (g_1^1, g_2^1)$ by requiring that $f_1^1(z,\xi,w)$ contains only one term ξ^N, while f_1^1, f_2^1, g_1^1, g_2^1 have weight at most N and f_3^1 has weight at most $N+1$. In the process, we also show that $[f_1^1(z,\xi,w)]_2^{N+s}$ depends only on z, if we do not want to impose the restriction on f_1^1. Moreover, the coefficient of ξ^{N+s-1} of f_1^1 can still be arbitrarily chosen without changing the normalization achieved for $[\widehat{a}^-]_{N+s-1}$ via $[f_1^1]_N$. However, by achieving $[\widehat{a}^-]_{N+s-1} \in w\mathcal{N}^*_{N+s-1}$ via F^1 and G^1, we may destroy the preliminary normalization achieved via F_0 and G_0. We will then restore the preliminary normalization via $F^2 = I + (f_1^2, f_2^2, f_3^2), G^2 = I + (g_1^2, g_2^2)$ satisfying $g_1^2(z,0) = 0$. This amounts to determining $g_1^2 = g_1$ and $f_1^2 = f_1$ via (204) and (205) for which the terms of weight at most $N+s$ have been determined by (237), and then $f_2^2 = f_2$, $f_3^2 = f_3$, $g_2^2 = g_2$ are determined by (211), (212), and (182), respectively. This allows us to repeat the procedure to achieve the normalization in any higher weight.

We will then remove the restriction that the normalizing mappings must be tangent to the identity. This will alter the normal form only by suitable linear dilations.

Suppose that $b_N \neq 0$. Let us verify that

$$Lb_N \neq 0. \tag{240}$$

We will also identify one of nonzero coefficients to describe the normalizing condition on $\hat{a}$. We write the two invariant polynomials

$$\bar{z}^N + (-\bar{z}-\xi)^N = \lambda_N \xi^N + \sum_{j<N} p_{ijk} z^i \xi^j w^k, \tag{241}$$

$$\sum_i \bar{z}^i (-\bar{z}-\xi)^{N-1-i} = \lambda'_{N-1} \xi^{N-1} + \sum_{j<N-1} q_{ijk} z^i \xi^j w^k. \tag{242}$$

If we plug in $w = \bar{z}^2 + \bar{z}\xi$, we obtain two polynomial identities in the variables z, $\bar{z}$, and ξ:

$$\bar{z}^N + (-\bar{z}-\xi)^N = \lambda_N \xi^N + \sum_{j<N} p_{ijk} z^i \xi^j (\bar{z}^2 + \bar{z}\xi)^k, \tag{243}$$

$$\sum_i \bar{z}^i (-\bar{z}-\xi)^{N-1-i} = \lambda'_{N-1} \xi^{N-1} + \sum_{j<N-1} q_{ijk} z^i \xi^j (\bar{z}^2 + \bar{z}\xi)^k. \tag{244}$$

If we set $\bar{z} = z = 0$, we obtain that

$$\lambda_N = \lambda'_N = (-1)^N. \tag{245}$$

Recall that j_* is the largest integer such that $(a^-)_{i_* j_* k_*} \neq 0$ and $i_* + j_* + 2k_* = s$. Since $w \mid [a^-]_s$, then $k_* > 0$. We obtain

$$(Lb_N)_{i_*(j_*+N-1)k_*} = (a^-)_{i_* j_* k_*} \bar{b}_N (-\lambda'_{N-1} - j_* \lambda_{N-1} + k_* \lambda_N) \neq 0. \tag{246}$$

Therefore, we can achieve

$$(\hat{a}^-)_{i_*(j_*+n)k_*} = 0, \quad n = 1, 2, \ldots. \tag{247}$$

This determines uniquely all $b_2, b_3, \ldots$.

We now remove the restriction that F and G are tangent to the identity. Suppose that both φ and $\hat{\varphi}$ are in the normal form. Suppose that $F\varphi = \hat{\varphi}G$. Then looking at the quadratic terms, we know that the linear parts of F and G must be dilations. In fact, the linear part of F must be the linear automorphism of the quadric. Thus the linear parts of F and G have the forms

$$G'\colon (z,\xi) = (\nu z, \bar{\nu}\xi), \quad F'(z,\xi,w) = (\nu z, \bar{\nu}\xi, \bar{\nu}^2 w). \tag{248}$$

Then $(F')^{-1}\hat{\varphi}G'$ is still in the normal form. Since $(F')^{-1}F$ is holomorphic and $(G')^{-1}G$ is CR, by the uniqueness of the normalization, we know that $F' = F$ and

$G' = G$. Therefore, F and G change the normal form a^- as follows,

$$a^-(z,\xi,w) = \bar{\nu}\hat{a}^-(\nu z, \bar{\nu}\xi, \bar{\nu}^2 w), \quad \nu \in \mathbb{C}\setminus\{0\}. \tag{249}$$

When $[\hat{a}^-]_s = [a^-]_s \neq 0$, we see that $|\nu| = 1$. Therefore, the formal automorphism group is discrete or one-dimensional. □

Coffman [2006] used an analogous method of even/odd function decomposition to obtain a quadratic normal form for non Levi-flat real-analytic m-submanifolds in $\mathbb{C}^n$ with a CR singularity satisfying certain nondegeneracy conditions, provided $\frac{2}{3}(n+1) \leq m < n$. He was able to achieve the convergent normalization by a rapid iteration method. Using the above decomposition of invariant and skew-invariant functions of the involution σ, one might achieve a convergent solution for approximate equations when M is formally equivalent to the quadric. However, when the iteration is employed, each new CR mapping $\hat{\varphi}$ might only be defined on a domain that is proportional to that of the previous φ by a *constant* factor. This is significantly different from the situations of [Moser 1985; Coffman 2006; 2010], where rapid iteration methods are applicable. Therefore, even if M is formally equivalent to the quadric, we do not know if they are holomorphically equivalent.

15. Instability of Bishop-like submanifolds

Let us now discuss stability of Levi-flat submanifolds under small perturbations that keep the submanifolds Levi-flat, in particular, we discuss which quadratic invariants are stable when moving from point to point on the submanifold. The only stable submanifolds are A.n and C.1. The Bishop-like submanifolds (or even just the Bishop invariant) are not stable under perturbation, which we show by constructing examples.

Proposition 15.1. *Suppose that $M \subset \mathbb{C}^{n+1}$, $n \geq 2$, is a connected real-analytic real codimension-two submanifold that has a nondegenerate CR singular at the origin. M can be written in coordinates $(z,w) \in \mathbb{C}^n \times \mathbb{C}$ as*

$$w = A(z,\bar{z}) + B(\bar{z},\bar{z}) + O(3), \tag{250}$$

for quadratic A and B. In a neighborhood of the origin all complex tangents of M are nondegenerate, while ranks of A and B are upper semicontinuous. Suppose that M is Levi-flat (that is, M_{CR} is Levi-flat). The CR singular set of M that is not of type B.$\frac{1}{2}$ at the origin is a real-analytic subset of M of codimension at least two, while the CR singular set of M that is of type B.$\frac{1}{2}$ at the origin has codimension at least one. A.n has an isolated CR singular point at the origin and so does C.1 in $\mathbb{C}^3$. Let $S_0 \subset M$ be the set of CR singular points. There is a neighborhood U of the origin such that for $S = S_0 \cap U$ we have:

(i) *If M is of type A.k for $k \geq 2$ at the origin, then it is of type A.j at each point of S for some $j \geq k$.*

(ii) *If M is of type C.1 at the origin, then it is of type C.1 on S. If M is of type C.0 at the origin, then it is of type C.0 or C.1 on S.*

(iii) *There exists an M that is of type B.γ at one point and of C.1 at CR singular points arbitrarily near. Similarly there exists an M of type A.1 at $p \in M$ that is either of type C.1, or B.γ, at points arbitrarily near p. There also exists an M of type B.γ at every point but where γ varies from point to point.*

Proof. First we show that the rank of A and the rank of B are lower semicontinuous on S_0, without imposing the Levi-flatness condition. Similarly the real dimension of the range of $A(z,\bar{z})$ is lower semicontinuous on S_0. Write M as

$$w = \rho(z,\bar{z}), \tag{251}$$

where ρ vanishes to second order at zero. If we move to a different point of S_0 via an affine map $(z,w) \mapsto (Z+z_0, W+w_0)$. Then we have

$$W + w_0 = \rho(Z+z_0, \bar{Z}+\bar{z}_0). \tag{252}$$

We compute the Taylor coefficients

$$\begin{aligned} W = \frac{\partial\rho}{\partial z}(z_0,\bar{z}_0)\cdot Z + \frac{\partial\rho}{\partial\bar{z}}(z_0,\bar{z}_0)\cdot\bar{Z} + Z^*\left[\frac{\partial^2\rho}{\partial z\partial\bar{z}}(z_0,\bar{z}_0)\right]Z \\ + \tfrac{1}{2}Z^t\left[\frac{\partial^2\rho}{\partial z\partial z}(z_0,\bar{z}_0)\right]Z + \tfrac{1}{2}Z^*\left[\frac{\partial^2\rho}{\partial\bar{z}\partial\bar{z}}(z_0,\bar{z}_0)\right]\bar{Z} + O(3). \end{aligned} \tag{253}$$

The holomorphic terms can be absorbed into W. If $(\partial\rho/\partial\bar{z})(z_0,\bar{z}_0)\cdot\bar{Z}$ is nonzero, then this complex defining function has a linear term in W and a linear term in $\bar{Z}$ and the submanifold is CR at this point. Therefore the set of complex tangents of M is defined by

$$\frac{\partial\rho}{\partial\bar{z}} = 0, \tag{254}$$

and each complex tangent point is nondegenerate. At a complex tangent point at the origin, A is given by $[(\partial^2\rho/\partial z\partial\bar{z})(z_0,\bar{z}_0)]$ and B is given by $\frac{1}{2}[(\partial^2\rho/\partial\bar{z}\partial\bar{z})(z_0,\bar{z}_0)]$. These matrices change continuously as we move along S. We first conclude that all CR singular points of M in a neighborhood of the origin are nondegenerate. Further holomorphic transformations act on A and B using Proposition 2.1. Thus the ranks of A and B and the real dimension of the range of $A(z,\bar{z})$ are lower semicontinuous on S_0 as claimed. Furthermore as M is real-analytic, the points where the rank drops lie on a real-analytic subvariety of S_0, or in other words a thin set. Let U be a small enough neighborhood of the origin so that $S = S_0 \cap U$ is connected.

Imposing the condition that M is Levi-flat, we apply Theorem 1.1. By a simple computation, unless M is of type B.$\frac{1}{2}$, the set of complex tangents of M has codimension at least two; and A.n has an isolated CR singular point and so does

C.1 in $\mathbb{C}^3$. Item (i) follows as A.k are the only types where the rank of B is greater than one, and the theorem says M must be one of these types. For (ii), note that since A is of rank one when M is C.x at a point, M cannot be of type A.k nearby. If M is of type C.1 at a point then the range of A must be of real dimension two in a neighborhood, and hence on this neighborhood M cannot be of type B.γ.

Examples proving (iii) are given below. □

Example 15.2. Define M via

$$w = |z_1|^2 + \gamma \bar{z}_1^2 + \bar{z}_1 z_2 z_3. \tag{255}$$

It is Levi-flat by Proposition 6.2. At the origin M is of type B.γ, but at a point where $z_1 = z_2 = 0$ and $z_3 \neq 0$, the submanifold is CR singular and it is of type C.1.

Example 15.3. Similarly we obtain a CR singular Levi-flat M via

$$w = \bar{z}_1^2 + \bar{z}_1 z_2 z_3; \tag{256}$$

this M is A.1 at the origin, but C.1 at nearby CR singular points.

Example 15.4. If we define M via

$$w = \gamma \bar{z}_1^2 + |z_1|^2 z_2, \tag{257}$$

then M is a CR singular Levi-flat type A.1 submanifold at the origin, but type B.γ at points where $z_1 = 0$ but $z_2 \neq 0$.

Example 15.5. The Bishop invariant can vary from point to point. Define M via

$$w = |z_1|^2 + \bar{z}_1^2(\gamma_1(1 - z_2) + \gamma_2 z_2), \tag{258}$$

where $\gamma_1, \gamma_2 \geq 0$. It is not hard to see that M is Levi-flat. Again it is an image of $\mathbb{C}^2 \times \mathbb{R}^2$ in a similar way as above.

At the origin, the submanifold is Bishop-like with Bishop invariant γ_1. When $z_1 = 0$ and $z_2 = 1$, the Bishop invariant is γ_2. In fact when $z_1 = 0$, the Bishop invariant at that point is $|\gamma_1(1 - z_2) + \gamma_2 z_2|$.

Proposition 6.2 says that this submanifold possesses a real-analytic foliation extending the Levi foliation through the singular points. Proposition 6.1 says that if a foliation on M extends to a (nonsingular) holomorphic foliation, then the submanifold would be a simple product of a Bishop submanifold and $\mathbb{C}$. Therefore, if $\gamma_1 \neq \gamma_2$ then the Levi foliation on M cannot extend to a holomorphic foliation of a neighborhood of M.

References

[Baouendi et al. 1999] M. S. Baouendi, P. Ebenfelt, and L. P. Rothschild, *Real submanifolds in complex space and their mappings*, Princeton Mathematical Series **47**, Princeton University Press, 1999. MR 2000b:32066 Zbl 0944.32040

[Bedford 1977] E. Bedford, "Holomorphic continuation of smooth functions over Levi-flat hypersurfaces", *Trans. Amer. Math. Soc.* **232** (1977), 323–341. MR 58 #1246 Zbl 0382.32009

[Bishop 1965] E. Bishop, "Differentiable manifolds in complex Euclidean space", *Duke Math. J.* **32** (1965), 1–21. MR 34 #369 Zbl 0154.08501

[Brunella 2007] M. Brunella, "Singular Levi-flat hypersurfaces and codimension one foliations", *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **6**:4 (2007), 661–672. MR 2009c:32065 Zbl 1214.32012 arXiv math/0701607

[Burcea 2013] V. Burcea, "A normal form for a real 2-codimensional submanifold in $\mathbb{C}^{N+1}$ near a CR singularity", *Adv. Math.* **243** (2013), 262–295. MR 3062747 Zbl 1279.32024 arXiv 1110.1118

[Burns and Gong 1999] D. Burns and X. Gong, "Singular Levi-flat real analytic hypersurfaces", *Amer. J. Math.* **121**:1 (1999), 23–53. MR 2000j:32062 Zbl 0931.32009

[Cerveau and Lins Neto 2011] D. Cerveau and A. Lins Neto, "Local Levi-flat hypersurfaces invariants by a codimension one holomorphic foliation", *Amer. J. Math.* **133**:3 (2011), 677–716. MR 2012e:32052 Zbl 1225.32038

[Coffman 2006] A. Coffman, "Analytic stability of the CR cross-cap", *Pacific J. Math.* **226**:2 (2006), 221–258. MR 2007j:32038 Zbl 1123.32018

[Coffman 2009] A. Coffman, "CR singularities of real fourfolds in $\mathbb{C}^3$", *Illinois J. Math.* **53**:3 (2009), 939–981. MR 2011m:32067 Zbl 1233.32027

[Coffman 2010] A. Coffman, *Unfolding CR singularities*, Memoirs of the American Mathematical Society **205:**962, American Mathematical Society, Providence, RI, 2010. MR 2011f:32077 Zbl 1194.32016

[Diederich and Fornaess 1978] K. Diederich and J. E. Fornaess, "Pseudoconvex domains with real-analytic boundary", *Ann. Math.* (2) **107**:2 (1978), 371–384. MR 57 #16696 Zbl 0378.32014

[Dolbeault et al. 2005] P. Dolbeault, G. Tomassini, and D. Zaitsev, "On boundaries of Levi-flat hypersurfaces in $\mathbb{C}^n$", *C. R. Math. Acad. Sci. Paris* **341**:6 (2005), 343–348. MR 2006e:32048 Zbl 1085.32019

[Dolbeault et al. 2011] P. Dolbeault, G. Tomassini, and D. Zaitsev, "Boundary problem for Levi flat graphs", *Indiana Univ. Math. J.* **60**:1 (2011), 161–170. MR 2952414 Zbl 1244.32019

[Ebeling 2007] W. Ebeling, *Functions of several complex variables and their singularities*, Graduate Studies in Mathematics **83**, American Mathematical Society, Providence, RI, 2007. MR 2008c:32001 Zbl 1188.32001

[Fernández-Pérez 2013] A. Fernández-Pérez, "On Levi-flat hypersurfaces with generic real singular set", *J. Geom. Anal.* **23**:4 (2013), 2020–2033. MR 3107688 Zbl 1277.32042

[Garrity 2000] T. Garrity, "Global structures on CR manifolds via Nash blow-ups", *Michigan Math. J.* **48** (2000), 281–294. MR 2001h:32058 Zbl 0995.32023

[Gong 1994a] X. Gong, "Normal forms of real surfaces under unimodular transformations near elliptic complex tangents", *Duke Math. J.* **74**:1 (1994), 145–157. MR 95c:32014 Zbl 0834.32003

[Gong 1994b] X. Gong, "On the convergence of normalizations of real analytic surfaces near hyperbolic complex tangents", *Comment. Math. Helv.* **69**:4 (1994), 549–574. MR 95j:32027 Zbl 0826.32012

[Gong 2004] X. Gong, "Existence of real analytic surfaces with hyperbolic complex tangent that are formally but not holomorphically equivalent to quadrics", *Indiana Univ. Math. J.* **53**:1 (2004), 83–95. MR 2005a:32042 Zbl 1060.32020

[Huang 1998] X. Huang, "On an n-manifold in $\mathbb{C}^n$ near an elliptic complex tangent", *J. Amer. Math. Soc.* **11**:3 (1998), 669–692. MR 98m:32026 Zbl 0904.32016

[Huang and Krantz 1995] X. Huang and S. G. Krantz, “On a problem of Moser”, *Duke Math. J.* **78**:1 (1995), 213–228. MR 96f:32026 Zbl 0846.32010

[Huang and Yin 2009a] X. Huang and W. Yin, “A Bishop surface with a vanishing Bishop invariant”, *Invent. Math.* **176**:3 (2009), 461–520. MR 2010f:32039 Zbl 1171.53045

[Huang and Yin 2009b] X. Huang and W. Yin, “A codimension two CR singular submanifold that is formally equivalent to a symmetric quadric”, *Int. Math. Res. Not.* **2009**:15 (2009), 2789–2828. MR 2010f:32038 Zbl 1182.32014

[Huang and Yin 2012] X. Huang and W. Yin, “Flattening of CR singular points and analyticity of local hull of holomorphy”, preprint, 2012. arXiv 1210.5146

[Kenig and Webster 1982] C. E. Kenig and S. M. Webster, “The local hull of holomorphy of a surface in the space of two complex variables”, *Invent. Math.* **67**:1 (1982), 1–21. MR 84c:32014 Zbl 0489.32007

[Kenig and Webster 1984] C. E. Kenig and S. M. Webster, “On the hull of holomorphy of an n-manifold in $\mathbb{C}^n$”, *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (4) **11**:2 (1984), 261–280. MR 86d:32019 Zbl 0558.32006

[Kohn 1979] J. J. Kohn, “Subellipticity of the $\bar\partial$-Neumann problem on pseudo-convex domains: sufficient conditions”, *Acta Math.* **142**:1-2 (1979), 79–122. MR 80d:32020 Zbl 0395.35069

[Lebl 2013] J. Lebl, “Singular set of a Levi-flat hypersurface is Levi-flat”, *Math. Ann.* **355**:3 (2013), 1177–1199. MR 3020158 Zbl 06149475 arXiv 1012.5993

[Lebl et al. 2014] J. Lebl, A. Minor, R. Shroff, D. Son, and Y. Zhang, “CR singular images of generic submanifolds under holomorphic maps”, *Ark. Mat.* **52**:2 (2014), 301–327. MR 3255142 arXiv 1205.5309

[Moser 1985] J. K. Moser, “Analytic surfaces in $\mathbb{C}^2$ and their local hull of holomorphy”, *Ann. Acad. Sci. Fenn. Ser. A I Math.* **10** (1985), 397–410. MR 87c:32024 Zbl 0585.32007

[Moser and Webster 1983] J. K. Moser and S. M. Webster, “Normal forms for real surfaces in $\mathbb{C}^2$ near complex tangents and hyperbolic surface transformations”, *Acta Math.* **150**:3-4 (1983), 255–296. MR 85c:32034 Zbl 0519.32015

[Whitney 1972] H. Whitney, *Complex analytic varieties*, Addison-Wesley, Reading, MA, 1972. MR 52 #8473 Zbl 0265.32008

Received March 3, 2014.

Xianghong Gong
Department of Mathematics
University of Wisconsin - Madison
Madison, WI 53706-1388
United States

gong@math.wisc.edu

Jiří Lebl
Department of Mathematics
Oklahoma State University
Stillwater, OK 74078
United States

lebl@math.okstate.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.167

MEASUREMENTS OF RIEMANNIAN TWO-DISKS AND TWO-SPHERES

FLORENT BALACHEFF

We prove that any Riemannian two-sphere having area at most 1 can be continuously mapped onto a tree in such a way that the topology of the fibers is controlled and their length is less than 7.6. This result improves previous estimates and relies on a similar statement for Riemannian two-disks.

1. Introduction

In this article we are interested to describe the possible geometries of Riemannian two-disks and two-spheres in the same way a tailor determines the geometry of a body: by taking some relevant measurements. We denote by $A(\cdot)$ the area functional and $|\cdot|$ the length functional. Our main result deals with measurements of two-disks:

Theorem 1.1. *If D is a Riemannian two-disk, then for any $\epsilon > 0$ we can find a continuous map to a trivalent tree such that the preimage of a terminal vertex is either an interior point or the boundary ∂D, the preimage of an interior point of an edge is homeomorphic to a circle, the preimage of a trivalent vertex is homeomorphic to the θ figure, and fibers have length at most*

$$(1+\epsilon)\max\bigl\{|\partial D| + \sqrt{A(D)},\ (4 + 11\sqrt{3}/4)\sqrt{A(D)}\bigr\}.$$

This theorem should be compared to a result by Y. Liokumovich [2014] which states that any Riemannian two-disk D admits a Morse function $f : D \to \mathbb{R}$ which is constant on the boundary and whose fibers have length at most $52\sqrt{A(D)} + |\partial D|$.

Using Theorem 1.1, we are able to estimate the measurements of two-spheres in terms of their area.

Theorem 1.2. *If M is a Riemannian two-sphere with area less than 1, then it admits a continuous map to a trivalent tree such that the preimage of a terminal vertex is a point, the preimage of an interior point of an edge is homeomorphic to a circle, the preimage of a trivalent vertex is homeomorphic to the θ figure and fibers have length at most $2\sqrt{3} + 33/8 \simeq 7.6$.*

MSC2010: 53C23.

Keywords: Curvature-free inequalities, Bers constant, closed geodesic, isoperimetric inequalities, width.

This improves a previous estimate by Liokumovich [2014] proving such a result with $8\sqrt{3}+12 \simeq 26$ as upper bound on the length of the fibers. Note nevertheless that the main result of Liokumovich is stronger: he proved the existence of a Morse function $f : M \to \mathbb{R}$ whose fibers have length at most 52. Also note that L. Guth [2005] proved the existence of maps such as in Theorem 1.2 with the upper bound $120/(2\sqrt{\pi}) \simeq 34$ on the length of the fibers under the weaker assumption that the 1-hypersphericity is less than $1/(2\sqrt{\pi})$. Finally we point out that the constant 7.6 in Theorem 1.2 is within a factor at most 6 from the optimal one; see Remark 2.8.

The interest in obtaining precise measurements for Riemannian two-spheres is illustrated by the fact that we can derive upper bounds on the shortest length of a closed geodesic, on the shortest length of a simple loop dividing the sphere into two subdisks of area at least $A/3$, and on the maximal length of a shortest pants decomposition for punctured spheres. More precisely, we are first able to recover the result of C. Croke [1988] on the existence of short closed geodesics for Riemannian two-spheres: we will deduce from Theorem 1.2 that any Riemannian two-sphere with unit area carries a closed geodesic of length at most $\simeq 10.1$; see Theorem 2.7. This is not as good as the current best constant, due to R. Rotman [2006] and equal to $4\sqrt{2} \simeq 5.7$, but it is not too far from it. Moreover, using Theorem 1.2, we can also recover Theorem VI of [Alvarez Paiva et al. 2013] on the existence of a short closed geodesic for Finsler (eventually nonreversible) two-spheres. The precise statement is the following.

Theorem 1.3. *Let M be a Finsler (eventually nonreversible) two-sphere with Holmes–Thompson area less than* 1. *Then it carries a closed geodesic of length at most* $2\sqrt{\pi}\,(11\sqrt{3}+16) \simeq 31.1$.

This improves the current best constant, due to Liokumovich [2014, Theorem 4]. We also easily deduce from Theorem 1.2 the following result, which also improves one of Liokumovich [2014, Theorem 1].

Theorem 1.4. *Let M be a Riemannian two-sphere. Then there exists a simple loop of length at most* $(2\sqrt{3}+33/8)\sqrt{A(M)}$ *dividing M into two subdisks of area at least $A(M)/3$.*

Finally, it is straightforward to see that Theorem 1.2 implies the following.

Theorem 1.5. *Let M be a Riemannian punctured two-sphere with area less than* 1. *Then there exists a decomposition of M into* 3*-holed spheres such that each boundary curve has length at most* $2\sqrt{3}+33/8 \simeq 7.6$.

This improves the current best bound, even for hyperbolic metrics; compare with [Balacheff and Parlier 2012].

The paper is organized as follows. The first section presents Besicovich's lemma and some of its useful corollaries: Papasoglu's lemma [2009] and the disk subdivision lemma of Liokumovich, A. Nabutovsky and Rotman [Liokumovich et al. 2014]. We also define an invariant called the θ-width, reformulate Theorem 1.1 in terms of this invariant, and show how to prove Theorems 1.2 and 1.3 from Theorem 1.1. In the second section, we prove Theorem 1.1. Our strategy is inspired by the proof of [Liokumovich et al. 2014, Theorem 1.6] where it is shown that the boundary of Riemannian two-disks with uniformly bounded diameter and area can always be contracted through closed curves of bounded length. We show that it is enough to consider the case where the length of the boundary is short in comparison with the area. This step is performed using Besicovich's lemma. Then we use the disk subdivision lemma to argue by induction on the area.

2. Preliminaries

As we deal only with surfaces, we will use the terms of disk and sphere for two-disk and two-sphere. We denote by $A(\cdot)$ the area functional and $|\cdot|$ the length functional.

Besicovich's lemma and consequences. In order to prove our results, we will use the following fundamental result of metric geometry as well as some of its consequences.

Lemma 2.1 [Besicovitch 1952]. *Let $\mathcal{S}$ be a Riemannian square. Then there exists a simple geodesic path connecting two opposites sides of length at most $\sqrt{A(\mathcal{S})}$.*

In particular, any Riemannian disk D whose boundary satisfies $|\partial D| > 4\sqrt{A(D)}$ can be subdivided into two subdisks of smaller perimeters (divide its boundary into four equal parts and apply Besicovich's lemma).

P. Papasoglu used Besicovich's lemma to derive the following estimate.

Lemma 2.2 [Papasoglu 2009]. *Let M be a Riemannian two-sphere. For any $\delta > 0$ there exists a simple closed curve of length at most $2\sqrt{3}\sqrt{A(M)} + \delta$ and subdividing M into two disks of area at least $A(M)/4$.*

Liokumovich, Nabutovsky and Rotman [2014, Proposition 3.2] apply Papasoglu's result to cut Riemannian disks into two parts of sufficiently big area by a curve of controlled length. We reformulate their result as follows.

Lemma 2.3 (disk subdivision). *Let D be a Riemannian two-disk. For any $\lambda < 1/4$ and $\delta > 0$ there exists a subdisk $D' \subset D$ such that $\lambda A(D) \leq A(D') \leq (1-\lambda)A(D)$ and $|\partial D' \setminus \partial D| \leq 2\sqrt{3}\sqrt{A(D)} + \delta$.*

Technical width. We now introduce our main tool, the θ-width, reformulate Theorem 1.1 in terms of this invariant, and show how to derive Theorems 1.2 and 1.3.

Definition 2.4 (θ-width). Let M be a compact Riemannian surface (possibly with nonempty boundary). We define the θ-width, denoted by $W_\theta(M)$, as the infimum of the $L > 0$ such that there exists a continuous map f from M to a trivalent tree T satisfying the following conditions:

(W1) $f(\partial M) \subset \partial T$ and the preimage of a terminal vertex is either an interior point or a connected component of ∂M.

(W2) The preimage of an interior point of an edge is homeomorphic to a circle.

(W3) The preimage of a trivalent vertex is homeomorphic to the letter θ.

(W4) The preimage of any point has length at most L.

Observe that in particular $W_\theta(M) \geq |\partial M|$.

Our results are consequences of the following estimate.

Theorem 2.5. *Let D be a Riemannian two-disk. Then*

$$W_\theta(D) \leq \max\bigl\{|\partial D| + \sqrt{A(D)},\ (4 + 11\sqrt{3}/4)\sqrt{A(D)}\bigr\}.$$

We will prove this theorem in Section 3. It is straightforward to check that it implies Theorem 1.1. Observe that it also implies the following statement, of which Theorem 1.2 is a direct consequence.

Corollary 2.6. *Let M be a Riemannian two-sphere. Then*

$$W_\theta(M) \leq (2\sqrt{3} + 33/8)\sqrt{A(M)}.$$

Proof of Corollary 2.6. Let M be a Riemannian two-sphere. First divide M into two disks D_1 and D_2 of area at least $A(M)/4$ by a simple closed curve of length at most $3\sqrt{3}\sqrt{A(M)}$ by choosing $\delta = \sqrt{3}\sqrt{A(M)}$ in Papasoglu's result (Lemma 2.2). Observe that choosing a better constant than $3\sqrt{3}$ does not lead to any improvement in our final estimate. Now for each subdisk we have the following bound according to Theorem 2.5:

$$\begin{aligned} W_\theta(D_i) &\leq \max\bigl\{|\partial D_i| + \sqrt{A(D_i)},\ (4 + 11\sqrt{3}/4)\sqrt{A(D_i)}\bigr\} \\ &\leq \bigl(2\sqrt{3} + 33/8\bigr)\sqrt{A(M)} \end{aligned}$$

as $A(D_i) \leq (3/4)A(M)$ for $i = 1, 2$ and $|\partial D_1| = |\partial D_2| \leq 3\sqrt{3}\sqrt{A(M)}$. It is straightforward to check that

$$W_\theta(M) \leq \max\bigl\{W_\theta(D_1),\ W_\theta(D_2)\bigr\} \leq (2\sqrt{3} + 33/8)\sqrt{A(M)}. \qquad \square$$

Existence of short closed geodesics. It is classic to derive for spheres the existence of short closed geodesics from bounds on the θ-width. In particular:

Theorem 2.7. (1) *A Riemannian two-sphere with area* 1 *carries a closed geodesic of length at most* $8/\sqrt{3}+11/2 \simeq 10.1$.

(2) *A Finsler reversible two-sphere with Holmes–Thompson area* 1 *carries a closed geodesic of length at most* $\sqrt{\pi/2}\,(8/\sqrt{3}+11/2) \simeq 12.7$.

(3) *A Finsler possibly nonreversible two-sphere with Holmes–Thompson area* 1 *carries a closed geodesic of length at most* $\sqrt{3\pi}\,(8/\sqrt{3}+11/2) \simeq 31.1$.

Proof. It follows from [Alvarez Paiva et al. 2013, Section 4.4] that:

- If any Riemannian sphere M with unit area satisfies $W_\theta(M) \le C$, then any reversible Finsler sphere M' with unit Holmes–Thompson area satisfies $W_\theta(M') \le \sqrt{\pi/2}\,C$.
- If any reversible Finsler sphere M with unit Holmes–Thompson area satisfies $W_\theta(M) \le C$, then any Finsler sphere M'' with unit Holmes–Thompson area satisfies $W_\theta(M') \le \sqrt{6}\,C$.

Now fix a Finsler sphere M. We denote by $\mathrm{sys}(M)$ the systole of M, defined as the length of a shortest closed geodesic. By Corollary 2.6 it remains to prove that

$$\mathrm{sys}(M) \le \tfrac{4}{3}\,W_\theta(M). \tag{2-1}$$

The existence of a closed geodesic on M can be proved through a minimax argument on the one-cycle space $\mathcal{Z}_1(M;\mathbb{Z})$. We refer the reader to [Balacheff and Sabourau 2010] and the references therein for additional information. Loosely speaking, this space arising from geometric measure theory is made of multiple curves (unions of oriented loops) endowed with some special topology. This space allows us to define a minimax process on the Finsler sphere M using F. Almgren's isomorphism between the relative fundamental group $\pi_1(\mathcal{Z}_1(M;\mathbb{Z}),\{0\})$ and the second homology group $H_2(M;\mathbb{Z}) \simeq \mathbb{Z}$. From a result of J. Pitts, the minimax quantity

$$\inf_{(z_t)} \sup_{0\le t\le 1} |z_t|,$$

where (z_t) runs over the families of one-cycles inducing a nontrivial element of $\pi_1(\mathcal{Z}_1(M;\mathbb{Z}),\{0\})$, bounds from above the systole.

We argue by contradiction. Suppose that $\mathrm{sys}(M) > \frac{4}{3}\,W_\theta(M)$. Fix $\epsilon > 0$ such that $\mathrm{sys}(M) > \frac{4}{3}\,W_\theta(M)+\epsilon$. By definition there exists a continuous map f from M to a trivalent tree T satisfying (W1)–(W4) with length $L = W_\theta(M)+\epsilon$.

Let v be a trivalent vertex. Its preimage, denoted by $\theta(v)$, is made of three disjoint oriented arcs α_1, α_2, and α_3 with the same endpoints, ordered such that $|\alpha_1| \le |\alpha_2| \le |\alpha_3|$. Denote by β_{ij} for $1 \le i < j \le 3$ the concatenation of the oriented arcs α_i with $-\alpha_j$. As $\operatorname{sys}(M) > |\beta_{ij}|$ we can continuously contract each of the β_{ij} to a point curve through a length decreasing homotopy by using a Birkhoff process; see [Croke 1988, pp. 4–5]. We denote by $\{\beta_{ij}^t\}_{t\in[0,1]}$ this homotopy with the convention that $\beta_{ij}^0 = \beta_{ij}$. We define an element of $\pi_1(\mathscr{Z}_1(M;\mathbb{Z}), \{0\})$ by

$$f_v(t) = \begin{cases} -\beta_{12}^{1-2t} + \beta_{13}^{1-2t} & \text{if } t \in \left[0, \frac{1}{2}\right], \\ \beta_{23}^{2t-1} & \text{if } t \in \left[\frac{1}{2}, 1\right]. \end{cases}$$

This gives rise to an element $[f_v] \in H_2(M, \mathbb{Z})$ such that $|f_v(t)| \leq \frac{4}{3} W_\theta(M) + \epsilon$ for any $t \in [0, 1]$.

Now fix an edge $e = [v_0, v_1] \simeq [0, 1]$ which is not terminal. We denote by α_t the preimage of an interior point of e corresponding to the parameter $t \in]0, 1[$ and orient it in a coherent way. For $i = 0, 1$, denote by α_i the oriented curve obtained as the limit of the curves α_t when $t \to i$. The curve α_i is a simple closed curve contained in $\theta(v_i)$. As before we can contract α_i to a point through a homotopy $\{\alpha_i^t\}_{t\in[0,1]}$. We define an element of $\pi_1(\mathscr{Z}_1(M;\mathbb{Z}), \{0\})$ by

$$f_e(t) = \begin{cases} \alpha_0^{1-3t} & \text{for } t \in \left[0, \frac{1}{3}\right], \\ \alpha_{3t-1} & \text{for } t \in \left[\frac{1}{3}, \frac{2}{3}\right], \\ \alpha_1^{3t-2} & \text{for } t \in \left[\frac{2}{3}, 1\right]. \end{cases}$$

This gives rise to an element $[f_e] \in H_2(M, \mathbb{Z})$ such that $|f_e(t)| \leq W_\theta(M) + \epsilon$ for any $t \in [0, 1]$.

Finally, fix a terminal edge $e = [v_0, v_1] \simeq [0, 1]$, with the terminal vertex corresponding to 0. With the same notation as above, the curve α_0 is reduced to a point curve. We define an element of $\pi_1(\mathscr{Z}_1(M;\mathbb{Z}), \{0\})$ by

$$f_e(t) = \begin{cases} \alpha_{2t} & \text{for } t \in [0, \frac{1}{2}], \\ \alpha_1^{2t-1} & \text{for } t \in \left[\frac{1}{2}, 1\right]. \end{cases}$$

This gives rise to an element $[f_e] \in H_2(M, \mathbb{Z})$ such that $|f_e(t)| \leq W_\theta(M) + \epsilon$ for any $t \in [0, 1]$.

It is straightforward to see (compare with [Balacheff 2003–2004, Section 1.3]):

$$[S^2] = \sum_{e\in E(T)} \varepsilon_e \cdot [f_e] + \sum_{v\in V(T)} \varepsilon_v \cdot [f_v]$$

for some choice of coefficients ε_v and ε_e in $\{-1, 1\}$. Here $E(T)$ and $V(T)$ denote the set of edges and the set of vertices of T, respectively. This implies that there exists an edge e such that $[f_e] \neq 0$ or a vertex v such that $[f_v] \neq 0$. According to the minimax principle on the one-cycle space, we conclude that

$$\operatorname{sys}(M) \leq \tfrac{4}{3} W_\theta(M) + \epsilon,$$

which is a contradiction. □

Remark 2.8. Using the estimate (2-1), we observe that the flat metric with three conical singularities of angle $2\pi/3$ on the two-sphere obtained by gluing two flat equilateral triangles of side 1 along their boundary satisfies

$$\frac{W_\theta}{\sqrt{A}} \geq \tfrac{3}{4} \cdot 2^{\frac{1}{2}} 3^{\frac{1}{4}} \geq 1.39.$$

This proves that the constant in Theorem 1.2 is within a factor at most 6 from the optimal one.

3. The θ-width of a Riemannian disk

In this section we prove Theorem 2.5. For this we adapt the strategy of the proof of [Liokumovich et al. 2014, Theorem 1.6] to control our invariant W_θ.

Reduction to the short boundary case.

Lemma 3.1. *Let D be a Riemannian two-disk and $C \geq 0$. Suppose that there exists $\eta > 0$ such that for any subdisk $D' \subset D$ for which*

$$|\partial D'| < (4+\eta)\sqrt{A(D')},$$

we have

$$W_\theta(D') \leq (1+\eta) \max\bigl\{|\partial D'| + \sqrt{A(D')},\, C\sqrt{A(D')}\bigr\}.$$

Then, for any subdisk $D' \subset D$,

$$W_\theta(D') \leq (1+\eta) \max\bigl\{|\partial D'| + \sqrt{A(D')},\, C\sqrt{A(D')}\bigr\}.$$

In the sequel, we will use this lemma with constant $C = 0$ (small area case) and

$$C = C_{\lambda,\eta} := 4 + 2\eta + 2\sqrt{3} + \frac{1-\lambda}{\sqrt{3}(1-2\eta)} + \sqrt{1-\lambda}$$

for $0 < \lambda < \frac{1}{4}$ and $\eta > 0$ (general case).

Proof. For any subdisk $D' \subset D$ we define $n(D')$ to be the smallest integer n such that

$$|\partial D'| < \bigl(4 + \eta\bigl(\tfrac{4}{3}\bigr)^n\bigr)\sqrt{A(D')}.$$

Let D' be a subdisk such that $n(D') = 0$. Equivalently, we have that $|\partial D'| < (4+\eta)\sqrt{A(D')}$, and so we are done by assumption.

Now fix an integer n and suppose that for any subdisk $D' \subset D$ such that $n(D') \leq n-1$, we have proven that

$$W_\theta(D') \leq (1+\eta) \max\bigl\{|\partial D'| + \sqrt{A(D')},\, C\sqrt{A(D')}\bigr\}.$$

Let $D' \subset D$ be a subdisk with $n(D') = n$. In particular we have $|\partial D'| > 4\sqrt{A(D')}$, so we can subdivide D' into two subdisks D'_1 and D'_2 of smaller perimeters using a Besicovich cut α of length $\sqrt{A(D')}$ (Lemma 2.1). More precisely,

$$\begin{aligned}|\partial D_i'| &\le \tfrac{3}{4}|\partial D'| + \sqrt{A(D')}\\ &< \tfrac{3}{4}\big(\eta\left(\tfrac{4}{3}\right)^n + 4\big)\sqrt{A(D')} + \sqrt{A(D')}\\ &< \big(\eta\left(\tfrac{4}{3}\right)^{n-1} + 4\big)\sqrt{A(D')}\end{aligned}$$

so $n(D_i') \le n - 1$.

Let $\epsilon > 0$ be small enough so that all points of D' at a distance at least ϵ from $\partial D'$ form a subdisk denoted by $D'' \subset D'$. The subdisk D'' is itself subdivided by the Besicovich's cut α into two subdisks $D_i'' \subset D_i'$ for $i = 1, 2$. By considering ϵ smaller if necessary, we can suppose that $\partial D_i''$ is sufficiently close to $\partial D_i'$ so that $|\partial D_i''| < |\partial D'|$ and $n(D_i'') \le n - 1$. In particular,

$$W_\theta(D_i'') \le (1+\eta)\max\big\{|\partial D_i''| + \sqrt{A(D_i'')},\ C\sqrt{A(D_i'')}\big\}$$

for $i = 1, 2$ by the induction assumption, which implies that

$$W_\theta(D_i'') \le (1+\eta)\max\big\{|\partial D'| + \sqrt{A(D')},\ C\sqrt{A(D')}\big\}.$$

Claim 3.2. $W_\theta(D') \le \max\big\{|\partial D'| + \sqrt{A(D')} + o(\epsilon),\ W_\theta(D_1''),\ W_\theta(D_2'')\big\}$.

Proof of Claim 3.2. Indeed for any $\delta > 0$ and $i = 1, 2$, let $f_i : D_i'' \to T_i$ be a continuous map to a trivalent tree T_i satisfying conditions (W1)–(W4) with length strictly less than $W_\theta(D_i'') + \delta$. Denote by v_i the terminal vertex of T_i corresponding to the boundary $\partial D_i''$. Consider a new edge $e \simeq [0, 1]$ and define a new trivalent tree T obtained from T_1, T_2, and e by identifying v_1, v_2, and the vertex of e corresponding to $\{1\}$ into the same vertex denoted by v. The trees T_1 and T_2 can be thought as subgraphs of T. Denote by $\{\gamma_t\}_{t\in[0,1]}$ a monotone isotopy from $\partial D'$ to $\partial D''$ formed by level sets of the distance function to $\partial D'$. It satisfies $|\gamma_t| \le |\partial D'| + o(\epsilon)$.

We define a new map $f : D' \to T$ as follows:

$$f(x) = \begin{cases} f_i(x) & \text{if } x \in D_i'' \setminus \partial D_i'',\\ v & \text{if } x \in \partial D_1'' \cup \partial D_2'',\\ t & \text{if } x \in \gamma_t \text{ for } t \in [0, 1].\end{cases}$$

By construction we have that the length of the preimages is always strictly less than

$$\max\big\{|\partial D'| + \sqrt{A(D')} + o(\epsilon),\ W_\theta(D_1'') + \delta,\ W_\theta(D_2'') + \delta\big\}.$$

It is easy to check that $f : D' \to T$ satisfies conditions (W1)–(W3), which yields the claim if we let $\delta \to 0$. □

Now Claim 3.2 implies

$$W_\theta(D') \le (1+\eta)\max\big\{|\partial D'| + \sqrt{A(D')},\ C\sqrt{A(D')}\big\}$$

by letting $\epsilon \to 0$, and we are done by induction. □

The small area case.

Lemma 3.3. *Let D be a Riemannian two-disk and $\eta > 0$. There exists $\epsilon > 0$ such that any subdisk $D' \subset D$ with $A(D') \le \epsilon$ satisfies*

$$W_\theta(D') \le (1+\eta)\bigl(|\partial D'| + \sqrt{A(D')}\bigr).$$

Proof. According to Lemma 3.1 with $C = 0$, it is enough to prove the lemma for subdisks D' with

$$|\partial D'| < (4+\eta)\sqrt{\epsilon}.$$

As observed in the proof of [Liokumovich et al. 2014, Lemma 2.3], for r small enough, every ball of radius r is $(1 + O(r))$-bilipschitz homeomorphic to a convex subset of $\mathbb{R}^2$. Hence for ϵ small enough the condition $|\partial D'| < (4+\eta)\sqrt{\epsilon}$ ensures that D' is $(1+\eta)$-bilipschitz to a subset $U \subset \mathbb{R}^2$ with analytic boundary. It is easy to continuously contract the boundary of U into a point through a continuous one-parameter family of closed multicurves — that is, the union of a finite number of closed curves — of U with decreasing length. For this, consider a supporting line ℓ of U. We linearly translate this line in the inner orthogonal direction until we sweep out U and denote by $\{\ell_t\}_{t\in[0,1]}$ this family of translated lines (with the convention that $\ell_0 = \ell$). For each $t \in [0, 1]$ the intersection $\ell_t \cap U$ consists of a finite number of disjoint segments. By transversality we can assume that this number of disjoint segments changes at each step by at most 1, and because the boundary is analytic the number of such steps is finite.

Consider the family of closed multicurves γ_t defined as the boundary of the union $\bigcup_{s\in[t,1]} U \cap \ell_t$. This is a continuous one-parameter family of closed multicurves of U with decreasing length that contracts ∂U to a point. The multicurves involved in this family are not disjoint, but it can be done by slightly perturbing the family in the neighborhood of ∂U without significantly increasing their length. Finally, it is classic to derive from this family a map $f : U \to T$ with T a trivalent tree and satisfying conditions (W1)–(W4), with L as close as wanted from $|\partial U|$; compare with [Gromov 1983, p. 128]. In particular $W_\theta(U) \le |\partial U|$ which in turn implies that $W_\theta(D') \le (1+\eta)|\partial D'|$. □

The general case.

Let D be a Riemannian disk. Fix $\eta > 0$ and $0 < \lambda < \frac{1}{4}$ and define

$$C_{\lambda,\eta} = 4 + 2\eta + 2\sqrt{3} + \frac{1-\lambda}{\sqrt{3}(1-2\eta)} + \sqrt{1-\lambda}.$$

We will argue by induction and prove that for any subdisk $D' \subset D$,

$$W_\theta(D') \le (1+\eta)\max\bigl\{|\partial D'| + \sqrt{A(D')},\ C_{\lambda,\eta}\sqrt{A(D')}\bigr\}.$$

This implies the conclusion of Theorem 2.5 by letting $\eta \to 0$ and $\lambda \to \frac{1}{4}$. According to Lemma 3.1 with $C = C_{\lambda,\eta}$, it is enough to estimate the θ-width of D' under the stronger assumption that $|\partial D'| < (4+\eta)\sqrt{A(D')}$.

Let $\epsilon > 0$ such that the conclusion of Lemma 3.3 holds. For any subdisk $D' \subset D$ we define $m(D')$ to be the smallest integer m such that

$$A(D') \leq \epsilon \left(\frac{1}{1-\lambda}\right)^m.$$

Let D' be a subdisk such that $m(D') = 0$. Equivalently, $A(D') \leq \epsilon$ and we are done according to Lemma 3.3.

Now fix a positive integer m and suppose that for any subdisk $D' \subset D$ with $m(D') \leq m-1$ we have proven that

$$W_\theta(D') \leq (1+\eta)\max\bigl\{|\partial D'| + \sqrt{A(D')},\ C_{\lambda,\eta}\sqrt{A(D')}\bigr\}.$$

Let $D' \subset D$ be a subdisk with $m(D') = m$.

By Lemma 2.3 there exists a subdisk $D'_0 \subset D'$ such that

$$\lambda A(D') \leq A(D'_0) \leq (1-\lambda)A(D') \quad \text{and} \quad |\partial D'_0 \setminus \partial D'| \leq (2\sqrt{3}+\eta)\sqrt{A(D')}.$$

First case. If $\partial D'_0 \cap \partial D' \neq \varnothing$, then D' decomposes into an union of subdisks $D'_0, \dots, D'_k$ with disjoint interiors such that $A(D'_i) \leq (1-\lambda)A(D')$ for $i = 0, \dots, k$. In particular for each i,

$$W_\theta(D'_i) \leq (1+\eta)\max\bigl\{|\partial D'_i| + \sqrt{A(D'_i)},\ C_{\lambda,\eta}\sqrt{A(D'_i)}\bigr\}$$

as $m(D'_i) \leq m-1$, and the inductive assumption applies.

Using a similar argument to that of Claim 3.2, it is straightforward to check that

$$W_\theta(D') \leq (1+\eta)\max\bigl\{|\partial D'| + |\partial D'_0 \setminus \partial D'| + \sqrt{1-\lambda}\sqrt{A(D')},\ C_{\lambda,\eta}\sqrt{1-\lambda}\sqrt{A(D')}\bigr\}$$

as $|\partial D'_i| \leq |\partial D'| + |\partial D'_0 \setminus \partial D'|$ and $A(D'_i) \leq (1-\lambda)A(D')$ for $i = 0, \dots, k$.

Combined with the fact that $|\partial D'| < (4+\eta)\sqrt{A(D')}$, this implies that

$$\begin{aligned} W_\theta(D') &\leq (1+\eta)\max\bigl\{(4+2\eta+2\sqrt{3}+\sqrt{1-\lambda})\sqrt{A(D')},\ C_{\lambda,\eta}\sqrt{A(D')}\bigr\} \\ &\leq (1+\eta)\max\bigl\{|\partial D'| + \sqrt{A(D')},\ C_{\lambda,\eta}\sqrt{A(D')}\bigr\}, \end{aligned}$$

as claimed.

Second case. If $\partial D'_0 \cap \partial D' = \varnothing$, then D' decomposes into the union of the disk D'_0 and an annulus $\mathscr{A}$. Recall that

$$\begin{aligned} |\partial D'| &< (4+\eta)\sqrt{A(D')}, & A(\mathscr{A}) &\leq (1-\lambda)A(D'), \\ |\partial D'_0| &\leq (2\sqrt{3}+\eta)\sqrt{A(D')}, & A(D'_0) &\leq (1-\lambda)A(D'). \end{aligned}$$

Thus $m(D_0') \le m-1$, so that, by the inductive assumption,

$$W_\theta(D_0') \le \max\bigl\{|\partial D_0'| + \sqrt{A(D_0')},\ C_{\lambda,\eta}\sqrt{A(D')}\bigr\} \le C_{\lambda,\eta}\sqrt{A(D')}.$$

Denote by $h(\mathscr{A})$ the height of the annulus, that is, the distance between its two boundary curves. We say that $\mathscr{A}$ *decomposes into a stack of annuli* if there exist a finite number of annuli $\mathscr{A}_1, \dots, \mathscr{A}_k$ with disjoint interiors such that $\mathscr{A} = \bigcup_{i=1}^k \mathscr{A}_i$ and $\mathscr{A}_i \cap \mathscr{A}_{i+1} = \beta_i$ is a common boundary simple closed curve for $i = 1, \dots, k-1$. The following lemma will help us to estimate the θ-width of D'.

Lemma 3.4. *The Riemannian annulus $\mathscr{A}$ decomposes into a stack of annuli $\mathscr{A}_1$, $\mathscr{A}_2, \dots, \mathscr{A}_k$ such that*

$$h(\mathscr{A}_i) \le \frac{\sqrt{1-\lambda}}{2\sqrt{3}(1-2\eta)}\sqrt{A(\mathscr{A})} \quad \textit{for } i = 1, \dots, k,$$

$$|\beta_i| \le \frac{2\sqrt{3}+\eta}{\sqrt{1-\lambda}}\sqrt{A(\mathscr{A})} \qquad \textit{for } i = 1, \dots, k-1.$$

Proof. Suppose that

$$h(\mathscr{A}) > \frac{\sqrt{1-\lambda}}{2\sqrt{3}(1-2\eta)}\sqrt{A(\mathscr{A})}.$$

Consider for every $0 < t < h(\mathscr{A})$ the 1-cycle c_t formed by points of $\mathscr{A}$ at distance t of β_0. By the coarea formula

$$A\bigl(\{c_t \mid t \in [\eta h(\mathscr{A}), (1-\eta)h(\mathscr{A})]\}\bigr) = \int_{\eta h(\mathscr{A})}^{(1-\eta)h(\mathscr{A})} |c_t|\, dt \le A(\mathscr{A}),$$

so that there exists some $t \in [\eta h(\mathscr{A}), (1-\eta)h(\mathscr{A})]$ such that

$$|c_t| \le \frac{2\sqrt{3}}{\sqrt{1-\lambda}}\sqrt{A(\mathscr{A})};$$

otherwise, we derive a contradiction. The cycle c_t can be approximated by a union of smooth closed simple curves with total length at most

$$\frac{2\sqrt{3}+\eta}{\sqrt{1-\lambda}}\sqrt{A(\mathscr{A})}.$$

So $\mathscr{A}$ decomposes into a stack of two annuli $\mathscr{A}_1$ and $\mathscr{A}_2$ such that $\mathscr{A}_1 \cap \mathscr{A}_2$ is a simple closed curve of length at most

$$\frac{2\sqrt{3}+\eta}{\sqrt{1-\lambda}}\sqrt{A(\mathscr{A})}$$

and such that $h(\mathscr{A}_i) \le (1-\eta)h(\mathscr{A})$ for $i = 1, 2$. By iterating this process, we derive the lemma. □

We suppose that the stack decomposition is ordered in such a way that D'_0 and $\mathscr{A}_1$ are adjacent. In the sequel we will denote by β_0 the boundary curve of $\mathscr{A}$ corresponding to $\partial D'_0$ and by β_k the one corresponding to $\partial D'$. Observe in particular that

$$|\beta_i| \leq (2\sqrt{3}+\eta)\sqrt{A(D')}$$

for $i = 0, \dots, k-1$ and that

$$|\beta_k| \leq (4+\eta)\sqrt{A(D')}.$$

Now it remains to estimate the θ-width of D' using this stack decomposition. For each annulus $\mathscr{A}_i$ of the decomposition, choose a minimizing simple path α_i between its two boundary curves. Cutting then along the curve α_i yields to a disk we denote by D'_i whose boundary consists in the concatenation of β_{i-1}, a copy of α_i, β_i and another copy of α_i. Observe that

$$\begin{aligned}|\partial D'_i| &\leq (4+\eta)\sqrt{A(D')} + (2\sqrt{3}+\eta)\sqrt{A(D')} + 2\left(\frac{\sqrt{1-\lambda}}{2\sqrt{3}(1-2\eta)}\right)\sqrt{A(\mathscr{A}_i)}\\ &\leq \left(4+2\eta+2\sqrt{3}+\frac{1-\lambda}{\sqrt{3}(1-2\eta)}\right)\sqrt{A(D')}.\end{aligned}$$

Since $A(D'_i) = A(\mathscr{A}_i) \leq (1-\lambda)A(D')$, we have $m(D'_i) \leq n-1$ for $i = 1, \dots, k$, so that

$$\begin{aligned}W_\theta(D'_i) &\leq \max\left\{|\partial D'_i| + \sqrt{A(D'_i)},\, C_{\lambda,\eta}\sqrt{A(D'_i)}\right\}\\ &\leq \max\left\{\left(4+2\eta+2\sqrt{3}+\frac{1-\lambda}{\sqrt{3}(1-2\eta)}+\sqrt{1-\lambda}\right)\sqrt{A(D')},\, C_{\lambda,\eta}\sqrt{A(D')}\right\}\\ &\leq C_{\lambda,\eta}\sqrt{A(D')}.\end{aligned}$$

by the inductive assumption.

Lemma 3.5. *For $i = 1, \dots, k$,*

$$W_\theta(D'_0 \cup \dots \cup D'_i) \leq \max\left\{W_\theta(D'_0 \cup \dots \cup D'_{i-1}),\, W_\theta(D'_i)\right\}.$$

In particular, $W_\theta(D') \leq \max\{W_\theta(D'_0), \dots, W_\theta(D'_k)\} \leq C_{\lambda,\eta}\sqrt{A(D)}$, which concludes the proof of Theorem 2.5.

Proof. Fix $\delta > 0$ and $i \in [\![1,k]\!]$. Choose a trivalent tree T_i (resp. T'_i) together with a continuous map $f_i : D'_i \to T_i$ (resp. $f'_i : D'_0 \cup \dots \cup D'_{i-1} \to T'_i$) satisfying conditions (W1)–(W4) with associated length strictly less than $W_\theta(D'_i) + \delta$ (resp. $W_\theta(D'_0 \cup \dots \cup D'_{i-1}) + \delta$).

We now fix some notation; see Figure 1. Let v_i denote the terminal vertex of T_i whose preimage is $\partial D'_i$, and v'_i the terminal vertex of T'_i whose preimage is $\beta_{i-1} = \partial(D'_0 \cup \dots \cup D'_{i-1})$. Denote by $U_i \subset T_i$ a small neighborhood of $v_i \in T_i$, by

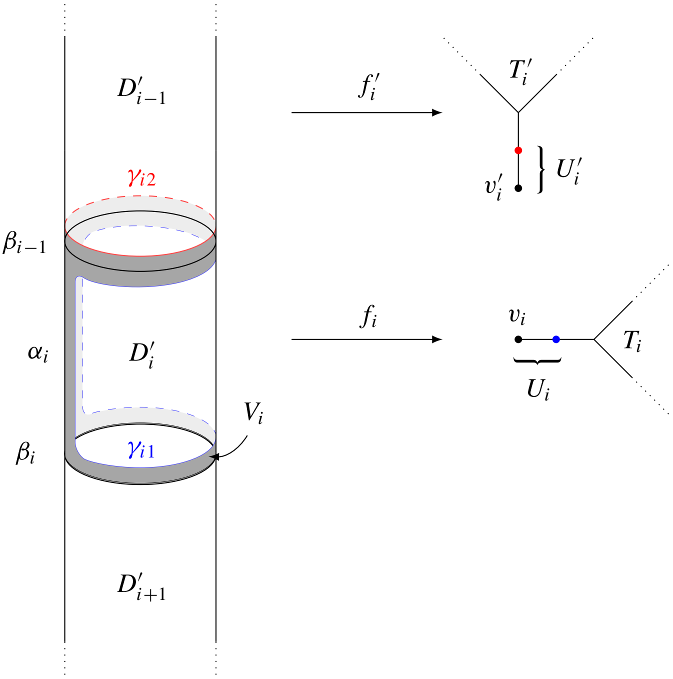


Figure 1. The annulus $\mathscr{A}$ near D'_i.

$U'_i \subset T'_i$ a small neighborhood of $v'_i \in T'_i$, and by V_i the closure of the union of the preimages $f_i^{-1}(U_i)$ and ${f'}_i^{-1}(U'_i)$. The set V_i is isomorphic to a sphere with three boundary components. One of these components is β_i; the other two are denoted by γ_{i1} and γ_{i2}, as in Figure 1.

Observe that γ_{i1} is a small deformation of $\partial D'_i$ viewed as a curve in D'_i, while γ_{i2} is a small deformation of $\beta_{i-1} \subset D'_0 \cup \cdots \cup D'_{i-1}$. In particular,

$$|\gamma_{i1}| = |\partial D'_i| + o(\epsilon) \quad \text{and} \quad |\gamma_{i2}| = |\beta_{i-1}| + o(\epsilon).$$

We will define a new map from $D'_0 \cup \cdots \cup D'_i$ to a trivalent tree by using the restriction of the previous maps f_i and f'_i on the complementary regions of V_i, and completing it on V_i using the following map, whose straightforward construction is depicted in Figure 2 on the next page.

Claim 3.6. *There exists a map $f : V_i \to Y$ where Y is a tripod* (*a trivalent tree with only three edges*) *and satisfying conditions* (W1)–(W4) *with associated length* $|\partial D'_i| + o(\epsilon)$. □

We apply the claim as follows. Consider the trivalent tree T''_i obtained from the disjoint union of $T_i \setminus U_i$, $T'_i \setminus U'_i$, and Y after identification of the terminal vertices of $T_i \setminus U_i$ and Y corresponding to γ_{i1} and the one of $T'_i \setminus U'_i$ and Y corresponding

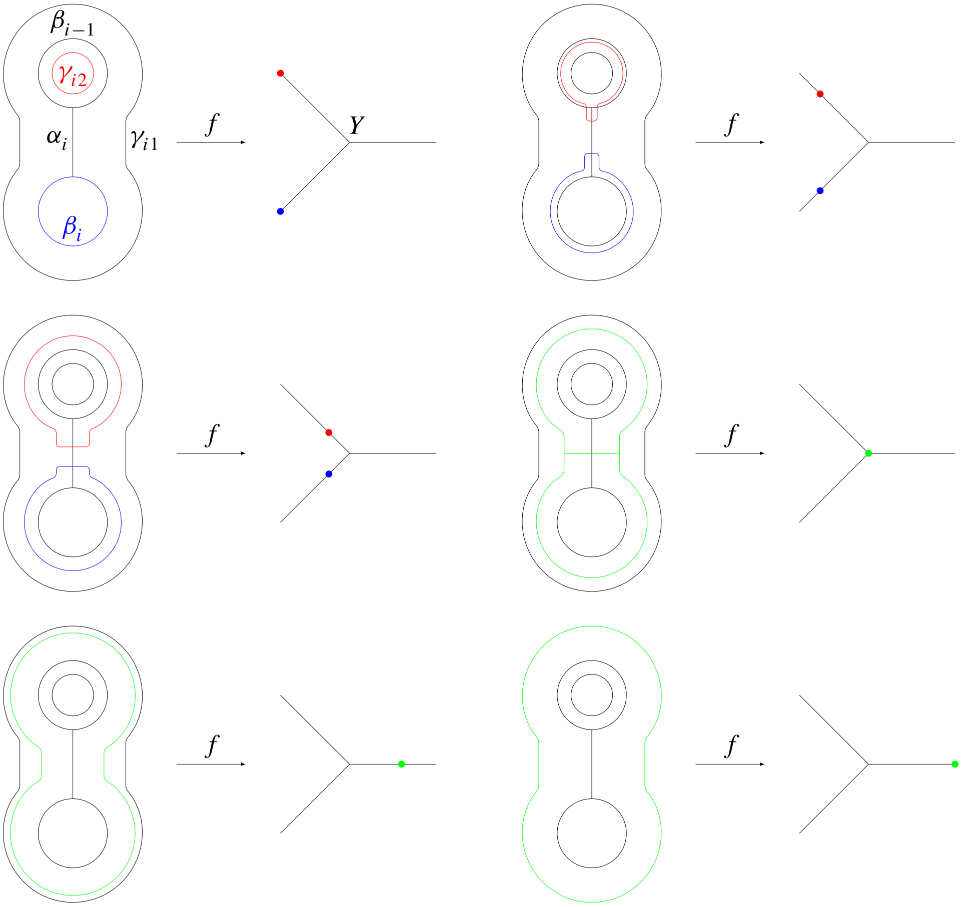


Figure 2. The map $f : V_i \to Y$.

to γ_{i2}. We then define $f_i'' : D_0' \cup \cdots \cup D_i' \to T_i''$ as follows:

$$f_i''(x) = \begin{cases} f_i(x) & \text{if } x \in D_i' \setminus V_i, \\ f_i'(x) & \text{if } x \in D_0' \cup \cdots \cup D_{i-1}' \setminus V_i, \\ f(x) & \text{if } x \in V_i. \end{cases}$$

By construction we have that the length of the preimages is always less than

$$\max\{W_\theta(D_0' \cup \cdots \cup D_{i-1}') + \delta,\ W_\theta(D_i') + \delta,\ |\partial D_i'| + o(\epsilon)\}.$$

This concludes the proof by letting $\epsilon \to 0$ and $\delta \to 0$ as $W_\theta(M) \geq |\partial M|$ for any Riemannian surface M. □

Acknowledgments

The author gratefully thanks Y. Liokumovich, A. Nabutovsky, R. Rotman and K. Tzanev for valuable discussions, and the anonymous referee for useful comments. The author acknowledges partial support for this work by the grant ANR12-BS01-0009 FINSLER. The paper was mainly written during a visit at the CRM of Barcelona. The author would like to thank the institute for its kind hospitality.

References

[Alvarez Paiva et al. 2013] J.-C. Alvarez Paiva, F. Balacheff, and K. Tzanev, "Isosystolic inequalities for optical hypersurfaces", preprint, 2013. arXiv 1308.5522

[Balacheff 2003–2004] F. Balacheff, "Sur des problèmes de la géométrie systolique", *Sémin. Théor. Spectr. Géom.* **22** (2003–2004), 71–82. MR 2006a:53038 Zbl 1083.53043

[Balacheff and Parlier 2012] F. Balacheff and H. Parlier, "Bers' constants for punctured spheres and hyperelliptic surfaces", *J. Topol. Anal.* **4**:3 (2012), 271–296. MR 2982444 Zbl 1262.30046

[Balacheff and Sabourau 2010] F. Balacheff and S. Sabourau, "Diastolic and isoperimetric inequalities on surfaces", *Ann. Sci. Éc. Norm. Supér.* (4) **43**:4 (2010), 579–605. MR 2011k:53046 Zbl 1226.53041

[Besicovitch 1952] A. S. Besicovitch, "On two problems of Loewner", *J. London Math. Soc.* **27** (1952), 141–144. MR 13,831d Zbl 0046.05304

[Croke 1988] C. B. Croke, "Area and the length of the shortest closed geodesic", *J. Differential Geom.* **27**:1 (1988), 1–21. MR 89a:53050 Zbl 0642.53045

[Gromov 1983] M. Gromov, "Filling Riemannian manifolds", *J. Differential Geom.* **18**:1 (1983), 1–147. MR 85h:53029 Zbl 0515.53037

[Guth 2005] L. Guth, "Lipshitz maps from surfaces", *Geom. Funct. Anal.* **15**:5 (2005), 1052–1099. MR 2007e:53029 Zbl 1101.53021

[Liokumovich 2014] Y. Liokumovich, "Slicing a 2-sphere", *J. Topol. Anal.* **6**:4 (2014), 573–590. MR 3238098 Zbl 1296.30053

[Liokumovich et al. 2014] Y. Liokumovich, A. Nabutovsky, and R. Rotman, "Contracting the boundary of a Riemannian 2-disc", preprint, 2014. arXiv 1205.5474

[Papasoglu 2009] P. Papasoglu, "Cheeger constants of surfaces and isoperimetric inequalities", *Trans. Amer. Math. Soc.* **361**:10 (2009), 5139–5162. MR 2010m:20064 Zbl 1183.53030

[Rotman 2006] R. Rotman, "The length of a shortest closed geodesic and the area of a 2-dimensional sphere", *Proc. Amer. Math. Soc.* **134**:10 (2006), 3041–3047. MR 2007f:53039 Zbl 1098.53035

Received March 12, 2014. Revised September 9, 2014.

FLORENT BALACHEFF
LABORATOIRE PAUL PAINLEVÉ
UNIVERSITÉ DES SCIENCES ET TECHNOLOGIES
BÂTIMENT M2
CITÉ SCIENTIFIQUE
59655 VILLENEUVE-D'ASCQ CEDEX
FRANCE
florent.balacheff@math.univ-lille1.fr

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.183

HARMONIC MAPS FROM $\mathbb{C}^n$ TO KÄHLER MANIFOLDS

JIANMING WAN

We prove that a harmonic map from $\mathbb{C}^n$ ($n \geq 2$) to any Kähler manifold must be holomorphic under an assumption on energy density. This can be considered as a complex analogue of the Liouville-type theorem for harmonic maps obtained by Sealey.

1. Introduction

The classical Liouville theorem says that a bounded harmonic function on $\mathbb{R}^n$ (or holomorphic function on $\mathbb{C}^n$) has to be constant. Sealey [1982] (see also [Xin 1996]) gave an analogue for harmonic maps. He proved that a harmonic map of finite energy from $\mathbb{R}^n$ ($n \geq 2$) to any Riemannian manifold must be a constant map. In this paper we consider the complex analogue of Sealey's result by asking: *Must a harmonic map with finite $\bar{\partial}$-energy from $\mathbb{C}^n$ ($n \geq 2$) to any Kähler manifold be holomorphic?*

On the other hand, from Siu and Yau's proof of the Frankel conjecture [1980] (the key is to prove that a stable harmonic map from S^2 to $\mathbb{CP}^n$ is holomorphic or conjugate holomorphic), we know that it is very important to study the holomorphicity of harmonic maps. So the above question is obviously interesting. We hope that it is true. But we do not know how to prove it. Our partial result can be stated as follows:

Theorem 1.1. *Let f be a harmonic map from $\mathbb{C}^n$ ($n \geq 2$) to any Kähler manifold. Let $e(f)$ be the energy density and $e''(f)$ the $\bar{\partial}$-energy density. If*

$$e(f)e''(f)(p) = O\left(\frac{1}{R^{4n+\alpha}}\right)^{1} \tag{1-1}$$

for some $\alpha > 0$, where R denotes the distance from the origin to p, then f is a holomorphic map.

The research is supported by the National Natural Science Foundation of China 11301416.
MSC2010: 53C55.
Keywords: harmonic maps, holomorphic maps.

[1] The notation O means $e(f)e''(f)(p) \leq C/R^{4n+\alpha}$ for some $C > 0$ and sufficiently large R.

The condition (1-1) implies that the $\bar{\partial}$-energy is finite. Since

$$(e''(f))^2 \le e(f)e''(f) = O\left(\frac{1}{R^{4n+\alpha}}\right),$$

one has

$$e''(f) = O\left(\frac{1}{R^{2n+\alpha/2}}\right).$$

This leads to

$$\int_{\mathbb{C}^n} e''(f)\,dv < \infty.$$

Note that we do not have any curvature assumption for the target manifold.

We should mention some other related results on holomorphicity of harmonic maps. For instance, Dong [2013] established many holomorphicity results under the assumption of the target manifolds having strongly seminegative curvature. Xin [1985] obtained some holomorphicity results on harmonic maps from a complete Riemann surface into $\mathbb{CP}^n$.

If the target manifold is $\mathbb{C}^m$ (in this case every component of the map is a harmonic function), then the answer of above question is positive (see [Wan 2010]).

The main idea of the proof of Theorem 1.1 is to consider a one-parameter family of maps and study the $\bar{\partial}$-energy variation.

The rest of the paper is organized as follows. Section 2 contains some basic materials on harmonic maps. In Section 3, we study the first variation of the $\bar{\partial}$-energy. Theorem 1.1 is proved in Section 4.

2. Preliminaries

The materials in this section may be found in [Xin 1996].

2A. ***Basic concepts of harmonic maps.*** Let f be a smooth map between two Riemannian manifolds (M, g) and (N, h). We can define the energy density of f by

$$e(f) = \frac{1}{2}\,\mathrm{trace}\,|df|^2 = \frac{1}{2}\sum_{i=1}^{m}\langle f_*e_i, f_*e_i\rangle,$$

where $\{e_i\}$ $(i = 1, \dots, m = \dim M)$ is a local orthonormal frame field of M. The energy integral is defined by

$$E(f) = \int_M e(f)\,dv.$$

If we choose local coordinates $\{x^i\}$ and $\{y^\alpha\}$ in M and N, respectively, the energy density can be written as

$$e(f)(x) = \frac{1}{2}g^{ij}(x)\frac{\partial f^\alpha(x)}{\partial x^i}\frac{\partial f^\beta(x)}{\partial x^j}h_{\alpha\beta}(f(x)). \tag{2-1}$$

The tension field of f is

$$\tau(f) = (\nabla_{e_i} df)(e_i),$$

where ∇ is the induced connection on the pullback bundle $f^{-1}TN$ over M from those of M and N.

Definition 2.1. We say that f is a harmonic map if $\tau(f) = 0$.

From the variation point of view, a harmonic map can be seen as the critical point of the energy integral functional. Let f_t be a one-parameter family of maps. We can regard it as a smooth map $M \times (-\epsilon, \epsilon) \to N$. Let $f_0 = f$ and $(df_t/dt)|_{t=0} = v$. Then we have the first-variation formula (see [Xin 1996])

$$\frac{d}{dt}E(f_t)\big|_{t=0} = \int_M \operatorname{div} W\, dv - \int_M \langle v, \tau(f)\rangle\, dv, \tag{2-2}$$

where $W = \langle v, f_* e_j\rangle e_j$. If M is compact, then $\int_M \operatorname{div} W\, dv = 0$. We know that a harmonic map is the critical point of the energy functional.

2B. $\bar\partial$*-energy.* Let us consider the complex case. Let f be a smooth map from $\mathbb{C}^n$ to a Kähler manifold N. Let J be the standard complex structure of $\mathbb{C}^n$ and J' the complex structure of N. Let ω and ω^N be the corresponding Kähler forms of $\mathbb{C}^n$ and N (i.e., $\omega(\,\cdot\,,\cdot\,) = \langle J\cdot, \cdot\,\rangle$ and $\omega^N(\,\cdot\,,\cdot\,) = \langle J'\cdot, \cdot\,\rangle$). The $\bar\partial$-energy density is defined by

$$\begin{aligned} e''(f) = |\bar\partial f|^2 &= |f_*J - J'f_*|^2 \\ &= \tfrac{1}{4}(|f_*e_i|^2 + |f_*Je_i|^2 - 2\langle J'f_*e_i, f_*Je_i\rangle) \\ &= \tfrac{1}{2}(e(f) - \langle f^*\omega^N, \omega^M\rangle), \end{aligned}$$

where $\{e_i, Je_i\}$ $(i = 1, \dots, n)$ is the Hermitian frame of $\mathbb{C}^n$ and $\langle f^*\omega^N, \omega\rangle$ denotes the induced norm. We say that f is holomorphic if $f_*J = J'f_*$. Obviously, f is holomorphic if and only if $|\bar\partial f|^2 \equiv 0$.

It is well known that a holomorphic map between two Kähler manifolds must be harmonic (see [Xin 1996]).

We denote the $\bar\partial$-energy by

$$E_{\bar\partial}(f) = \int_{\mathbb{C}^n} |\bar\partial f|^2\, dv.$$

3. $\bar\partial$-energy variation

Let us consider the one-parameter family of maps $f_t(x) = f(tx) : \mathbb{C}^n \to N$, $t \in (1-\epsilon, 1+\epsilon)$ and $f_1 = f$. Let B_R denote the Euclidean ball in $\mathbb{C}^n$ of radius R around 0. We write

$$E(R, t) = \int_{B_R} |\bar\partial f_t|^2\, dv.$$

Lemma 3.1. $E(R, t) = t^{2-2n} E(Rt, 1)$.

Proof. Under the standard Hermitian metric of $\mathbb{C}^n$, $g^{ij} = \delta_{ij}$, from (2-1) we have

$$e(f_t)(x) = t^2 e(f)(tx).$$

By using the natural coordinates, it is easy to show that

$$\langle f_t^* \omega^N, \omega \rangle(x) = t^2 \langle f^* \omega^N, \omega \rangle(tx).$$

So we get

$$|\bar{\partial} f_t|^2(x) = t^2 |\bar{\partial} f|^2(tx).$$

It is easy to check that

$$\int_{B_R} |\bar{\partial} f_t|^2 \, dv = t^{2-2n} \int_{B_{Rt}} |\bar{\partial} f|^2 \, dv.$$

Thus we obtain the lemma. □

We now prove the following variation formula for $\bar{\partial}$-energy:

Lemma 3.2. $\displaystyle \frac{\partial E(R, t)}{\partial t}\bigg|_{t=1} = \frac{R}{2} \int_{\partial B_R} \left(\left| f_* \frac{\partial}{\partial r} \right|^2 - \left\langle J' f_* \frac{\partial}{\partial r},\, f_* J \frac{\partial}{\partial r} \right\rangle \right) dv.$

Proof. Let $\{e_1, \dots, e_{2n} = \partial/\partial r\}$ be a local orthonormal frame field, where $\partial/\partial r$ denotes the unit radial vector field. By the definition of $f_t(x)$, it is easy to see that the variation vector field of f_t at $t = 1$ is

$$v = \frac{df_t}{dt}\bigg|_{t=1} = r f_* \frac{\partial}{\partial r}.$$

The proof is separated into two steps.

Step 1. From (2-2), we have

$$\begin{aligned} \frac{d}{dt} \int_{B_R} e(f_t) \, dv \bigg|_{t=1} &= \int_{B_R} \operatorname{div} \langle v, f_* e_j \rangle e_j \, dv - \int_{B_R} \langle v, \tau(f) \rangle \, dv \\ &= \int_{\partial B_R} \left\langle v, f_* \frac{\partial}{\partial r} \right\rangle dv = R \int_{\partial B_R} \left| f_* \frac{\partial}{\partial r} \right|^2 dv. \end{aligned}$$

Since f is harmonic, we know that the tension field $\tau(f)$ is 0, and the second equality follows from the divergence theorem.

Step 2. On the other hand, from [Xin 1996] we know that $(d/dt)f_t^*\omega^N = d\theta_t$, where $\theta_t = f_t^* i(f_{t*}\partial/\partial t)\omega^N$. Since $(df_t/dt)|_{t=1} = rf_*\partial/\partial r$, we get $\theta_1 = \theta = rf^* i(f_*\partial/\partial r)\omega^N$. Then

$$\begin{aligned}
\frac{d}{dt}\int_{B_R}\langle f_t^*\omega^N, \omega\rangle\, dv\Big|_{t=1} &= \int_{B_R}\langle d\theta, \omega\rangle\, dv \\
&= \int_{B_R} d(\theta\wedge *\omega) + \int_{B_R}\langle\theta, \delta\omega\rangle\, dv \\
&= \int_{\partial B_R}\theta\wedge *\omega - \int_{B_R}\langle\theta, *d\omega^{n-1}\rangle\, dv \\
&= \int_{\partial B_R}\theta\wedge *\omega \\
&= -\int_{\partial B_R}\theta(e_i)\omega\Big(e_i, \frac{\partial}{\partial r}\Big)\, dv \\
&= -R\int_{\partial B_R}\omega^N\Big(f_*\frac{\partial}{\partial r}, f_*e_i\Big)\omega\Big(e_i, \frac{\partial}{\partial r}\Big)\, dv \\
&= -R\int_{\partial B_R}\Big\langle J' f_*\frac{\partial}{\partial r}, f_*e_i\Big\rangle\Big\langle Je_i, \frac{\partial}{\partial r}\Big\rangle dv \\
&= R\int_{\partial B_R}\Big\langle J' f_*\frac{\partial}{\partial r}, f_*J\frac{\partial}{\partial r}\Big\rangle dv.
\end{aligned}$$

Noting that $\langle d\theta, \omega\rangle\, dv = d\theta\wedge *\omega$, the second equality follows from the differential rules, where δ and $*$ are the codifferential and star operators. By Stokes' theorem and the definition of δ, the third equality holds. The fifth equality follows from direct computation. Since we may choose $e_1 = J\,\partial/\partial r$, the last equality holds.

Combining Steps 1 and 2, we obtain

$$\frac{d}{dt}\int_{B_R}|\bar\partial f_t|^2\, dv\Big|_{t=1} = \frac{R}{2}\int_{\partial B_R}\Big(\Big|f_*\frac{\partial}{\partial r}\Big|^2 - \Big\langle J' f_*\frac{\partial}{\partial r}, f_*J\frac{\partial}{\partial r}\Big\rangle\Big)\, dv. \qquad \square$$

Remark 3.3. If M is a compact manifold, $\int_M\langle f^*\omega^N, \omega^M\rangle\, dv$ is a homotopy invariant. This was observed first by Lichnerowicz [1970].

4. Proof of Theorem 1.1

We use a similar trick to [Sealey 1982].

By Lemma 3.1, we obtain

$$\frac{\partial E(R,t)}{\partial t}\Big|_{t=1} = (2-2n)E(R,1) + R\frac{\partial E(R,1)}{\partial R}.$$

On the other hand, from Lemma 3.2 and the condition (1-1), one has

$$\begin{aligned}\frac{\partial E(R,t)}{\partial t}\Big|_{t=1} &= \frac{R}{2}\int_{\partial B_R}\Big(\Big|f_*\frac{\partial}{\partial r}\Big|^2 - \Big\langle J' f_*\frac{\partial}{\partial r},\, f_* J\frac{\partial}{\partial r}\Big\rangle\Big)\,dv\\ &\geq \frac{R}{2}\int_{\partial B_R}\Big(\Big|f_*\frac{\partial}{\partial r}\Big|^2 - \Big|f_*\frac{\partial}{\partial r}\Big|\Big|f_* J\frac{\partial}{\partial r}\Big|\Big)\,dv\\ &\geq -\frac{R}{2}\int_{\partial B_R}\Big|f_*\frac{\partial}{\partial r}\Big|\Big|\Big|f_*\frac{\partial}{\partial r}\Big| - \Big|f_* J\frac{\partial}{\partial r}\Big|\Big|\,dv\\ &\geq -\frac{R}{2}R^{2n-1}\frac{1}{R^{2n+\alpha/2}}C\\ &= -\frac{C}{2}R^{-\alpha/2},\end{aligned}$$

where C is a positive constant. Hence for any $\epsilon > 0$, there exists an R_0 such that

$$\frac{\partial E(R,t)}{\partial t}\Big|_{t=1} \geq -\epsilon$$

for all $R \geq R_0$. Therefore

$$R\frac{\partial E(R,1)}{\partial R} \geq -\epsilon + (2n-2)E(R,1)$$

for $R \geq R_0$.

If $E(\infty,1) = \int_{\mathbb{C}^n} |\bar\partial f|^2\,dv = E > 0$, then there exists an R_1 such that for all $R \geq R_1$ we have $E(R,1) \geq E_0 > 0$. Since $n \geq 2$, we can choose a sufficiently small ϵ such that

$$R\frac{\partial E(R,1)}{\partial R} \geq A = -\epsilon + (2n-2)E_0 > 0$$

when $R \geq R_2 = \max(R_0, R_1)$. Then

$$E(\infty,1) = \int_{\mathbb{C}^n} |\bar\partial f|^2\,dv \geq \int_{R_2}^{\infty}\frac{A}{R}\,dR = \infty.$$

This is a contradiction. Therefore $\int_{\mathbb{C}^n} |\bar\partial f|^2\,dv = 0$. Hence f is a holomorphic map.

Remark 4.1. Compared with the real case [Sealey 1982], Lemma 3.2 has the term $\langle J' f_*(\partial/\partial r),\, f_* J(\partial/\partial r)\rangle$. We need to use condition (1-1) to control it.

Remark 4.2. If we consider the ∂-energy density $e'(f) = |\partial f|^2 = |f_* J + J' f_*|^2$, the corresponding result of Theorem 1.1 also holds; i.e., if the condition (1-1) is replaced by $e(f)e'(f)(p) = O(1/R^{4n+\alpha})$, then the conclusion is that f is a conjugate holomorphic map ($|\partial f|^2 \equiv 0$).

References

[Dong 2013] Y. Dong, "Monotonicity formulae and holomorphicity of harmonic maps between Kähler manifolds", *Proc. Lond. Math. Soc.* (3) **107**:6 (2013), 1221–1260. MR 3149846 Zbl 1295.53070

[Lichnerowicz 1970] A. Lichnerowicz, "Applications harmoniques et variétés Kähleriennes", pp. 341–402 in *Symposia Mathematica* (Rome, 1968/69), vol. III, Academic Press, London, 1970. MR 41 #7598 Zbl 0193.50101

[Sealey 1982] H. C. J. Sealey, "Some conditions ensuring the vanishing of harmonic differential forms with applications to harmonic maps and Yang–Mills theory", *Math. Proc. Cambridge Philos. Soc.* **91**:3 (1982), 441–452. MR 83i:58038 Zbl 0494.58002

[Siu and Yau 1980] Y. T. Siu and S. T. Yau, "Compact Kähler manifolds of positive bisectional curvature", *Invent. Math.* **59**:2 (1980), 189–204. MR 81h:58029 Zbl 0442.53056

[Wan 2010] J. Wan, *Harmonic maps and harmonic complex structures*, thesis, Zhejiang University, Hangzhou, 2010.

[Xin 1985] Y. L. Xin, "Holomorphicity of a class of harmonic maps", *Acta Math. Sinica* **28**:3 (1985), 382–386. In Chinese. MR 87c:58035 Zbl 0598.58018

[Xin 1996] Y. L. Xin, *Geometry of harmonic maps*, Progress in Nonlinear Differential Equations and their Applications **23**, Birkhäuser, Boston, 1996. MR 97h:58052 Zbl 0848.58014

Received May 6, 2014. Revised August 31, 2014.

JIANMING WAN
DEPARTMENT OF MATHEMATICS
NORTHWEST UNIVERSITY
XI'AN, 710127
CHINA
wanj_m@aliyun.com

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.191

EIGENVARIETIES AND INVARIANT NORMS

CLAUS M. SORENSEN

We give a proof of the Breuil–Schneider conjecture in a large number of cases, which complement the indecomposable case, which we dealt with earlier. In this paper, we view the conjecture from a broader global perspective. If $U_{/F}$ is any definite unitary group, which is an inner form of GL(n) over $\mathcal{K}$, we point out how the eigenvariety $\mathbb{X}(K^p)$ parametrizes a global p-adic Langlands correspondence between certain n-dimensional p-adic semisimple representations ρ of Gal($\overline{\mathbb{Q}}|\mathcal{K}$) (or what amounts to the same, pseudorepresentations) and certain Banach–Hecke modules $\mathcal{B}$ with an admissible unitary action of $U(F \otimes \mathbb{Q}_p)$, when p splits. We express the locally regular-algebraic vectors of $\mathcal{B}$ in terms of the Breuil–Schneider representation of ρ. As an application, we give a weak form of local–global compatibility in the crystalline case, showing that the Banach space representations $B_{\xi,\zeta}$ of Schneider and Teitelbaum fit the picture as predicted.

1. Introduction

Let $\mathcal{K}$ be a number field. The Fontaine–Mazur conjecture [1995] predicts a characterization of all (irreducible) Galois representations $r : \Gamma_{\mathcal{K}} = \mathrm{Gal}(\overline{\mathbb{Q}}|\mathcal{K}) \to \mathrm{GL}_n(\mathbb{Q}_p)$ occurring naturally — by which we mean some Tate twist of r is a subquotient of the étale cohomology $H^\bullet(X, \mathbb{Q}_p)$ of some smooth projective variety $X_{/\mathcal{K}}$. It is a major result (due to Tsuji and others) that every such r is *geometric*, which means it is unramified at all but finitely many places, and potentially semistable at all places

MSC2010: 11F33.
Keywords: Galois representations, p-adic Langlands, eigenvarieties, automorphic forms.

above p. Fontaine and Mazur assert the converse: that every geometric r occurs in cohomology (up to a Tate twist). The potentially semistable representations are now more or less completely understood, by work of Colmez and Fontaine [2000]. They are given by admissibly filtered (ϕ, N)-modules (with Galois action), which are objects of a more concrete combinatorial nature. The p-adic Langlands program, still in its initial stages, attempts to link p-adic Hodge theory with nonarchimedean functional analysis. Locally, if K is a fixed finite extension of $\mathbb{Q}_p$, and $L|\mathbb{Q}_p$ is another sufficiently large finite extension (the coefficient field), one hopes to pair certain Galois representations $r : \Gamma_K \to \mathrm{GL}_n(L)$ with certain Banach L-spaces with a unitary admissible $\mathrm{GL}_n(K)$-action. This is now well understood for $\mathrm{GL}_2(\mathbb{Q}_p)$ thanks to recent work of Berger, Breuil, Colmez, Paskunas, and others. See [Berger 2011] for a nice survey.

The goal of this paper is to shed some light on a *global* analogue, for any n, and any CM field $\mathcal{K}$. To give the flavor, if $\mathcal{K}|\mathbb{Q}$ is a quadratic imaginary field in which p splits, we will set up a bijection between certain Galois representations $r : \Gamma_{\mathcal{K}} \to \mathrm{GL}_n(L)$ (actually, pseudorepresentations) and certain Banach–Hecke modules with a unitary admissible $\mathrm{GL}_n(\mathbb{Q}_p)$-action. This is most likely folklore. We emphasize that this bijection is based on matching Satake parameters and Frobenius eigenvalues at places *away* from p. More importantly, we relate the algebraic vectors to the p-adic Hodge theory on the Galois side. The word *certain* here has a precise meaning. It means those representations which *come from an eigenvariety*, of some fixed tame level K^p. We will be precise below.

We model the discussion on the $\mathrm{GL}_2(\mathbb{Q}_p)$-case: To any continuous Galois[1] representation $\rho : \Gamma_{\mathbb{Q}_p} \to \mathrm{GL}_2(L)$, the p-adic Langlands correspondence associates a unitary Banach L-space representation $B(\rho)$ of $\mathrm{GL}_2(\mathbb{Q}_p)$. Moreover, ρ is de Rham with distinct Hodge–Tate weights precisely when there are nonzero locally algebraic vectors: $B(\rho)^{\mathrm{alg}} \neq 0$. Here we use the notation from [Breuil 2010, p. 7]. Some authors prefer to write $B(\rho)^{\text{l-alg}}$. (Recall that a vector is locally algebraic if some open subgroup acts polynomially. They were studied in detail in [Schneider and Teitelbaum 2001]. See Section 3.2 below for more details.) Conversely, if ρ is de Rham with (distinct) Hodge–Tate weights, $\{0, 1-k\}$ (with the convention that the cyclotomic character has weight -1), then the locally algebraic vectors are given by

$$B(\rho)^{\mathrm{alg}} = \mathrm{Sym}^{k-2}(L^2) \otimes \pi(\rho)$$

for a smooth generic representation $\pi(\rho)$, possibly reducible, obtained by a slight modification of the classical local Langlands correspondence.

The Breuil–Schneider conjecture. The local p-adic Langlands program is somewhat vague, and a precise conjectural framework is still developing beyond the case

[1] In the introduction, we will denote global Galois representations by r, local ones by ρ.

of $\mathrm{GL}_2(\mathbb{Q}_p)$, where pretty much everything is known. However, there is a weak (but precise) version formulated in [Breuil and Schneider 2007], which we now recall. We keep our finite extension $K|\mathbb{Q}_p$, and a finite Galois extension thereof, $K'|K$. Pick a third field of coefficients $L \subset \overline{\mathbb{Q}}_p$, finite over $\mathbb{Q}_p$, but large enough so that it contains the Galois closures of K and K'_0 (the maximal unramified subfield of K'). The roles of these fields are the following. We consider potentially semistable representations $\rho : \Gamma_K \to \mathrm{GL}_n(L)$, which become semistable when restricted to $\Gamma_{K'}$. As mentioned above, such ρ correspond to $(\phi, N) \times \mathrm{Gal}(K'|K)$-modules D with an admissible filtration. This makes use of Fontaine's period ring B_{st},

$$D = (B_{\mathrm{st}} \otimes_{\mathbb{Q}_p} \rho)^{\Gamma_{K'}}.$$

This is a finite, free $K'_0 \otimes_{\mathbb{Q}_p} L$-module of rank n with a semilinear Frobenius ϕ, a (nilpotent) monodromy operator N such that $N\phi = p\phi N$, a commuting action of $\mathrm{Gal}(K'|K)$, and an admissible Galois-stable filtration on $D_{K'}$. Note

$$K' \otimes_{\mathbb{Q}_p} L \simeq \prod_{\tau \in \mathrm{Hom}(K,L)} K' \otimes_{K,\tau} L.$$

Accordingly, $D_{K'} \simeq \prod_\tau D_{K',\tau}$, and each $K' \otimes_{K,\tau} L$-module $D_{K',\tau}$ is filtered.

- *Hodge–Tate numbers.* For each $\tau : K \hookrightarrow L$, we let $i_{1,\tau} \leq \cdots \leq i_{n,\tau}$ denote the jumps in the Hodge filtration (listed with multiplicity). That is,

$$\mathrm{gr}^i(D_{K',\tau}) \neq 0 \iff i \in \{i_{1,\tau}, \dots, i_{n,\tau}\}.$$

  We will denote this multiset of integers by $HT_\tau(\rho) = \{i_{j,\tau} : j = 1, \dots, n\}$.

- *Weil–Deligne representation.* If we forget about the filtration, the resulting $(\phi, N) \times \mathrm{Gal}(K'|K)$-module corresponds to a Weil–Deligne representation, once we fix an embedding $K'_0 \hookrightarrow L$; see Proposition 4.1 in [Breuil and Schneider 2007] for details on this correspondence. Basically, one looks at the n-dimensional L-vector space

$$D_L = D \otimes_{K'_0 \otimes_{\mathbb{Q}_p} L} L,$$

  with the induced N coming from B_{st}, and with $r : W_K \to \mathrm{GL}(D_L)$ defined by $r(w) = \phi^{-d(w)} \circ \overline{w}$. Here $\overline{w}$ denotes the image of w in $\mathrm{Gal}(K'|K)$, and $d(w)$ gives the power of arithmetic Frobenius which w induces. The ensuing Weil–Deligne representation becomes unramified upon restriction to $W_{K'}$. We will denote it by $\mathrm{WD}(\rho) = (r, N, D_L)$ throughout the text.

The Breuil–Schneider conjecture asks for a characterization of the data arising in this fashion, assuming all Hodge–Tate numbers are distinct. To state it, start with abstract data. Firstly, for each embedding $\tau : K \hookrightarrow L$, say we are given n distinct integers $HT_\tau = \{i_{1,\tau} < \cdots < i_{n,\tau}\}$. Secondly, say we are given some

n-dimensional Weil–Deligne representation WD, with coefficients in L, which becomes unramified after restriction to $W_{K'}$. With this data, we will associate a locally algebraic representation BS of $\mathrm{GL}_n(K)$, with coefficients in L. The algebraic part is defined in terms of the HT_τ, the smooth part in terms of WD. Our data should come from a Galois representation ρ, as above, precisely when BS has a $\mathrm{GL}_n(K)$-stable $\mathcal{O}_L$-lattice, the unit ball of an invariant norm.

The following was also announced as Conjecture 4.1 in Breuil's [2010] ICM address:

The Breuil–Schneider conjecture [2007, Conjecture 4.3]. *The following are equivalent*:

(1) *The data HT_τ and* $\mathrm{WD}^{\text{F-ss}}$ *arise from a potentially semistable ρ.*

(2) BS *admits a norm $\|\cdot\|$, invariant under the action of* $\mathrm{GL}_n(K)$.

Before we recall the status of the conjecture, we return to the definition of BS.

- *Algebraic part.* Introduce $b_{j,\tau} = -i_{n+1-j,\tau} - (j-1)$. That is, write the $i_{j,\tau}$ in the opposite order, change signs, and subtract $(0, 1, \dots, n-1)$. We let ξ_τ be the irreducible algebraic L-representation of GL_n, of highest weight
$$b_{1,\tau} \le b_{2,\tau} \le \cdots \le b_{n,\tau}$$
relative to the *lower* triangular Borel. Their tensor product $\xi = \bigotimes_\tau \xi_\tau$, with τ running over $\mathrm{Hom}(K, L)$, is then an irreducible algebraic representation of $\mathrm{GL}_n(K \otimes_{\mathbb{Q}_p} L)$ over L. We will view ξ as a representation of $\mathrm{GL}_n(K)$.
- *Smooth part.* By the classical local Langlands correspondence [Harris and Taylor 2001], the Frobenius semisimplification $\mathrm{WD}^{\text{F-ss}}$ is isomorphic to $\mathrm{rec}_n(\pi^\circ)$ for some irreducible admissible smooth representation π° of $\mathrm{GL}_n(K)$, defined over $\bar{\mathbb{Q}}_p$. Here rec_n is normalized as in [loc. cit.]. To define it over $\bar{\mathbb{Q}}_p$, we need to fix a square root $q^{1/2}$, where $q = \#\mathbb{F}_K$. By the Langlands classification, one has
$$\mathrm{Ind}_P(Q(\Delta_1) \otimes \cdots \otimes Q(\Delta_r)) \xrightarrow{!} \!\!\!\!\!\!\rightarrow \pi^\circ,$$
a unique irreducible quotient, where the $Q(\Delta_i)$ are generalized Steinberg representation built from the Δ_i, which are segments of supercuspidals, suitably ordered. The smooth part of BS is now defined to be
$$\pi = \mathrm{Ind}_P(Q(\Delta_1) \otimes \cdots \otimes Q(\Delta_r)) \otimes |\det|^{(1-n)/2},$$
or rather its model over L, which is independent of the choice of $q^{1/2}$. Note that $\pi \simeq \pi^\circ \otimes |\det|^{(1-n)/2}$ if and only if π° is generic (that is, has a Whittaker model). For that reason, the association $\mathrm{WD} \mapsto \pi$ is often called the *generic* local Langlands correspondence.

We let $\mathrm{BS} = \xi \otimes_L \pi$, following [Breuil and Schneider 2007] (although they do not use the notation BS). In fact, we will find it more convenient to work with a different normalization. In the above construction there is a choice of a *sign*, essentially reflected in whether one twists by $|\det|^{(1-n)/2}$ or its inverse. The latter is more commonly used in the references we rely on. The resulting representation is just a twist of BS by a harmless explicit p-adically unitary continuous character. Namely,

$$\widetilde{\mathrm{BS}} = \mathrm{BS} \otimes_L \mu^{n-1}, \quad \mu(g) = N_{K|\mathbb{Q}_p}(\det g)^\times,$$

where $a^\times = a|a|_p = \mathrm{BS}(\chi_{\mathrm{cyc}})(a) \in \mathbb{Z}_p^\times$ denotes the unit factor of an $a \in \mathbb{Q}_p^*$. Of course, $\widetilde{\mathrm{BS}}$ has an invariant norm if and only if BS does, so it makes no real difference. It reflects a Tate twist: $\widetilde{\mathrm{BS}}(\rho)$ is nothing but $\mathrm{BS}(\rho \otimes \chi_{\mathrm{cyc}}^{n-1})$.

The implication (2) ⇒ (1) in the conjecture is in fact completely known. After many cases were worked out in [Schneider and Teitelbaum 2006; Breuil and Schneider 2007], the general case was settled by Y. Hu in his thesis [2009]. In fact, Hu proves that (1) is equivalent to the *Emerton condition*, which is a purely group-theoretical condition:

$$(3) \qquad J_P(\mathrm{BS})^{Z_M^+ = \chi} \neq 0 \quad \Longrightarrow \quad \forall z \in Z_M^+ : |\delta_P^{-1}(z)\chi(z)|_p \leq 1.$$

Here J_P is Emerton's generalization of the Jacquet functor [2006a; 2007]. The heart of Hu's proof is to translate (3) into finitely many inequalities relating the Hodge polygon to the Newton polygon. In the vein of [Fontaine and Rapoport 2005], he is then able to show the existence of an admissible filtration compatible with the given data. The implication (2) ⇒ (3) is relatively easy.

What remains is to produce an invariant norm on $\mathrm{BS}(\rho)$, for any potentially semistable ρ (with distinct Hodge–Tate weights). One of the main motivations for writing this paper was to make progress in this direction, (1) ⇒ (2). The supercuspidal case was dealt with in [Breuil and Schneider 2007] by purely *local* methods (see Theorem 5.2 in [loc. cit.]). The desired norm can be found either by compact induction, or by looking at matrix coefficients.

When $\mathrm{WD}(\rho)$ is indecomposable (in other words, $\pi^\circ = Q(\Delta)$ is generalized Steinberg), we proved (1) ⇒ (2) in [Sorensen 2013]. Here (as in the supercuspidal case) the Emerton condition boils down to just integrality of the central character, and in fact the resulting conjecture was stated explicitly as Conjecture 5.5 in [Breuil and Schneider 2007]. The key point of [Sorensen 2013] was to make use of the fact that $Q(\Delta)$ is a discrete series representation, and therefore admits a *pseudocoefficient*. Inserting this as a test function in the trace formula for a certain definite unitary group, one can pass to a global setup (à la Grunwald–Wang). Finally, the desired norm was found by relating classical algebraic modular forms to the completed cohomology for the definite unitary group, $\widetilde{H}^0$. (This is within the framework of [Emerton 2006b], in which completed cohomology was defined and studied extensively for the tower of locally symmetric spaces — for any reductive group. In our case, however, these constructions boil down to just continuous functions on profinite sets, and Emerton's general machinery is not needed.) The argument in [Sorensen 2013] is purely group-theoretical, and in fact carries over to any connected reductive group over $\mathbb{Q}_p$, exploiting a compact form (using a Galois cohomological computation of Borel and Harder, which shows the existence of locally prescribed forms).

The purpose of this paper is to prove results which *complement* those of [Sorensen 2013]. The idea of relating algebraic modular forms to $\widetilde{H}^0$, already present in [Emerton 2006b], can be pushed further, now that local–global compatibility at $p = \ell$ is available in the "book project" context. This was proved recently by Barnet-Lamb, Gee, Geraghty, and Taylor in the so-called Shin-regular case [Barnet-Lamb et al. 2014], and this regularity hypothesis was then shown to be unnecessary by Caraiani, as part of her Harvard Ph.D. thesis — see [Caraiani 2012]. This results in the following somewhat vague Theorem A, which we will make more precise in Theorem B below.

Theorem A. *The Breuil–Schneider conjecture holds for representations ρ that come from stable automorphic representations π on definite unitary groups, of regular weight (that is, restrictions $\rho \simeq r_{\pi,\iota}|_{\Gamma_{\mathcal{K}_w}}$, at places $w|p$, of irreducible automorphic Galois representations $r_{\pi,\iota}$ of regular weight).*

Eigenvarieties. We will combine the approaches of [Chenevier 2009; Emerton 2006b]. Thus let $\mathcal{K}$ be a CM field, with maximal totally real subfield F. Let D be a central simple $\mathcal{K}$-algebra of $\dim_{\mathcal{K}}(D) = n^2$, equipped with an anti-involution $\star$ of the second kind (that is, $\star|_{\mathcal{K}}$ is conjugation). We introduce the unitary F-group $U = U(D, \star)$, an outer form of $\mathrm{GL}(n)$, which becomes the inner form $D^\times$ over $\mathcal{K}$. It will be convenient to also introduce $G = \mathrm{Res}_{F|\mathbb{Q}}(U)$. We will always assume that U is a totally definite group. In other words, we assume that $G(\mathbb{R})$ is a *compact* Lie group, which is therefore a product of copies of $U(n)$.

We will fix a prime number p such that every place $v|p$ of F splits in $\mathcal{K}$, and such that $D_w^\times \simeq \mathrm{GL}_n(\mathcal{K}_w)$ for every $w|v$. To be safe, we fix an isomorphism $D_w \xrightarrow{\sim} M_n(\mathcal{K}_w)$ once and for all (uniquely determined up to inner automorphisms). To keep track of various identifications, it is customary to *choose* a place $\tilde{v}$ of $\mathcal{K}$ above every $v|p$. Once and for all, we also choose an isomorphism $\iota : \mathbb{C} \xrightarrow{\sim} \overline{\mathbb{Q}}_p$. This gives rise to an identification

$$\mathrm{Hom}(F, \mathbb{R}) = \mathrm{Hom}(F, \mathbb{C}) \simeq \mathrm{Hom}(F, \overline{\mathbb{Q}}_p) = \bigsqcup_{v|p} \mathrm{Hom}(F_v, \overline{\mathbb{Q}}_p),$$

and similarly for $\mathrm{Hom}(\mathcal{K}, \mathbb{C})$. By assumption $F_v \simeq \mathcal{K}_w$ for $w|v$, so the choices $\{\tilde{v}\}$ just amount to fixing a CM-type Φ, which is ordinary for ι, in the sense of [Katz 1978]. This will ensure that the various identifications we make are compatible.

The eigenvariety for G depends on the choice of a tame level $K^p \subset G(\mathbb{A}_f^p)$. It is a reduced rigid analytic space $\mathbb{X}_{/E}$, where we take E to be the Galois closure of F in $\overline{\mathbb{Q}}_p$, with additional structure:

$$\chi : \mathbb{X} \to \widehat{T}, \quad \lambda : \mathcal{H}(K^p)^{\mathrm{sph}} \to \mathcal{O}(\mathbb{X}).$$

Here $\widehat{T}_{/E}$ is weight space, parametrizing locally analytic characters of $T(\mathbb{Q}_p)$, and $\mathcal{H}(K^p)^{\mathrm{sph}}$ is the spherical central subalgebra of the Hecke E-algebra $\mathcal{H}(K^p)$. Finally, $\mathcal{O}(\mathbb{X})$ is the Banach algebra of rigid functions on $\mathbb{X}$. Moreover, there is a Zariski-dense subset $X_{\mathrm{cl}} \subset \mathbb{X}(\overline{\mathbb{Q}}_p)$ such that the evaluation

$$\mathbb{X}(\overline{\mathbb{Q}}_p) \longrightarrow (\widehat{T} \times \operatorname{Spec} \mathcal{H}(K^p)^{\mathrm{sph}})(\overline{\mathbb{Q}}_p), \quad x \mapsto (\chi_x, \lambda_x),$$

identifies X_{cl} with the set of *classical* points: roughly this means that, first of all, $\chi_x = \psi_x \cdot \theta_x$ is locally algebraic (ψ_x is the algebraic part, θ_x is the smooth part), and there exists an automorphic representation π of weight ψ_x such that $\pi_p \hookrightarrow \operatorname{Ind}_B^G(\theta_x)$, and $\mathcal{H}(K^p)^{\mathrm{sph}}$ acts on $\pi_f^{K^p} \neq 0$ by the character λ_x. (The condition that π_p embeds in a principal series is the analogue of the "finite slope" requirement showing up in the classical works of Coleman, Mazur and others. The choice of a θ_x is called a refinement of π.) Thus a classical point x carries a little more information than just an automorphic representation π_x; it keeps track of the refinement θ_x. We refer to Section 4 below for more details on eigenvarieties, which contains references to their various constructions.

It is of utmost importance to us that the eigenvariety carries a family of Galois representations. To be more precise, if we let $\Sigma = \Sigma(K^p)$ be the set of ramified places, there is a unique continuous n-dimensional pseudorepresentation into the unit ball $\mathcal{O}(\mathbb{X})^{\leq 1}$ of $\mathcal{O}(\mathbb{X})$, say

$$\mathcal{T} : \Gamma_{\mathcal{K},\Sigma} \to \mathcal{O}(\mathbb{X})^{\leq 1},$$

associated with $\lambda : \mathcal{H}(K^p)^{\mathrm{sph}} \to \mathcal{O}(\mathbb{X})$, in the sense that for all places $w \notin \Sigma$,

$$\mathcal{T}(\mathrm{Frob}_w) = \lambda(b_{w|v}(h_w)).$$

Here h_w is the element of the spherical Hecke algebra for $\mathrm{GL}_n(\mathcal{K}_w)$, which acts via the sum of the (integral) Satake parameters on spherical vectors, and

$$b_{w|v} : \mathcal{H}(\mathrm{GL}_n(\mathcal{K}_w), \mathrm{GL}_n(\mathcal{O}_{\mathcal{K}_w})) \longrightarrow \mathcal{H}(U(F_v), K_v)$$

is the standard base change homomorphism between the pertaining spherical Hecke algebras, see (4.2) on p. 17 in [Mínguez 2011], and his Theorem 4.1. Here K_v is the factor of K^p at v. It is a hyperspecial maximal compact subgroup of $U(F_v)$ (note that this does not determine K_v up to conjugation when n is even, but our K^p is fixed from the outset). In particular, for each $x \in \mathbb{X}(\overline{\mathbb{Q}}_p)$ there is a unique semisimple Galois representation

$$r_x : \Gamma_{\mathcal{K},\Sigma} \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p), \quad \mathcal{T}_x = \mathrm{tr}(r_x).$$

In fact, the way $\mathcal{T}$ is constructed is by first defining r_x for *regular* classical points $x \in X_{\mathrm{cl}}$, by which we mean the dominant character ψ_x is given by a strictly decreasing sequence of integers (at some place). We will denote this subset of points by $X_{\mathrm{reg}} \subset X_{\mathrm{cl}}$. Thanks to [White 2012, Theorem 6.1, p. 28] (now superseded by [Kaletha et al. 2014]) this guarantees that π has a base change to $\mathrm{GL}_n(\mathbb{A}_{\mathcal{K}})$ of the form $\Pi = \boxplus \Pi_i$, where the Π_i are cohomological *cuspidal* (as opposed to just discrete) automorphic representations to which one can attach Galois representations. Now, X_{reg} can be shown to be Zariski dense, and a formal argument in [Chenevier 2004] interpolates the pseudocharacters $\mathrm{tr}(\rho_x)$ for $x \in X_{\mathrm{reg}}$ by a unique $\mathcal{T}$, which one can then specialize at *any* point $x \in \mathbb{X}(\overline{\mathbb{Q}}_p)$.

We can now rephrase the statement in Theorem A in terms of eigenvarieties: if $x \in X_{\mathrm{reg}}$ is a (classical) point such that r_x is irreducible (as a representation of $\Gamma_{\mathcal{K}}$), and $w|p$ is a place of $\mathcal{K}$, then $r_x|_{\Gamma_{\mathcal{K}_w}}$ is potentially semistable, *and* its locally algebraic representation $\mathrm{BS}(r_x|_{\Gamma_{\mathcal{K}_w}})$ admits a $\mathrm{GL}_n(\mathcal{K}_w)$-invariant norm.

Our main result. The actual construction of an invariant norm $\|\cdot\|$ is more interesting than its mere existence. It comes out of a much more precise result, which we now describe. Fix a finite extension $L|E$. At each point $x \in \mathbb{X}(L)$, we have a pseudorepresentation $\mathcal{T}_x : \Gamma_{\mathcal{K},\Sigma} \to L$ (the trace of an actual representation ρ_x, which may or may not be defined over L). On the other hand, to $x \in \mathbb{X}(L)$ we associate the Banach L-space

$$\mathcal{B}_x = (L \otimes_E \widetilde{H}^0(K^p))^{\mathfrak{h}=\lambda_x},$$

where $\mathfrak{h} = \mathcal{H}(K^p)^{\mathrm{sph}}$ is shorthand notation. This space is really very concrete. The completed cohomology $\widetilde{H}^0(K^p)$ is here nothing but the space of all *continuous* functions

$$f : Y(K^p) \to E, \quad Y(K^p) = \varprojlim_{K_p} Y(K_p K^P), \quad Y(K) = G(\mathbb{Q})\backslash G(\mathbb{A}_f)/K$$

with supremum norm. The superscript $\mathfrak{h} = \lambda_x$ means we take the eigenspace for the character $\lambda_x : \mathfrak{h} \to L$ (not the generalized eigenspace). Note that $\mathcal{B}_x$ is much more than just a Banach L-space: For one thing, it is a Banach module for the Banach–Hecke algebra $\widehat{\mathcal{H}}(K^p)$ (see [Schneider and Teitelbaum 2006] for a detailed discussion of these). For another thing, there is a natural $\widehat{\mathcal{H}}(K^p)$-linear action of $G(\mathbb{Q}_p)$ by *isometries* of $\mathcal{B}_x$, which is admissible (meaning that its mod p reduction $\overline{\mathcal{B}}_x$ is a smooth admissible representation of $G(\mathbb{Q}_p)$ over $\mathbb{F}_L$, in the usual sense). Now, for two points $x, x' \in \mathbb{X}(L)$,

$$\mathcal{T}_x = \mathcal{T}_{x'} \iff \lambda_x = \lambda_{x'} \iff \mathcal{B}_x = \mathcal{B}_{x'},$$

since each $b_{w|v}$ is onto; see Corollary 4.2 in [Mínguez 2011] (a fact also used on p. 10 of [Clozel et al. 2011]). In other words, the set of all pairs $(\mathcal{T}_x, \mathcal{B}_x)$ is the graph of a bijection between the set of pseudorepresentations $\{\mathcal{T}_x : x \in \mathbb{X}(L)\}$, and the set of Banach representations $\{\mathcal{B}_x : x \in \mathbb{X}(L)\}$. That is, there is a one-to-one correspondence:

$$\left\{ \begin{array}{c} n\text{-dimensional pseudo-} \\ \text{representations } \mathcal{T} : \Gamma_{\mathcal{K},\Sigma} \to L \\ \text{coming from } \mathbb{X}(L) \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{Banach } \hat{\mathcal{H}}_L(K^p)\text{-modules } \mathcal{B} \\ \text{with admissible unitary} \\ G(\mathbb{Q}_p)\text{-action, coming from } \mathbb{X}(L) \end{array} \right\}.$$

Here $\mathcal{T} \leftrightarrow \mathcal{B}$ means there is a point $x \in \mathbb{X}(L)$ such that $\mathcal{T} = \mathcal{T}_x$ and $\mathcal{B} = \mathcal{B}_x$. (We say that a pseudocharacter $\mathcal{T} : \Gamma_{\mathcal{K},\Sigma} \to L$ comes from $\mathbb{X}(L)$ if it is of the form $\mathcal{T}_x$ for a point $x \in \mathbb{X}(L)$, and similarly for Banach modules.)

To ease the exposition, let us assume we have *split ramification*. That is, $S(K^p) \subset \mathrm{Spl}_{\mathcal{K}|F}$. Then local base change is defined everywhere, and there is a unique automorphic representation π_x associated with a point $x \in X_{\mathrm{cl}}$ such that r_x is irreducible (indeed its global base change is cuspidal and determined almost everywhere). Experts have informed us that $m(\pi_x) = 1$, but we have not been able to locate it in the literature. Our main result in this paper is the following, which will be proved in Section 6.

Theorem B. *Assume $S(K^p) \subset \mathrm{Spl}_{\mathcal{K}|F}$. For each* classical *point $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}}$, defined over L, such that $m(\pi_x) = 1$, there is a unique* (*up to topological equivalence*) *Banach space $B(r_x)$ over L with an admissible unitary $G(\mathbb{Q}_p)$-action such that*:

(1) *$B(r_x)^{\mathrm{ralg}} \simeq \widetilde{\mathrm{BS}}(r_x) := \bigotimes_{v|p} \widetilde{\mathrm{BS}}(r_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}})$ is dense in $B(r_x)$.*

(2) *There is a $G(\mathbb{Q}_p) \times \hat{\mathcal{H}}(K^p)$-equivariant topological isomorphism,*

$$B(r_x) \otimes \Big(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v} \Big) \xrightarrow{\sim} \overline{\mathcal{B}_x^{\mathrm{ralg}}}.$$

(*Here $\overline{\mathcal{B}_x^{\mathrm{ralg}}}$ denotes the closure of the locally regular-algebraic*[2] *vectors in $\mathcal{B}_x$.*)

(3) *If r_x is crystalline above p, there is a continuous $G(\mathbb{Q}_p)$-equivariant map with dense image,*

$$B_{\xi_x,\zeta_x} \longrightarrow B(r_x),$$

which restricts to an isomorphism $H_{\xi_x,\zeta_x} \xrightarrow{\sim} B(r_x)^{\mathrm{ralg}}$. (*Here $H_{\xi,\zeta}$ and $B_{\xi,\zeta}$ are the spaces introduced by Schneider and Teitelbaum* [2006], *and we take ξ_x of highest weight ψ_x, and ζ_x to be the eigensystem of θ_x.*)

[2] Apparently "regularity" is no longer an issue here (see Section 2.3 below). We warn the reader that the terminology "regular-algebraic" can be misleading here. It is *stronger* than "cohomological". It means the Hodge–Tate gaps (for some embedding τ) are at least *two*. The condition arises in endoscopy when base-changing to GL_n.

Since Theorem B is exclusively about *classical* points, it can be formulated purely in terms of automorphic Galois representations (thus refining and extending Theorem A): for instance, it says that $\bigotimes_{v|p} \widetilde{\mathrm{BS}}(r_{\pi,\iota}|_{\Gamma_{\mathcal{K}_{\tilde{v}}}})$ admits a unitary Banach completion $B(r_{\pi,\iota})$ such that

$$B(r_{\pi,\iota}) \otimes (\pi_f^p)^{K^p} \xrightarrow{\sim} \overline{\widetilde{H}^0(K^p)[\pi]^{\mathrm{ralg}}}.$$

The eigenvariety formulation of Theorem B is more amenable to generalization, and is meant to signal what we expect to be true. Namely, say $x \in \mathbb{X}(L)$ is a point which is not *a priori* known to be classical, but which behaves like a classical point — in that $r_x|_{\Gamma_{\mathcal{K}_w}}$ is potentially semistable of regular weight, for all $w|p$. Then Theorem B should hold verbatim for x, which essentially means x is *necessarily* classical (cf. the Fontaine–Mazur conjecture). Furthermore, if a p-adic local Langlands correspondence exists in this generality, we believe that $B(r_x) = \widehat{\bigotimes}_{v|p} B(r_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}})$ should satisfy something along the lines of (2), for *any* point x: the representations $\pi_{x,v}$ are still defined for $v \nmid p$, via classical local Langlands, and we would hope that $B(r_x) \otimes \big(\bigotimes_{v\nmid p} \pi_{x,v}^{K_v}\big)$ at least embeds into $\mathcal{B}_x$ (perhaps assuming the local restrictions $r_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}}$ are irreducible for $v|p$).

Caraiani, Emerton, Gee, Geraghty, Paskunas, and Shin [Caraiani et al. 2014] have recently announced spectacular work in the principal series case. They employ a delicate variant of the Taylor–Wiles–Kisin patching method (allowing the weight and p-level to vary freely), and construct a candidate for the p-adic local Langlands correspondence for $\mathrm{GL}_n(F)$ via deformations of Galois representations. This makes use of auxiliary global data, and they are unable to show the proposed candidate only depends on the local data at p. However, they are able to say enough about their construction to reduce the Breuil–Schneider conjecture to folklore conjectures in the automorphy lifting world. For instance, they prove that $\mathrm{BS}(\rho)$ admits an admissible unitary Banach completion for potentially crystalline $\rho : \Gamma_F \to \mathrm{GL}_n(E)$ which *lie on an automorphic component* of a certain potentially crystalline deformation ring — which is expected to always hold. *Morally*, we show Breuil–Schneider for $\rho = r_{\pi,\iota}|_{\Gamma_{F_w}}$, and [Caraiani et al. 2014] shows it for potentially crystalline lifts ρ of $\bar{r}_{\pi,\iota}|_{\Gamma_{F_w}}$ — which of course is more general, but their techniques rely on heavy machinery.

Organization of the paper. Section 2 sets up notation used throughout the paper, and recalls how to attach Galois representations to automorphic forms on definite unitary groups. We work out the relation between Breuil–Schneider's normalization BS and our preferred normalization $\widetilde{\mathrm{BS}}$, which is what occurs naturally in cohomology. In Section 3 we briefly discuss completed cohomology in this context, which boils down to just $\widetilde{H}^0$ — continuous functions on a profinite set, and relate its locally algebraic vectors to algebraic modular forms. This is crucial for the general strategy of our paper (and its prequel [Sorensen 2013]). At the end we prove Theorem A. Section 4 introduces eigenvarieties $\mathbb{X}$ and the (semisimple) Galois representations ρ_x they carry — fundamental notions in Section 5, which discusses the Banach representations $\mathcal{B}_x$ associated with arbitrary points x on $\mathbb{X}$, and sets up the bijection $\rho_x \leftrightarrow \mathcal{B}_x$ in Theorem B. Section 6 goes into detail about how one can naturally complete $\widetilde{\mathrm{BS}}$ and get a "rough" candidate for a p-adic local Langlands correspondence $B(r_x) \hookrightarrow \mathcal{B}_x$ when x is classical, and explains to what it extent it satisfies local–global compatibility. We specialize to the crystalline case, and show that our candidate $B(r_x)$ is (almost) a quotient of the purely locally defined Banach representation B_{ξ_x,ζ_x} of Schneider and Teitelbaum — as predicted. Section 7 is logically independent of the rest of the paper, and we include it here only for future reference. It gives an in-depth treatment of "Zariski density of crystalline points" in this context, by expanding on an argument of Emerton (building on work of Katz) in the case of $\mathrm{GL}(2)_{/\mathbb{Q}}$. Density should be important in future work on extending parts of Theorem B to (a priori) *non*classical points.

2. Automorphic Galois representations

We start out by summarizing what is currently known about attaching Galois representations to automorphic representations of definite unitary groups. Due to the work of many people, we now have an almost complete understanding of this, and below we merely navigate the existing literature. We claim no originality in this section. Our goal is simply to state the precise result. Particularly, we want to emphasize the local–global compatibility at $p = \ell$, recently proved in [Barnet-Lamb et al. 2014; Caraiani 2012], which is fundamental for this paper.

2.1. ***Definite unitary groups.*** Throughout this article, we fix a totally real field F, and a CM extension $\mathcal{K}$. We let c denote the nontrivial element of $\mathrm{Gal}(\mathcal{K}|F)$. The places of F will usually be denoted by v, and those of $\mathcal{K}$ by w. We are interested in outer forms U of $\mathrm{GL}(n)_F$, which become an inner form $D^\times$ over $\mathcal{K}$. Here D is a central simple $\mathcal{K}$-algebra, of $\dim_{\mathcal{K}}(D) = n^2$. These forms are unitary groups $U = U(D, \star)$, where $\star$ is an anti-involution on D of the second kind ($\star|_{\mathcal{K}} = c$). Thus, for any F-algebra R,

$$U(R) = \{x \in (D \otimes_F R)^\times : xx^\star = 1\}.$$

We will always assume from now on that $U(F \otimes_{\mathbb{Q}} \mathbb{R})$ is *compact*. Thus, by making a choice of a CM-type Φ, the group may be identified with $U(n)^{\mathrm{Hom}(F,\mathbb{R})}$ (up to conjugation). It will be convenient to work over the rationals, and introduce $G = \mathrm{Res}_{F|\mathbb{Q}}(U)$. With the same Φ one identifies $G(\mathbb{C})$ with $\mathrm{GL}_n(\mathbb{C})^{\mathrm{Hom}(F,\mathbb{R})}$.

2.2. ***Weights of automorphic representations.*** Following standard notation in the subject, $(\mathbb{Z}^n)_+^{\mathrm{Hom}(\mathcal{K},\mathbb{C})}$ will denote the set of tuples $a = (a_\tau)_{\tau \in \mathrm{Hom}(\mathcal{K},\mathbb{C})}$, where each

$a_\tau = (a_{\tau,j})$ itself is a decreasing tuple,

$$a_\tau = (a_{\tau,1} \geq a_{\tau,2} \geq \cdots \geq a_{\tau,n}),$$

of integers. In the obvious way, we can identify a_τ with a dominant weight for $\mathrm{GL}(n)$, relative to the upper triangular Borel. We say a_τ is *regular* if all the inequalities above are strict. We say a is regular if a_τ is regular for *some* τ.

Now, let $\pi = \pi_\infty \otimes \pi_f$ be an automorphic representation of $U(\mathbb{A}_F)$. We will define what it means for π to have weight a: Every embedding $\tau : \mathcal{K} \hookrightarrow \mathbb{C}$ restricts to a $\sigma : F \hookrightarrow \mathbb{R}$, which corresponds to an infinite place $v = v(\sigma)$ of F. With this notation, τ identifies $U(F_v) \simeq U(n)$, under which π_v should be equivalent to the contragredient $\breve{V}_{a_\tau}$, or rather its restriction. Here V_{a_τ} is the irreducible algebraic representation of $\mathrm{GL}_n(\mathbb{C})$ of highest weight a_τ.

Remark. We must have $V_{a_{\tau c}} = \breve{V}_{a_\tau}$. In other words, $a_{\tau c,j} = -a_{\tau,n+1-j}$.

2.3. ***Associating Galois representations.*** We have introduced enough notation in order to formulate the following main result, the foundation for our work. As mentioned already, this is the culmination of collaborative efforts of a huge group of outstanding mathematicians, as will become clear below.

Theorem 1. *Choose a prime p, and an isomorphism $\iota : \mathbb{C} \xrightarrow{\sim} \bar{\mathbb{Q}}_p$. Let π be an automorphic representation of $U(\mathbb{A}_F)$ such that π_∞ has* regular *weight a. Then there exists a unique continuous semisimple Galois representation*

$$\rho_{\pi,\iota} : \Gamma_{\mathcal{K}} = \mathrm{Gal}(\bar{\mathbb{Q}}|\mathcal{K}) \to \mathrm{GL}_n(\bar{\mathbb{Q}}_p)$$

such that the following properties are satisfied:

(a) $\breve{\rho}_{\pi,\iota} \simeq \rho^c_{\pi,\iota} \otimes \epsilon_{\mathrm{cyc}}^{n-1}$,

(b) *For* every *finite place v, and every $w|v$ (even those above p),*

$$\mathrm{WD}(\rho_{\pi,\iota}|_{\Gamma_{\mathcal{K}_w}})^{\text{F-ss}} \simeq \iota\,\mathrm{rec}(\mathrm{BC}_{w|v}(\pi_v) \otimes |\det|_w^{(1-n)/2})$$

whenever $\mathrm{BC}_{w|v}(\pi_v)$ is defined, namely if π_v is unramified or $v = ww^c$ splits.

(c) $\rho_{\pi,\iota}|_{\Gamma_{\mathcal{K}_w}}$ *is potentially semistable for all $w|p$, with Hodge–Tate numbers*

$$HT_{\iota\tau}(\rho_{\pi,\iota}|_{\Gamma_{\mathcal{K}_w}}) = \{a_{\tau,j} + (n-j) : j = 1, \dots, n\}$$

for every $\tau : \mathcal{K} \hookrightarrow \mathbb{C}$ such that $\iota\tau$ lies above w. A word about our normalization here: $\rho_{\pi,\iota} \otimes_{\iota\tau,\mathcal{K}_w} \mathbb{C}_{\mathcal{K}_w}(i)$ has no $\Gamma_{\mathcal{K}_w}$-invariants unless i is of the above form, in which case they form a line. Thus $HT_{\iota\tau}(\epsilon_{\mathrm{cyc}}) = \{-1\}$.

Proof. Ngo's proof of the fundamental lemma makes endoscopic transfer widely available. In particular, weak base change from any unitary group associated with $\mathcal{K}|F$ to $\mathrm{GL}_n(\mathbb{A}_{\mathcal{K}})$ has matured. Building on work of Clozel and Labesse, White [2012] worked out the cohomological case completely. In our given setup, π_v is automatically discrete series for all $v|\infty$, in which case Theorem 6.1 in [White 2012] — or rather the pertaining remarks 6.2 and 6.3 — yields an automorphic representation

$$\Pi = \Pi_1 \boxplus \cdots \boxplus \Pi_t$$

of $\mathrm{GL}_n(\mathbb{A}_{\mathcal{K}})$ which is an isobaric sum of mutually nonisomorphic conjugate self-dual cuspidal automorphic representations Π_i of some $\mathrm{GL}_{n_i}(\mathbb{A}_{\mathcal{K}})$ such that

$$\Pi_w = \mathrm{BC}_{w|v}(\pi_v)$$

for all $w|v$, where v is split or archimedean, or π_v is unramified. The *regularity* of π_∞ ensures that the Π_i are cuspidal (as opposed to just discrete), which in turn implies the previous equality at the archimedean places $w|v$. Let us spell it out in that case: Fix an embedding $\tau : \mathcal{K} \hookrightarrow \mathbb{C}$ inducing $\mathcal{K}_w \simeq \mathbb{C}$. Then,

$$\phi_{\Pi_w} : \mathbb{C}^* = W_{\mathbb{C}} \simeq W_{\mathcal{K}_w} \to \mathrm{GL}_n(\mathbb{C})$$

maps

$$z \mapsto \begin{pmatrix} (z/\bar{z})^{-h_1+(n-1)/2} & & \\ & \ddots & \\ & & (z/\bar{z})^{-h_n+(n-1)/2} \end{pmatrix},$$

for certain $h_j \in \mathbb{Z}$, which are given in terms of the weight by $h_j = a_{\tau,j} + (n-j)$. This last formula is worked out in [Bergeron and Clozel 2005], for example (see their Proposition 5.3.1, p. 63, which gives the Langlands parameters of cohomological representations of $U(a,b)$). These h_j are distinct, so each $\Pi_i \otimes |\det|_{\mathcal{K}}^{(n_i-n)/2}$ is regular-algebraic, *essentially* conjugate self-dual, and cuspidal. By Theorem A of [Barnet-Lamb et al. 2014], and the references therein, we can associate a Galois representation $r_{\Pi_i,\iota}$ to it satisfying the properties analogous to (a)–(c). As a remark, in [loc. cit.] local–global compatibility *at p* is proved assuming Shin regularity, which is much weaker than regularity. In any case, the Shin regularity assumption was removed by Caraiani [2012]. It is then straightforward to check that the representation

$$\rho_{\pi,\iota} = r_{\Pi_1,\iota} \oplus \cdots \oplus r_{\Pi_t,\iota}$$

has the desired properties. It is uniquely determined by (b), by Tchebotarev. □

Remark. It appears within reach to extend the previous argument to the *irregular* case. By [White 2012], one still has a weak base change $\boxplus_{i=1}^t \Pi_i$, but the Π_i are only *discrete*, not cuspidal. By Shapiro's lemma in $(\mathfrak{g}, K)$-cohomology, these Π_i should still be cohomological (of Speh type). By the Moeglin–Waldspurger description of the discrete spectrum of $\mathrm{GL}(n_i)$, one can in turn express each Π_i as an isobaric sum of cusp forms, with which one can associate Galois representations.

After having consulted several experts in the field, we are quite optimistic about this line of argument, and that the ubiquitous regularity[3] assumption (appearing throughout this paper) can safely be dropped. However, we have not made any serious attempt to work out the details. We are hopeful that the recent joint work of Kaletha, Mínguez, Shin, and White [Kaletha et al. 2014], which complements [Mok 2014], should provide the strengthenings of [White 2012] needed.

2.4. ***The Breuil–Schneider recipe.*** We attach to a potentially semistable representation $\rho : \Gamma_{\mathcal{K}_w} \to \mathrm{GL}_n(\bar{\mathbb{Q}}_p)$, with distinct Hodge–Tate numbers, a locally algebraic representation $\mathrm{BS}(\rho)$ of $\mathrm{GL}_n(\mathcal{K}_w)$ on a $\bar{\mathbb{Q}}_p$-vector space, as in the introduction. The Breuil–Schneider conjecture is the mere existence of an invariant norm on $\mathrm{BS}(\rho)$. Our first goal is to prove the conjecture for $\rho = \rho_{\pi,\iota}|\Gamma_{\mathcal{K}_w}$ for any place w of $\mathcal{K}$ above p. We will achieve this below. For now, we will compute $\mathrm{BS}(\rho_{\pi,\iota}|\Gamma_{\mathcal{K}_w})$ explicitly, by relating it to the classical local Langlands correspondence.

In fact, we prefer to use a slightly different normalization: There is a choice of a *sign* involved in the recipe on p. 16 in [Breuil and Schneider 2007]. Instead of twisting by $|\det|_w^{(1-n)/2}$, we prefer to twist by $|\det|_w^{(n-1)/2}$ to make it more compatible with the previous notation. Consequently, $\mathrm{BS}(\rho)$ becomes twisted by an integral character.

Definition 2. For $a \in \mathbb{Q}_p^*$ we let $a^\times = a|a|_p$ denote its unit part. We introduce

$$\mu : \mathrm{GL}_n(\mathcal{K}_w) \xrightarrow{\det} \mathcal{K}_w^* \xrightarrow{N_{\mathcal{K}_w|\mathbb{Q}_p}} \mathbb{Q}_p^* \longrightarrow \mathbb{Q}_p^*/p^{\mathbb{Z}} \simeq \mathbb{Z}_p^\times.$$

That is, $\mu(g) = N_{\mathcal{K}_w|\mathbb{Q}_p}(\det g)^\times$. We will normalize $\mathrm{BS}(\rho)$ as follows:

$$\widetilde{\mathrm{BS}}(\rho) := \mathrm{BS}(\rho) \otimes_{\bar{\mathbb{Q}}_p} \mu^{n-1}.$$

(Of course, this has an invariant norm precisely when $\mathrm{BS}(\rho)$ does.)

Lemma 3. $$\widetilde{\mathrm{BS}}(\rho) = \mathrm{BS}(\rho(n-1)).$$

Proof. Note that the character $a \mapsto a^\times$ (which maps $p \mapsto 1$, and is the identity on $\mathbb{Z}_p^\times$) corresponds to the p-adic cyclotomic character $\chi_{\mathrm{cyc}} : \Gamma_{\mathbb{Q}_p} \to \mathbb{Z}_p^\times$ via local class field theory $\mathbb{Q}_p^* \to \Gamma_{\mathbb{Q}_p}^{\mathrm{ab}}$. For any p-adic field K, it follows that $\mathrm{BS}(\chi_{\mathrm{cyc}})$ is simply the character $a \mapsto N_{K|\mathbb{Q}_p}(a)^\times$. Consequently, $\widetilde{\mathrm{BS}}(\rho) = \mathrm{BS}(\rho \otimes \chi_{\mathrm{cyc}}^{n-1})$. □

We compute it in the automorphic case. Given the local–global compatibility results of [Barnet-Lamb et al. 2014] (generalized in [Caraiani 2012]), this is basically just “bookkeeping”.

[3] We emphasize that the use of “regularity” here is admittedly poor terminology, and should not be confused with “cohomological”. It is stronger, in that the gaps between the Hodge–Tate weights should be at least two. Thus dropping the regularity assumption does not signal that we can say anything for $\mathrm{GL}_2(\mathbb{Q}_p)$ about weight-one forms, for instance.

Corollary 4. *Let π be an automorphic representation of $U(\mathbb{A}_F)$ of regular weight a. Assume $\rho_{\pi,\iota}$ is absolutely* irreducible *(as a representation of the full Galois group $\Gamma_{\mathcal{K}}$). Let $v|p$ be a place of F, either split in $\mathcal{K}$, or such that π_v is unramified. Then, for any place $w|v$ of $\mathcal{K}$, we have*

$$\widetilde{\mathrm{BS}}(\rho_{\pi,\iota}|_{\Gamma_{\mathcal{K}_w}}) = \Big(\bigotimes_{\sigma:\mathcal{K}_w\hookrightarrow\bar{\mathbb{Q}}_p} \check{V}_{a_{\iota^{-1}\sigma}}\Big) \otimes_{\bar{\mathbb{Q}}_p} (\mathrm{BC}_{w|v}(\pi_v)\otimes_{\mathbb{C},\iota}\bar{\mathbb{Q}}_p).$$

(*We abuse notation, and let V_{a_τ} denote the irreducible algebraic representation of* $\mathrm{GL}_n(\bar{\mathbb{Q}}_p)$ *of highest weight a_τ, as opposed to the complex representation from earlier chapters.*)

Proof. What is denoted π^{unit} in [Breuil and Schneider 2007] equals, in our case, $\mathrm{BC}_{w|v}(\pi_v)\otimes|\det|_w^{(1-n)/2}$ (more precisely, $\bigotimes_{\mathbb{C},\iota}\bar{\mathbb{Q}}_p$). When it is generic, the smooth part of $\mathrm{BS}(\rho)$ is

$$\pi^{\mathrm{unit}}\otimes_{\bar{\mathbb{Q}}_p}|\det|_w^{(1-n)/2} = (\mathrm{BC}_{w|v}(\pi_v)\otimes|\det|_w^{1-n})\otimes_{\mathbb{C},\iota}\bar{\mathbb{Q}}_p.$$

In the nongeneric case, π^{unit} has to be replaced by a certain parabolically induced representation. However, if we assume $\rho_{\pi,\iota}$ is (globally) irreducible, we see that $\Pi = \mathrm{BC}_{\mathcal{K}|F}(\pi)$ must be cuspidal, and in particular Π_w is generic. The algebraic part of $\mathrm{BS}(\rho)$ is constructed out of the Hodge–Tate numbers: What is denoted $i_{j,\sigma}$ in [Breuil and Schneider 2007], for an embedding $\sigma:\mathcal{K}_w\hookrightarrow\bar{\mathbb{Q}}_p$, equals $a_{\tau,n+1-j}+(j-1)$ in our notation, where $\sigma=\iota\tau$. In (8) on p. 17 of [Breuil and Schneider 2007], the numbers become

$$b_{\tau,j} := -i_{n+1-j,\sigma}-(j-1) = -a_{\tau,j}-(n-1).$$

Breuil and Schneider's ρ_σ is the irreducible algebraic representation of $\mathrm{GL}_n(\bar{\mathbb{Q}}_p)$ of highest weight $b_{\tau,1}\le\cdots\le b_{\tau,n}$ relative to the *lower* triangular Borel. Relative to the upper triangular Borel, ρ_σ has highest weight $b_{\tau,n}\ge\cdots\ge b_{\tau,1}$, so that $\rho_\sigma\simeq\check{V}_{a_\tau}\otimes\det^{1-n}$ (more precisely, $\bigotimes_{\mathbb{C},\iota}\bar{\mathbb{Q}}_p$). Altogether, the algebraic part is

$$\xi = \bigotimes_{\sigma}\rho_\sigma \simeq \bigotimes_{\tau|w}(\check{V}_{a_\tau}\otimes\det^{1-n})$$

(the tensor product ranging over $\tau:\mathcal{K}\hookrightarrow\mathbb{C}$ such that $\iota\tau$ induces w). Here we abuse notation a bit, and use V_{a_τ} to denote the irreducible algebraic representation of $\mathrm{GL}_n(\bar{\mathbb{Q}}_p)$ of highest weight a_τ. As a representation of $\mathrm{GL}_n(\mathcal{K}_w)$, embedded diagonally in $\prod_{\sigma:\mathcal{K}_w\hookrightarrow\bar{\mathbb{Q}}_p}\mathrm{GL}_n(\bar{\mathbb{Q}}_p)$, the algebraic part becomes

$$\xi = \Big(\bigotimes_{\sigma:\mathcal{K}_w\hookrightarrow\bar{\mathbb{Q}}_p}\check{V}_{a_{\iota^{-1}\sigma}}\Big)\otimes(N_{\mathcal{K}_w|\mathbb{Q}_p}\circ\det)^{1-n},$$

which yields the result. □

3. Completed cohomology

In this section we will prove the Breuil–Schneider conjecture [2007, Conjecture 4.3], for the potentially semistable representations $\rho = \rho_{\pi,\iota}|_{\Gamma_{\mathcal{K}_w}}$ above. This will make heavy use of ideas of Emerton [2006b]. The basic idea is to view $\widetilde{\mathrm{BS}}(\rho)$ as a component of the p-adic automorphic representation $\tilde{\pi} = \tilde{\pi}_p \otimes \pi_f^p$ attached to π, which in turn embeds into the completed cohomology $\tilde{H}^0$ for G.

3.1. ***The p-adic automorphic representation.*** We keep our automorphic representation π of $U(\mathbb{A}_F)$ of regular weight a. Recall that we introduced the group $G = \mathrm{Res}_{F|\mathbb{Q}}(U)$. Interchangeably, below we will view π as an automorphic representation of $G(\mathbb{A})$. We will follow p. 52 in [Emerton 2006b] in attaching a p-adic automorphic representation to π. (The $\mathbb{G}$ there will be our G, and F there will be $\mathbb{Q}$.) This can be done for W-allowable π, where W is an irreducible algebraic representation of $G(\mathbb{C})$, which in this case (where G is compact at infinity) simply means $\pi_\infty \simeq W|_{G(\mathbb{R})}$. See Definition 3.1.3 in [Emerton 2006b].

To make this more explicit, in terms of the weight a, we need to make some identifications. Let us choose a CM-type Φ. For each $\sigma : F \hookrightarrow \mathbb{R}$ we let $\tilde{\sigma}$ denote its lift in Φ. Thus the two extensions to $\mathcal{K}$ are $\{\tilde{\sigma}, \tilde{\sigma}^c\}$. Via the choice of Φ,

$$G(\mathbb{C}) \xrightarrow{\sim}_{\Phi} \mathrm{GL}_n(\mathbb{C})^{\mathrm{Hom}(F,\mathbb{R})}, \quad G(\mathbb{R}) \xrightarrow{\sim}_{\Phi} U(n)^{\mathrm{Hom}(F,\mathbb{R})}.$$

We immediately infer that $W \simeq \bigotimes_{\sigma\in\mathrm{Hom}(F,\mathbb{R})} \check{V}_{a_{\tilde{\sigma}}}$ under these identifications. Via $\iota : \mathbb{C} \xrightarrow{\sim} \bar{\mathbb{Q}}_p$ we identify W with an algebraic representation of $G(\bar{\mathbb{Q}}_p)$. Now

$$G(\bar{\mathbb{Q}}_p) \xrightarrow{\sim}_{\Phi} \prod_{v|p} \mathrm{GL}_n(\bar{\mathbb{Q}}_p)^{\mathrm{Hom}(F_v,\bar{\mathbb{Q}}_p)}$$

allows us to factor our p-adic W accordingly, as $W \simeq \bigotimes_{v|p} W_v$, where we let

$$W_v = \bigotimes_{\substack{\sigma\in\mathrm{Hom}(F,\mathbb{R}) \\ \sigma|v}} \check{V}_{a_{\tilde{\sigma}}}.$$

In the same vein, $G(\mathbb{Q}_p) = \prod_{v|p} U(F_v)$. To go any further, from this point on we assume every $v|p$ splits in $\mathcal{K}$, and that $D_w \simeq M_n(\mathcal{K}_w)$ for each divisor $w|v$. Then $U(F_v) \xrightarrow{\sim} \mathrm{GL}_n(\mathcal{K}_w)$, defined up to conjugation. If we assume (as we may) that our CM-type Φ is ordinary at ι, in the sense of [Katz 1978], then Φ singles out a place $\tilde{v}$ of $\mathcal{K}$ above each $v|p$ of F. With this selection of places at hand,

$$G(\mathbb{Q}_p) \xrightarrow{\sim} \prod_{v|p} \mathrm{GL}_n(\mathcal{K}_{\tilde{v}}).$$

Moreover, the inclusion into $G(\bar{\mathbb{Q}}_p)$ corresponds to the diagonal embeddings

$$\mathrm{GL}_n(\mathcal{K}_{\tilde{v}}) = \mathrm{GL}_n(F_v) \hookrightarrow \mathrm{GL}_n(\bar{\mathbb{Q}}_p)^{\mathrm{Hom}(F_v,\bar{\mathbb{Q}}_p)}.$$

The following is Definition 3.1.5 in [Emerton 2006b], except that we are working with representations over $\bar{\mathbb{Q}}_p$ instead of descending to a finite extension of $\mathbb{Q}_p$.

Definition 5. The classical p-adic automorphic representation of $G(\mathbb{A}_f)$ over $\bar{\mathbb{Q}}_p$ attached to the W-allowable automorphic representation π of $G(\mathbb{A})$ is

$$\tilde{\pi} := \tilde{\pi}_p \otimes_{\bar{\mathbb{Q}}_p} \pi_f^p, \quad \tilde{\pi}_p := W \otimes_{\bar{\mathbb{Q}}_p} \pi_p.$$

Here $G(\mathbb{Q}_p)$ acts diagonally on $W \otimes_{\bar{\mathbb{Q}}_p} \pi_p$, and $G(\mathbb{A}_f^p)$ acts through the second factor π_f^p. (Abusing notation, we write π_p instead of $\pi_p \otimes_{\mathbb{C},\iota} \bar{\mathbb{Q}}_p$ and so on.)

At each $v|p$ we introduce $\tilde{\pi}_v = W_v \otimes_{\bar{\mathbb{Q}}_p} \mathrm{BC}_{\tilde{v}|v}(\pi_v)$, a locally algebraic representation of $\mathrm{GL}_n(\mathcal{K}_{\tilde{v}})$, which depends on the choice of an ordinary CM-type Φ. Moreover, $\tilde{\pi}_p \simeq \bigotimes_{v|p} \tilde{\pi}_v$ under the isomorphism $G(\mathbb{Q}_p) \simeq \prod_{v|p} \mathrm{GL}_n(\mathcal{K}_{\tilde{v}})$.

This leads to the main result of this section.

Proposition 6. *Suppose every $v|p$ of F splits in $\mathcal{K}$, and $D_w \simeq M_n(\mathcal{K}_w)$ for all $w|v$. For each $v|p$ of F pick a place $\tilde{v}|v$ of $\mathcal{K}$ (this amounts to choosing an ι-ordinary CM-type). Let π be an automorphic representation of $U(\mathbb{A}_F)$ of regular weight, and assume $\rho_{\pi,\iota}$ is (globally) irreducible. Then, for all $v|p$ of F,*

$$\widetilde{\mathrm{BS}}(\rho_{\pi,\iota}|_{\Gamma_{\mathcal{K}_{\tilde{v}}}}) \simeq \tilde{\pi}_v,$$

which embeds into $\tilde{\pi}|_{\mathrm{GL}_n(\mathcal{K}_{\tilde{v}})}$ (where we restrict via $U(F_v) \xrightarrow{\sim} \mathrm{GL}_n(\mathcal{K}_{\tilde{v}})$).

Proof. This follows from the preceding discussion, combined with the computation of the Breuil–Schneider representation in Corollary 4 above. □

3.2. ***Algebraic modular forms.*** We will study the space of modular forms for G of a given weight. To put things in a broader perspective, we will use the cohomological framework of [Emerton 2006b], although we will only work with H^0, which is explicit and of a combinatorial nature. In our situation, $G(\mathbb{R})$ is compact and connected, so things simplify tremendously, and we only have cohomology in degree zero. Indeed, for every compact open subgroup $K \subset G(\mathbb{A}_f)$, the corresponding arithmetic quotient is a *finite* set:

$$Y(K) = G(\mathbb{Q})\backslash G(\mathbb{A}_f)/K.$$

An irreducible algebraic representation W of $G(\mathbb{C})$ defines a local system $\mathcal{V}_W$ on each $Y(K)$, and $H^0(Y(K), \mathcal{V}_{\breve{W}})$ is identified with the space of modular forms of level K and weight W. That is, all functions $f : G(\mathbb{A}_f) \to \breve{W}$ which are K-invariant on the right and such that $f(\gamma g) = \gamma f(g)$ for all elements $\gamma \in G(\mathbb{Q})$. Then

$$H^0(\mathcal{V}_{\breve{W}}) := \varinjlim_K H^0(Y(K), \mathcal{V}_{\breve{W}}) \simeq \bigoplus_{\pi : \pi_\infty \simeq W} m_G(\pi)\pi_f$$

is a smooth, admissible, semisimple representation of $G(\mathbb{A}_f)$, which we wish to suitably p-adically complete. Via our choice of $\iota : \mathbb{C} \xrightarrow{\sim} \bar{\mathbb{Q}}_p$, we will view W as a representation of $G(\overline{\mathbb{Q}}_p)$ and so on. Occasionally it will be convenient to work over a field $E \subset \overline{\mathbb{Q}}_p$, finite over $\mathbb{Q}_p$. It suffices to take E large enough so that it contains the image of every embedding $F \hookrightarrow \overline{\mathbb{Q}}_p$. In that case G splits over E, and by highest weight theory W may be defined over E. Thus, from now on, $H^0(\mathcal{V}_{\check{W}})$ is an E-vector space with a smooth admissible $G(\mathbb{A}_f)$-action.

Definition 7. For each tame level $K^p \subset G(\mathbb{A}_f^p)$, following [Emerton 2006b], we introduce

$$H^0(K^p, \mathcal{O}_E/\varpi_E^s) := \varinjlim_{K_p} H^0(Y(K_pK^p), \mathcal{O}_E/\varpi_E^s),$$

and

$$\tilde{H}^0(K^p) := E \otimes_{\mathcal{O}_E} \varprojlim_{s} H^0(K^p, \mathcal{O}_E/\varpi_E^s).$$

The latter is an E-Banach space with a unitary $G(\mathbb{Q}_p)$-action, commuting with the action of the Hecke algebra $\mathcal{H}(K^p)$ of compactly supported K^p-biinvariant E-valued functions on $G(\mathbb{A}_f^p)$. In fact, it becomes a Banach module over the completion $\widehat{\mathcal{H}}(K^p)$. Also,

$$\tilde{H}^0 := \varinjlim_{K^p} \tilde{H}^0(K^p),$$

a locally convex E-vector space with an action of $G(\mathbb{A}_f)$.

In our simple setup, they can all be realized very explicitly. For example,

$$\tilde{H}^0(K^p) = \{\text{continuous } Y(K^p) \xrightarrow{f} E\}, \quad Y(K^p) = \varprojlim_{K_p} Y(K_pK^p)$$

with the supremum norm $\|\cdot\|$. The connection to modular forms is via their locally algebraic vectors. We recall their definition:

Definition 8. Let V be a continuous representation of $G(\mathbb{Q}_p)$ over E, and let W be an absolutely irreducible algebraic representation of $G(\mathbb{Q}_p)$ over E. We assume E is large enough for $\mathrm{End}_{G(\mathbb{Q}_p)}(W) = E$ to hold. The space of locally W-algebraic vectors $V_{W\text{-alg}} \subset V$ is the image of the natural "evaluation" map

$$W \otimes_E \mathrm{Hom}_{G(\mathbb{Q}_p)}(W, V) \xrightarrow{\sim} V_{W\text{-alg}} \subset V.$$

The space of locally algebraic vectors is $V_{\mathrm{alg}} = \bigoplus_W V_{W\text{-alg}}$. The subspace of locally *regular*-algebraic vectors is $V_{\mathrm{ralg}} = \bigoplus_W' V_{W\text{-alg}}$, with W ranging over representations of regular weight (in the sense of Section 2.2).

The key ingredient, which relates completed cohomology to modular forms, is the isomorphism:

Lemma 9. *For any absolutely irreducible algebraic representation W,*

$$W \otimes_E H^0(\mathcal{V}_{\check{W}}) \xrightarrow{\sim} (\widetilde{H}^0)_{W\text{-alg}}.$$

(*When W is only irreducible over E, tensor over* $\mathrm{End}_{\mathfrak{g}}(W)$, *where* $\mathfrak{g} = \mathrm{Lie}G(\mathbb{Q}_p)$.)

Proof. This is Corollary 2.2.25 in [Emerton 2006b] (also spelled out in [Sorensen 2013] for H^0). Let us briefly sketch the main idea. For any tame level K^p, one shows that

$$W \otimes_E H^0(K^p, \mathcal{V}_{\check{W}}) = \varinjlim_{K_p} W \otimes_E H^0(Y(K_pK^p), \mathcal{V}_{\check{W}}) \xrightarrow{\sim} \widetilde{H}^0(K^p)_{W\text{-alg}}.$$

This goes as follows: $H^0(Y(K_pK^p), \mathcal{V}_{\check{W}})$ is a space of classical p-adic modular forms, and it is an easy exercise to identify it with $\mathrm{Hom}_{K_p}(W, \widetilde{H}^0(K^p))$. Now,

$$W \otimes_E \mathrm{Hom}_{K_p}(W, \widetilde{H}^0(K^p)) \xrightarrow{\text{eval.}} \widetilde{H}^0(K^p)$$

is injective since W is absolutely irreducible, even when restricted to K_p (which is Zariski dense). The image of this evaluation map is the W-isotypic subspace of $\widetilde{H}^0(K^p)$. As K_p varies, the maps are compatible, and produces a map out of the direct limit onto $\widetilde{H}^0(K^p)_{W\text{-alg}}$, as desired. □

Remark. For higher-degree cohomology H^i there is an analogous canonical $G(\mathbb{A}_f)$-equivariant map, $W \otimes_E H^i(\mathcal{V}_{\check{W}}) \to (\widetilde{H}^i)_{W\text{-alg}}$, which occurs as the edge map of a certain spectral sequence, but the map is not known to be injective for groups other than $\mathrm{GL}(2)_{\mathbb{Q}}$ (and groups G which are compact at infinity modulo center). Injectivity is what makes the whole machinery of [Emerton 2006b] work; see his Theorem 0.7 and Proposition 2.3.8 on p. 47, for example. In particular, it is available in our case, where $G(\mathbb{R})$ is compact. In general, one would have to localize the spectral sequence at a "*cohomologically*" non-Eisenstein maximal ideal $\mathfrak{m}$ (which means it does not contribute to mod p cohomology outside the middle degree). This is expected to hold when the Galois representation $\bar{\rho}_{\mathfrak{m}}$ is absolutely irreducible (which is what it means for $\mathfrak{m}$ to be non-Eisenstein), but this is difficult to show. Partial results are now available for $U(2,1)$; see Theorem A in [Emerton and Gee 2013].

From the previous discussion, we get decompositions of completed cohomology:

Proposition 10. (1) $\overline{\mathbb{Q}}_p \otimes_E (\widetilde{H}^0)_{W\text{-alg}} \simeq \bigoplus_{\pi:\pi_\infty \simeq W} m_G(\pi)\tilde{\pi}$.

(2) $\overline{\mathbb{Q}}_p \otimes_E \widetilde{H}^0(K^p)_{W\text{-alg}} \simeq \bigoplus_{\pi:\pi_\infty \simeq W} m_G(\pi)\big(\tilde{\pi}_p \otimes_{\overline{\mathbb{Q}}_p} (\pi_f^p)^{K^p}\big)$.

Now, suppose $\mathfrak{h} \subset \mathcal{H}(K^p)$ is a central subalgebra. It then acts on $(\pi_f^p)^{K^p}$ by a character $\lambda_\pi : \mathfrak{h} \to \overline{\mathbb{Q}}_p$. Conversely, say we start out with $\lambda : \mathfrak{h} \to \overline{\mathbb{Q}}_p$. Then,

$$\overline{\mathbb{Q}}_p \otimes_E \widetilde{H}^0(K^p)_{W\text{-alg}}^{\mathfrak{h}=\lambda} \simeq \bigoplus_{\pi:\pi_\infty \simeq W, \lambda_\pi = \lambda} m_G(\pi)\big(\tilde{\pi}_p \otimes_{\overline{\mathbb{Q}}_p} (\pi_f^p)^{K^p}\big).$$

As always, we assume W has *regular* weight, so we know how to attach Galois representations. If $\mathfrak{h}$ contains the spherical part $\mathcal{H}(K^p)^{\mathrm{sph}}$, all the π contributing to the right-hand side have the same Galois representation ρ_λ, by Chebotarev, which we assume is *irreducible*. By Proposition 6, we may factor the above,

$$\overline{\mathbb{Q}}_p \otimes_E \widetilde{H}^0(K^p)^{\mathfrak{h}=\lambda}_{W\text{-alg}} \simeq \Big(\bigotimes_{v|p} \widetilde{\mathrm{BS}}(\rho_\lambda|_{\Gamma_{\mathcal{K}_{\tilde{v}}}})\Big) \otimes_{\overline{\mathbb{Q}}_p} \Big(\bigoplus_{\pi:\pi_\infty \simeq W, \lambda_\pi=\lambda} m_G(\pi)\pi_f^p\Big)^{K^p}.$$

This has the form of a $G(\mathbb{Q}_p) \simeq \prod_{v|p} \mathrm{GL}_n(\mathcal{K}_{\tilde{v}})$-representation tensor an $\mathcal{H}(K^p)$-module. In particular, since $\widetilde{H}^0(K^p)$ carries a $G(\mathbb{Q}_p)$-invariant norm, we finally deduce the Breuil–Schneider conjecture for automorphic Galois representations:

Theorem 11. *If π is an automorphic representation of $U(\mathbb{A}_F)$ of regular weight such that $\rho_{\pi,\iota}$ is irreducible, then $\mathrm{BS}(\rho_{\pi,\iota}|_{\Gamma_{\mathcal{K}_w}})$ admits a $\mathrm{GL}_n(\mathcal{K}_w)$-invariant norm for all places $w|p$ of $\mathcal{K}$.*

The discussion leading up to this theorem strongly suggests a better formulation in terms of eigenvarieties. We will employ this machinery in the next section.

4. Eigenvarieties

Eigenvarieties are rigid analytic spaces interpolating Hecke eigensystems occurring in spaces of automorphic forms of varying weight. Historically, the first example is the Coleman–Mazur eigencurve for $\mathrm{GL}(2)_{\mathbb{Q}}$, revisited by Buzzard, Emerton, Urban, and others. There are different constructions for any reductive group G, which each have their drawbacks and limitations. When $G(\mathbb{R})$ is compact, however, the theory is in good shape, and all constructions are compatible. Below we will combine the approach of [Emerton 2006b] with that of [Chenevier 2009] (for arbitrary totally real F), extending parts of [Bellaïche and Chenevier 2009] (when $F = \mathbb{Q}$).

4.1. ***The classical points.*** By our hypotheses, $G_{\mathbb{Q}_p} \simeq \prod_{v|p} \mathrm{Res}_{\mathcal{K}_{\tilde{v}}|\mathbb{Q}_p} \mathrm{GL}(n)$ is quasisplit, and we pick the Borel pair (B, T), defined over $\mathbb{Q}_p$, corresponding to the product of the upper triangular pairs in each $\mathrm{GL}_n(\mathcal{K}_{\tilde{v}})$.

As in [Emerton 2006b], let $\widehat{T}$ denote the *weight* space. That is, the rigid analytic variety (over the coefficient field E introduced in the introduction) which parametrizes the locally analytic characters on $T(\mathbb{Q}_p)$. In other words,

$$\widehat{T}(A) = \mathrm{Hom}_{la}(T(\mathbb{Q}_p), A^\times)$$

for any affinoid E-algebra A. It comes with a universal map $T(\mathbb{Q}_p) \to \mathcal{O}(\widehat{T})^\times$.

The eigenvariety depends on the choice of *tame* level $K^p \subset G(\mathbb{A}_f^p)$, which we will always assume is decomposable as $\prod_{v \nmid p} K_v$, where K_v is a compact open subgroup of $U(F_v)$ that is hyperspecial for all but finitely many v — say, for all $v \notin S(K^p)$. Correspondingly, the Hecke algebra factors as a tensor product,

$$\mathcal{H}(K^p) = \bigotimes_{v \nmid p} \mathcal{H}(K_v) = \mathcal{H}(K^p)^{\mathrm{ram}} \otimes_E \mathcal{H}(K^p)^{\mathrm{sph}}.$$

Here, $\mathcal{H}(K^p)^{\mathrm{sph}} = \bigotimes_{v \notin S(K^p)} \mathcal{H}(K_v)$ sits as a central subalgebra of $\mathcal{H}(K^p)$; hence it acts by a character on $\pi_f^{K^p}$ for any automorphic π with K^p-invariants.

We now make precise which points we wish to interpolate by an eigenvariety.

Definition 12. Let $E(0, K^p)_{\mathrm{cl}} \subset (\widehat{T} \times \operatorname{Spec} \mathcal{H}(K^p)^{\mathrm{sph}})(\overline{\mathbb{Q}}_p)$ be the subset of pairs $x = (\chi, \lambda)$ for which there exists an irreducible $G(\mathbb{A}_f)$-subquotient π_f of $\overline{\mathbb{Q}}_p \otimes_E H^0(\mathcal{V}_{\check{W}})$, where W is an irreducible algebraic representation of G_E, such that:

(a) $\chi = \psi\theta$, where ψ is the highest weight of W (relative to B), and θ is a smooth character of $T(\mathbb{Q}_p)$ such that $\pi_p \hookrightarrow \operatorname{Ind}_{B(\mathbb{Q}_p)}^{G(\mathbb{Q}_p)}(\theta)$,

(b) $\pi_f^{K^p} \neq 0$, and $\mathcal{H}(K^p)^{\mathrm{sph}}$ acts on it via λ.

This is the definition, and notation, used on p. 5 in [Emerton 2006b].

4.2. ***Eigenvariety conventions.*** Emerton defines the degree-zero cohomological eigenvariety of G, of tame level K^p, to be the rigid analytic closure of $E(0, K^p)_{\mathrm{cl}}$ in $\widehat{T} \times \operatorname{Spec} \mathcal{H}(K^p)^{\mathrm{sph}}$. By the uniqueness part of Theorem 1.6 in [Chenevier 2009], it coincides with the eigenvariety defined there. We will intertwine the two points of view. Thus, with $E(0, K^p)_{\mathrm{cl}}$ is associated a quadruple $(\mathbb{X}, \chi, \lambda, X_{\mathrm{cl}})$, consisting of the following data:

- $\mathbb{X}_{/E}$ is an equidimensional reduced, rigid, analytic variety,
- $\chi : \mathbb{X} \to \widehat{T}$ is a finite morphism (Theorem 0.7 (i) on p. 6 in [Emerton 2006b]),
- $\lambda : \mathcal{H}(K^p)^{\mathrm{sph}} \to \mathcal{O}(\mathbb{X})$ is an E-algebra homomorphism,
- $X_{\mathrm{cl}} \subset \mathbb{X}(\overline{\mathbb{Q}}_p)$ is a Zariski-dense subset,

satisfying various properties (listed in Theorem 1.6 in [Chenevier 2009], for example), the most important of which is the following: the canonical evaluation map

$$\mathbb{X}(\overline{\mathbb{Q}}_p) \longrightarrow (\widehat{T} \times \operatorname{Spec} \mathcal{H}(K^p)^{\mathrm{sph}})(\overline{\mathbb{Q}}_p), \quad x \mapsto (\chi_x, \lambda_x),$$

induces a *bijection*

$$X_{\mathrm{cl}} \xrightarrow{\sim} E(0, K^p)_{\mathrm{cl}}.$$

Moreover, there is a classicality criterion, analogous to Coleman's "noncritical slope implies classical", which we will not use directly (we will use that X_{cl} is Zariski dense, though). More properties will be recalled below when needed, such as the connection with Emerton's Jacquet functor.

Notation. Following standard usage, by $\mathbb{X}(\overline{\mathbb{Q}}_p)$ we mean the union (or direct limit) of all $\mathbb{X}(L) = \operatorname{Hom}_E(\operatorname{Sp}(L), \mathbb{X})$, where L ranges over all the finite extensions of E.

Remark. Loeffler [2011] spells out how Chenevier's construction is related to Emerton's (in the case where $G(\mathbb{R})$ is compact). In addition, he introduces so-called *intermediate* eigenvarieties, where one replaces B with an arbitrary parabolic subgroup (and drops the assumption that G should be quasisplit at p). It would be interesting to adapt our arguments to that setting, and thereby make progress towards the Breuil–Schneider conjecture when π_p does not embed in a principal series (induced from the Borel). This ought to put the results of this paper, and that of [Sorensen 2013], under the same roof. However, at this point we are only producing norms at classical points — where the eigenvariety formalism, strictly speaking, is unnecessary — but the goal is to reach the nonclassical points by somehow p-adically varying the norms $\|\cdot\|_x$ at classical x. We hope to return to these questions elsewhere.

4.3. ***The Galois pseudocharacter.*** At each point $x \in \mathbb{X}(\bar{\mathbb{Q}}_p)$ we will assign a continuous semisimple Galois representation $\rho_x : \Gamma_{\mathcal{K}} \to \mathrm{GL}_n(\bar{\mathbb{Q}}_p)$, which is unramified outside $\Sigma = \Sigma(K^p)$, the places of $\mathcal{K}$ above $S(K^p)$. This is first done at a dense set of classical points, then by a formal argument one interpolates $\mathrm{tr}(\rho_x)$ by a pseudocharacter. We refer to Chapter 1 of [Bellaïche and Chenevier 2009] for an extensive elegant introduction to pseudorepresentations, a notion going back to Wiles for GL(2), and to Taylor for GL(n).

Definition 13. Let $X_{\mathrm{reg}} \subset X_{\mathrm{cl}}$ be the subset of points x such that $\chi_x = \psi_x \theta_x$, where $\psi_x = \bigotimes_{\sigma \in \mathrm{Hom}(F, \bar{\mathbb{Q}}_p)} \psi_{x,\tilde{\sigma}}$ is a regular character of T. That is, some $\psi_{x,\tilde{\sigma}}$ is a regular dominant character of $T_{\mathrm{GL}(n)}$ in the usual sense.

This is a Zariski-dense subset of $\mathbb{X}(\bar{\mathbb{Q}}_p)$; see p. 18 in [Chenevier 2009] and the references given there. Now let $x \in X_{\mathrm{reg}}$, and look at the corresponding pair (χ_x, λ_x), where $\chi_x = \psi_x \theta_x$. There exists an irreducible $G(\mathbb{A}_f)$-summand π_f in $\bar{\mathbb{Q}}_p \otimes_E H^0(\mathcal{V}_{\breve{W}_x})$, where W_x has regular highest weight ψ_x, such that $\mathcal{H}(K^p)^{\mathrm{sph}}$ acts on $\pi_f^{K^p} \neq 0$ via λ_x, and $\pi_p \hookrightarrow \mathrm{Ind}_{B(\mathbb{Q}_p)}^{G(\mathbb{Q}_p)}(\theta_x)$. Thus $\iota^{-1}\pi_f$ is the finite part of an automorphic representation of $U(\mathbb{A}_F)$ of regular weight W_x, unramified outside $S(K^p)$, to which we can associate a continuous semisimple Galois representation

$$\rho_x : \Gamma_{\mathcal{K}} \to \mathrm{GL}_n(\bar{\mathbb{Q}}_p)$$

with the following properties:

(a) $\rho_x^\vee \simeq \rho_x^c \otimes \epsilon_{\mathrm{cyc}}^{n-1}$.

(b) For every finite place $v \nmid p$ of F *outside* $S(K^p)$, and every $w|v$ of $\mathcal{K}$, the local representation $\rho_x|_{\Gamma_{\mathcal{K}_w}}$ is unramified, and satisfies the identity

$$\mathrm{tr}\, \rho_x(\mathrm{Frob}_w) = \lambda_x(b_{w|v}(h_w)).$$

(Here, Frob_w is a geometric Frobenius and h_w is the element of the spherical Hecke algebra for $\mathrm{GL}_n(\mathcal{K}_w)$ acting on an unramified Π_w by $\sum \alpha_i$, where the α_i are the integral Satake parameters. Finally, the map

$$b_{w|v} : \mathcal{H}(\mathrm{GL}_n(\mathcal{K}_w), K_w) \to \mathcal{H}(U(F_v), K_v)$$

is the base change homomorphism between the spherical Hecke algebras. See [Mínguez 2011] for a careful useful discussion of this latter map.)

(c) For every finite place $v|p$ of F, the local representation $\rho_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}}$ is potentially semistable. Furthermore:

  (i) The semisimplification of the attached Weil–Deligne representation is

$$\mathrm{WD}(\rho_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}})^{ss} \simeq \bigoplus_{i=1}^{n} (\theta_{x,\tilde{v}}^{(i)} \circ \mathrm{Art}_{\mathcal{K}_{\tilde{v}}}^{-1}).$$

  (Here $\theta_x = \bigotimes_{v|p} \theta_{x,\tilde{v}}$, where $\theta_{x,\tilde{v}}$ is a smooth character of the diagonal torus $T_{\mathrm{GL}(n)}(\mathcal{K}_{\tilde{v}}) \simeq (\mathcal{K}_{\tilde{v}}^*)^n$, factored as a product $\theta_{x,\tilde{v}}^{(1)} \otimes \cdots \otimes \theta_{x,\tilde{v}}^{(n)}$.)

  (ii) The Hodge–Tate numbers are, for any embedding $\tau : \mathcal{K}_{\tilde{v}} \hookrightarrow \overline{\mathbb{Q}}_p$,

$$\mathrm{HT}_\tau(\rho_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}}) = \{a_{\tau,j} + (n-j) : j = 1, \dots, n\},$$

  where the tuple $(a_{\tau,j})$ corresponds to the dominant character $\psi_{x,v,\tau}$ of $T_{\mathrm{GL}(n)}$. (Here we factor $\psi_x = \bigotimes_{v|p} \otimes_{\tau:\mathcal{K}_{\tilde{v}} \hookrightarrow \overline{\mathbb{Q}}_p} \psi_{x,v,\tau}$.)

Observe that there may be many automorphic representations associated to a given point $x \in X_{\mathrm{cl}}$, but they are all isomorphic outside $S(K^p)$ (and of the same weight). In particular, by (b) and Chebotarev, the Galois representation is independent of the choice of π_f, justifying the notation ρ_x.

Proposition 14. *There exists a unique continuous n-dimensional pseudocharacter $\mathcal{T} : \Gamma_{\mathcal{K},\Sigma} \to \mathcal{O}(\mathbb{X})^{\leq 1}$ such that $\mathcal{T}(\mathrm{Frob}_w) = \lambda(b_{w|v}(h_w))$ for all places $w \notin \Sigma$.*

Proof. We are in the situation of Proposition 7.1.1 in [Chenevier 2004]: $\mathbb{X}$ is reduced, $\mathcal{O}(\mathbb{X})^{\leq 1}$ is a compact subring, and for all $x \in X_{\mathrm{reg}}$, a Zariski-dense subset, we have a representation ρ_x of $\Gamma_{\mathcal{K},\Sigma}$ such that $\mathrm{tr}\,\rho_x(\mathrm{Frob}_w) = \lambda(b_{w|v}(h_w))(x)$. □

Corollary 15. *For every $x \in \mathbb{X}(\overline{\mathbb{Q}}_p)$, there is a unique continuous semisimple Galois representation $\rho_x : \Gamma_{\mathcal{K},\Sigma} \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ such that* $\mathrm{tr}\,\rho_x(\mathrm{Frob}_w) = \lambda_x(b_{w|v}(h_w))$ *for all $w \notin \Sigma$.*

Proof. This follows from Theorem 1 of [Taylor 1991]. □

In particular, this applies to the classical point $x \in X_{\mathrm{cl}}$, *not* in X_{reg}. One of the goals of [Chenevier 2009] was to extend properties (a)–(c) above to this setting. This was partially accomplished; see his Theorems 3.3 and 3.5.

5. Banach space representations

With each point $x \in \mathbb{X}(L)$, we have associated an n-dimensional continuous pseudo-character $\mathcal{T}_x : \Gamma_{\mathcal{K}} \to L$, unramified outside $\Sigma(K^p)$. Here we will associate a Banach $\widehat{\mathcal{H}}_L(K^p)$-module $\mathcal{B}_x$, with an admissible unitary $G(\mathbb{Q}_p)$-action, such that the pairs $(\mathcal{T}_x, \mathcal{B}_x)$ form the graph of a one-to-one correspondence. We explicitly compute the locally (regular) algebraic vectors in $\mathcal{B}_x$ for $x \in X_{\mathrm{reg}}$ such that $\mathcal{T}_x$ is absolutely irreducible, in terms of the Breuil–Schneider representation attached to $\mathcal{T}_x$, or rather, its corresponding Galois representation ρ_x. As a result, we prove the Breuil–Schneider conjecture for such ρ_x.

A global p-adic Langlands correspondence. With the eigenvariety language set up, we can reformulate our findings at the end of Section 2. We let $X_{\mathrm{irr}} \subset \mathbb{X}(\overline{\mathbb{Q}}_p)$ be the points x for which ρ_x is irreducible.

Theorem 16. *Let $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}}$, corresponding to $(\psi_x\theta_x, \lambda_x)$. Let W_x be the irreducible algebraic representation of G_E of highest weight ψ_x. Then,*

$$\overline{\mathbb{Q}}_p \otimes_E \widetilde{H}^0(K^p)^{\mathfrak{h}=\lambda_x}_{W_x\text{-alg}} \simeq \Bigl(\bigoplus_{v|p} \widetilde{\mathrm{BS}}(\rho_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}})\Bigr) \otimes_{\overline{\mathbb{Q}}_p} \Bigl(\bigoplus_{\substack{\pi:\pi_\infty\simeq W_x,\\ \lambda_\pi=\lambda_x}} m_G(\pi)(\pi_f^p)^{K^p}\Bigr),$$

where we write $\mathfrak{h} = \mathcal{H}(K^p)^{\mathrm{sph}}$ for simplicity.

This formula suggests the following definition.

Definition 17. At each point $x \in \mathbb{X}(\overline{\mathbb{Q}}_p)$, we introduce the eigenspace

$$\mathcal{B}_x := (\overline{\mathbb{Q}}_p \otimes_E \widetilde{H}^0(K^p))^{\mathfrak{h}=\lambda_x}.$$

This is a Banach $\widehat{\mathcal{H}}(K^p)$-module with a (commuting) unitary $G(\mathbb{Q}_p)$-action.

We remind ourselves that $\mathcal{B}_x$ is nothing but the space of *continuous* λ_x-eigenforms $f : Y(K^p) \to \overline{\mathbb{Q}}_p$. This sets up a one-to-one correspondence $\rho_x \leftrightarrow \mathcal{B}_x$. That is,

$$\rho_x = \rho_{x'} \iff \lambda_x = \lambda_{x'} \iff \mathcal{B}_x = \mathcal{B}_{x'}$$

for any two $x, x' \in \mathbb{X}(\overline{\mathbb{Q}}_p)$. Let us say that a Galois representation ρ *comes from* $\mathbb{X}$ if $\rho \simeq \rho_x$ for some $x \in \mathbb{X}(\overline{\mathbb{Q}}_p)$, and similarly for Banach modules $\mathcal{B} \simeq \mathcal{B}_x$.

This leads to the main result of this section, which in some sense is the genesis of what follows.

Theorem 18. *The eigenvariety $\mathbb{X}$ mediates a one-to-one correspondence between*:

- *The set of continuous semisimple Galois representations $\rho : \Gamma_{\mathcal{K}} \to \mathrm{GL}_n(\overline{\mathbb{Q}}_p)$ coming from $\mathbb{X}$.* (*In particular*, *ρ is unramified outside $\Sigma(K^p)$.*)
- *The set of Banach $\widehat{\mathcal{H}}(K^p)$-modules $\mathcal{B}$, with unitary $G(\mathbb{Q}_p)$-action, from $\mathbb{X}$.*

We write $\rho \leftrightarrow \mathcal{B}$ when there is a point $x \in \mathbb{X}(\overline{\mathbb{Q}}_p)$ such that $\rho \simeq \rho_x$ and $\mathcal{B} \simeq \mathcal{B}_x$.

(1) *Let* $x \in \mathbb{X}(\overline{\mathbb{Q}}_p)$. *If there is a regular* W *for which* $\mathcal{B}_x^{W\text{-alg}} \neq 0$, *then* ρ_x *is potentially semistable at all places* $w|p$ *of* $\mathcal{K}$.

(2) *Let* $x \in X_{\mathrm{cl}}$. *Then* $\mathcal{B}_x^{W_x\text{-alg}} \neq 0$, *and* $\mathcal{B}_x^{W\text{-alg}} = 0$ *for all regular* $W \neq W_x$.

(3) *For* $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}}$, *the locally regular-algebraic vectors of* $\mathcal{B}_x$ *are*

$$\mathcal{B}_x^{\mathrm{ralg}} = \mathcal{B}_x^{W_x\text{-alg}} \simeq \Bigl(\bigotimes_{v|p} \widetilde{\mathrm{BS}}(\rho_x|\Gamma_{\mathcal{K}_{\tilde{v}}})\Bigr) \otimes_{\overline{\mathbb{Q}}_p} \Bigl(\bigoplus_{\substack{\pi:\pi_\infty \simeq W_x,\\ \lambda_\pi = \lambda_x}} m_G(\pi)(\pi_f^p)^{K^p}\Bigr).$$

Remark. This is a strengthening of Theorem 11, which in the notation of the above Theorem 18 merely says that $\widetilde{\mathrm{BS}}(\rho_x|\Gamma_{\mathcal{K}_{\tilde{v}}})$ admits an invariant norm. We stress that x is a *classical* point here (so that ρ_x is irreducible, of regular weight). Thus $\rho_x \simeq \rho_{\pi,\iota}$ for an automorphic π as in Theorem 11. Part (3) of Theorem 18 is stronger, in that it makes precise how $\widetilde{\mathrm{BS}}(\rho_x|\Gamma_{\mathcal{K}_{\tilde{v}}})$ factors into the Banach representation $\mathcal{B}_x$ of p-adic modular forms. Another key point is that Theorem 18 emphasizes the correspondence $\rho_x \leftrightarrow \mathcal{B}_x$, which is defined for *all* points x on the eigenvariety $\mathbb{X}$ (classical or not).

Proof. First, (1) follows from Proposition 10, which shows there is an automorphic π, with $\pi_\infty \simeq W$, such that $\mathfrak{h}$ acts on $\pi_f^{K^p}$ by λ_x. Since W is regular, we know how to associate a Galois representation $\rho_{\pi,\iota}$ with the usual local properties, which must be ρ_x by Tchebotarev.

For (2), we follow the same line of argument. Since $x \in X_{\mathrm{cl}}$, there is an automorphic π contributing to $\mathcal{B}_x^{W_x\text{-alg}}$. Moreover, if $\mathcal{B}_x^{W\text{-alg}} \neq 0$, there is an automorphic π, of regular weight W, for which $\rho_{\pi,\iota} \simeq \rho_x$. From $\rho_{\pi,\iota}$ we can recover W through its Hodge–Tate numbers, and similarly for ρ_x, even if x is not in X_{reg} (this is shown in Section 3.15 of [Chenevier 2009], based on results of Sen, Berger and Colmez). Therefore, $W = W_x$. □

Remark. As remarked earlier, we are optimistic that one can remove the regularity hypotheses in the theorem. Indeed it seems possible to attach Galois representations to automorphic π of $U(\mathbb{A}_F)$ of *irregular* weight. When π_p is of finite slope (that is, embeds in a principal series), this can be done by means of eigenvarieties, as in [Chenevier 2009]. In general, it seems likely that one can push the ideas from the proof of Theorem 1. By [White 2012], there is always a base change $\boxplus_{i=1}^t \Pi_i$, where the Π_i are *discrete* automorphic representations of $\mathrm{GL}_{n_i}(\mathbb{A}_{\mathcal{K}})$, which in turn (by the Moeglin–Waldspurger classification) are isobaric sums of cohomological, essentially conjugate, self-dual cusp forms, with which one can associate Galois representations. Local–global compatibility at p follows from [Caraiani 2012].

6. Compatibility with classical local Langlands

In this section we deduce from our previous results that $\widetilde{\mathrm{BS}}(\rho_x)$ admits an invariant norm such that the completion $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ satisfies local–global compatibility (see Corollary 19 below for the precise statement). However, we *cannot* show that this completion $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ only depends on the restrictions of ρ_x to places above p — which seems to be an extremely difficult problem at the heart of the p-adic Langlands program. In Section 6.2 we will restrict ourselves to the *unramified* case, and prove a "weak" version of local–global compatibility (somewhat similar to part (1) of Theorem 1.2.1 in [Emerton 2011]) — it is "weak" since we only get a nonzero map (with a huge kernel) instead of an embedding. We refer the reader to part (b) of Theorem 21: the p-adic local Langlands correspondence, still mysterious in higher rank, is replaced by the coarse version in [Schneider and Teitelbaum 2006], which associates a huge Banach representation $B_{\xi,\zeta}$ with a pair (ξ,ζ) satisfying the Emerton condition (here ξ is an irreducible algebraic representation, and ζ is a suitable Weyl-orbit in the dual torus). The philosophy propounded in [Schneider and Teitelbaum 2006; Breuil and Schneider 2007] is that the (almost) quotients of $B_{\xi,\zeta}$ should somehow correspond to the crystalline representations of type (ξ,ζ). This is well understood for $\mathrm{GL}_2(\mathbb{Q}_p)$, where the admissible filtration is usually unique (see Theorem 2.3.2, p. 8, of Berger's [2011] survey), and $B_{\xi,\zeta}$ essentially *is* the local p-adic Langlands correspondence in the (irreducible) crystalline case. We provide evidence supporting this philosophy of Breuil, Schneider, and Teitelbaum for $n > 2$.

6.1. *Completions of the space of algebraic vectors.*

Split ramification and the automorphic representation π_x. Throughout, we will make the assumption that we have *split ramification*. That is, $S(K^p) \subset \mathrm{Spl}_{\mathcal{K}|F}$. This has the effect that the local base change $\mathrm{BC}_{w|v}$ is defined at *all* places v. We fix a point $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}}$, as above. Under our ramification hypothesis, there is a *unique* automorphic representation π of $U(\mathbb{A}_F)$ contributing to the (regular) algebraic vectors $\mathcal{B}_x^{\mathrm{ralg}}$ in Theorem 18(3). Indeed, any such π has an irreducible Galois representation $\rho_{\pi,\iota} \simeq \rho_x$, and therefore $\mathrm{BC}_{\mathcal{K}|F}(\pi)$ must be cuspidal, and it is uniquely determined at the infinite places, and away from $\Sigma(K^p)$. By strong multiplicity one for GL_n, the base change is unique. Locally, $\mathrm{BC}_{w|v}$ is injective (see Corollary 4.2 in [Mínguez 2011]), and therefore π is uniquely determined. We denote it $\pi_x = \bigotimes \pi_{x,v}$. Its local components $\pi_{x,v}$ are given by

$$\mathrm{WD}(\rho_x|_{\Gamma_{\mathcal{K}_w}})^{\text{F-ss}} \simeq \mathrm{rec}(\mathrm{BC}_{w|v}(\pi_{x,v}) \otimes |\det|_w^{(1-n)/2}).$$

We think of $\{\pi_x\}$ as a family of automorphic representations interpolated by $\mathbb{X}$. In general (without split ramification) the π_x will be L-packets, not singletons.

With this notation, part (3) of Theorem 18 becomes: for all $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}}$,

$$\mathcal{B}_x^{\mathrm{ralg}} \simeq \widetilde{\mathrm{BS}}(\rho_x) \otimes \left(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v} \right)^{m(\pi_x)}.$$

Most likely, $m(\pi_x) = 1$, and this may already be in the literature. However, we have not been able to find a suitable reference. Now, since $\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}$ is a simple $\mathcal{H}(K^p)$-module, we may think of $\widetilde{\mathrm{BS}}(\rho_x)^{m(\pi_x)}$ as its multiplicity space in $\mathcal{B}_x^{\mathrm{ralg}}$,

$$\widetilde{\mathrm{BS}}(\rho_x)^{m(\pi_x)} \xrightarrow{\sim} \mathrm{Hom}_{\mathcal{H}(K^p)}\left(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}, \mathcal{B}_x^{\mathrm{ralg}} \right),$$

as representations of $G(\mathbb{Q}_p)$. We will view the right-hand side as sitting inside a Banach space of continuous transformations. For that purpose, we first look at each local component $\pi_{x,v}$, where $v \nmid p$. When v splits, it can be identified with a p-integral irreducible representation of $\mathrm{GL}_n(F_v)$. By Theorem 1 in [Vignéras 2004], it has a unique commensurability class of stable lattices. Correspondingly, $\pi_{x,v}$ has a unique equivalence class of $\mathrm{GL}_n(F_v)$-invariant norms $\|\cdot\|_v$. (By Theorem 1 in [Vignéras 2010], the completion $\hat{\pi}_{x,v}$ is a topologically irreducible unitary Banach space representation of $\mathrm{GL}_n(F_v)$.) When $\pi_{x,v}$ is unramified, its Satake parameters are p-units, and one easily finds a stable lattice in a suitable unramified principal series, again resulting in a $U(F_v)$-invariant (supremum) norm $\|\cdot\|_v$, which we may normalize so that a given spherical vector has norm one. The tensor product norm (see Proposition 17.4 in [Schneider 2002]) on $\bigotimes_{v \nmid p} \pi_{x,v}$ is then invariant under $G(\mathbb{A}_f^p)$. By restriction, the finite-dimensional space $\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}$ inherits a norm, and becomes a Banach-module for $\widehat{\mathcal{H}}(K^p)$. With this extra structure at hand,

$$\mathrm{Hom}_{\mathcal{H}(K^p)}\left(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}, \mathcal{B}_x^{\mathrm{ralg}} \right) \hookrightarrow \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}\left(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}, \mathcal{B}_x \right).$$

(Here $\mathcal{L}$ denotes the space of continuous linear transformations equipped with the usual transformation norm; see Corollary 3.2 in [Schneider 2002].) We have to check that any $\mathcal{H}(K^p)$-equivariant map $\bigotimes_{v \nmid p} \pi_{x,v}^{K_v} \xrightarrow{\phi} \mathcal{B}_x$ is automatically continuous. If $\phi \neq 0$, it must be injective (by simplicity), and thus $\|\phi(\cdot)\|_{\mathcal{B}_x}$ defines a norm on $\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}$. However, all norms on a finite-dimensional space are equivalent (Proposition 4.13 in [Schneider 2002]), so that $\|\phi(u)\|_{\mathcal{B}_x} \leq C\|u\|$ for some constant $C > 0$ and all u. Altogether, this embeds $\widetilde{\mathrm{BS}}(\rho_x)$ into a Banach space (Proposition 3.3 in [Schneider 2002]):

$$\widetilde{\mathrm{BS}}(\rho_x)^{m(\pi_x)} \hookrightarrow \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}\left(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}, \mathcal{B}_x \right). \tag{1}$$

If we restrict the transformation norm to $\widetilde{\mathrm{BS}}(\rho_x)^{m(\pi_x)}$, we arrive at:

Corollary 19. *Let $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}}$ be a point such that $m(\pi_x) = 1$. Then there is a $G(\mathbb{Q}_p)$-invariant norm $\|\cdot\|$ on $\widetilde{\mathrm{BS}}(\rho_x)$ such that the corresponding completion $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ satisfies the following: There is a topological isomorphism*

$$\widetilde{\mathrm{BS}}(\rho_x)^\wedge \otimes \Bigl(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}\Bigr) \xrightarrow{\sim} \overline{\mathcal{B}_x^{\mathrm{ralg}}},$$

where $\overline{\mathcal{B}_x^{\mathrm{ralg}}}$ is the closure of the regular-algebraic vectors $\mathcal{B}_x^{\mathrm{ralg}}$ in $\mathcal{B}_x$. Moreover:

- *$\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ is an* admissible *unitary Banach space representation of $G(\mathbb{Q}_p)$.*
- *Its regular-algebraic vectors $\widetilde{\mathrm{BS}}(\rho_x)$ form a* dense *subspace.*

Proof. We obtain $\|\cdot\|$ by restricting the transformation norm to $\widetilde{\mathrm{BS}}(\rho_x)$. Thus (1) becomes an isometry, and extends uniquely to an isometry

$$\widetilde{\mathrm{BS}}(\rho_x)^\wedge \hookrightarrow \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}\Bigl(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}, \mathcal{B}_x\Bigr).$$

To ease the notation, let us write $M = \bigotimes_{v \nmid p} \pi_{x,v}^{K_v}$ throughout this proof; M is a finite-dimensional simple $\mathcal{H}(K^p)$-module. We tensor the isometry by this M:

$$j : \widetilde{\mathrm{BS}}(\rho_x)^\wedge \otimes M \hookrightarrow \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}(M, \mathcal{B}_x) \otimes M \xrightarrow{\sim} \mathcal{B}_x[M].$$

Here $\mathcal{B}_x[M]$ denotes the closure of the sum of all closed $\mathcal{H}(K^p)$-submodules of $\mathcal{B}_x$ isomorphic to M (a topological direct sum of a subcollection, by Zorn). Note that $\mathrm{End}_{\mathcal{H}(K^p)}(M) = \overline{\mathbb{Q}}_p$. Note also that the tensor products (equipped with their tensor product norms, as on p. 110 in [Schneider 2002]), are already complete, as M is finite-dimensional. The above isomorphism with $\mathcal{B}_x[M]$ is a *topological* isomorphism by the open mapping theorem (Proposition 8.6 in [Schneider 2002]), but not necessarily isometric. Consequently, $\mathrm{im}(j) \subset \mathcal{B}_x$ is a closed subspace, containing $\mathcal{B}_x^{\mathrm{ralg}}$ by Theorem 18. In fact, $\mathrm{im}(j)$ is the closure of $\mathcal{B}_x^{\mathrm{ralg}}$ in $\mathcal{B}_x$, since $\widetilde{\mathrm{BS}}(\rho_x)$ is dense in the completion $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$. Again invoke the open mapping theorem to see that j is a topological isomorphism onto $\overline{\mathcal{B}_x^{\mathrm{ralg}}}$. Admissibility of $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ follows from admissibility of $\mathcal{B}_x$. □

Remark. Equivalently, there is a $G(\mathbb{Q}_p)$-equivariant topological isomorphism

$$\widetilde{\mathrm{BS}}(\rho_x)^\wedge \xrightarrow{\sim} \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}\Bigl(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v}, \overline{\mathcal{B}_x^{\mathrm{ralg}}}\Bigr).$$

We like to think of this Banach space representation $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ as a rough candidate for a p-adic local Langlands correspondence, at least when the various restrictions $\rho_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}}$ are irreducible. Of course, to really justify this point of view, one would need to show that the completion $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ only depends on the restrictions $\rho_x|_{\Gamma_{\mathcal{K}_{\tilde{v}}}}$

at p, and that it factors as a tensor product $\widehat{\bigotimes}_{v|p}$ of appropriate completions $\widetilde{\mathrm{BS}}(\rho_x|\Gamma_{\mathcal{K}_{\tilde{v}}})^\wedge$. Both appear to be very difficult questions.

6.2. ***Universal modules and the crystalline case.*** We now specialize to the *crystalline* case, where we can relate $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ to the Schneider–Teitelbaum universal modules $B_{\xi,\zeta}$, which are given by a purely local construction at p. They are expected to be quite large. However, for $n > 2$ it is not even known that $B_{\xi,\zeta} \neq 0$ (Conjecture 6.1, p. 24 in [Breuil and Schneider 2007]). For $n = 2$ this is a deep result of Berger and Breuil. We will prove nonvanishing when (ξ, ζ) "comes from an eigenvariety". This will be a by-product of a stronger result.

Definition 20. A classical point $x \in X_{\mathrm{cl}}$ is called *old* if ρ_x is crystalline at all places above p. That is, $\mathrm{Hom}_{G(\mathbb{Z}_p)}(W_x, \mathcal{B}_x) \neq 0$. Equivalently, $\pi_{x,v}$ is unramified for all $v|p$. We denote the set of old points by X_{old}.

Thus, from now on, we fix a point $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}} \cap X_{\mathrm{old}}$. By Proposition 6,

$$\widetilde{\mathrm{BS}}(\rho_x) = W_x \otimes \pi_{x,p} \xrightarrow{\sim} W_x \otimes \mathrm{Ind}_B^G(\theta_x),$$

where θ_x is unramified smooth. (Indeed, for any point x, $\pi_{x,p}$ embeds into the (unnormalized) principal series $\mathrm{Ind}_B^G(\theta_x)$. Since x is old, $\pi_{x,p}$ is unramified, and hence so is θ_x. Furthermore, as $x \in X_{\mathrm{irr}}$, the base change $\mathrm{BC}_{\mathcal{K}|F}(\pi_x)$ is cuspidal, and therefore generic. In particular, $\pi_{x,p}$ must be generic. As is well known, this implies that $\pi_{x,p}$ must be the full unramified principal series.)

As in [Schneider and Teitelbaum 2006], we express $\pi_{x,p} \simeq \mathrm{Ind}_B^G(\theta_x)$ in terms of the *universal module*. This goes back to Borel and Matsumoto, and is defined as follows. For any algebra character $\zeta : \mathcal{H}(G, K) \to \overline{\mathbb{Q}}_p$ (where $K = G(\mathbb{Z}_p)$ is hyperspecial when p is assumed to be unramified in F) we introduce the smooth representation

$$\mathcal{M}_\zeta = \text{c-Ind}_K^G(1) \otimes_{\mathcal{H}(G,K),\zeta} \overline{\mathbb{Q}}_p = \mathcal{C}_c(K\backslash G, \overline{\mathbb{Q}}_p) \otimes_{\mathcal{H}(G,K),\zeta} \overline{\mathbb{Q}}_p.$$

The pair $(\mathcal{M}_\zeta, 1_K)$ is a universal initial object in the category of pairs (V, v), where V is an unramified smooth representation of $G(\mathbb{Q}_p)$, and $v \in V^K$ is a nonzero vector on which $\mathcal{H}(G, K)$ acts via ζ. That is, there is a unique $G(\mathbb{Q}_p)$-map $\mathcal{M}_\zeta \to V$ which maps $1_K \mapsto v$. The image of this map is the span of the orbit Gv (since $\mathcal{M}_\zeta$ is generated by 1_K). In what follows we will take $\zeta_x = \widehat{\theta}_x$, the eigensystem of $\mathrm{Ind}_B^G(\theta_x)^K$. The choice of a spherical vector yields

$$\mathcal{M}_{\zeta_x} \to \mathrm{Ind}_B^G(\theta_x), \quad \zeta_x = \widehat{\theta}_x.$$

It is a general fact that the two representations have the same semisimplification (see the Ph.D. thesis of X. Lazarus [2000] for a thorough discussion in greater generality). Under our assumptions, $\mathrm{Ind}_B^G(\theta_x)$ is irreducible, and therefore the above must be an isomorphism. Consequently, we may identify

$$\widetilde{\mathrm{BS}}(\rho_x) \simeq W_x \otimes \mathcal{M}_{\zeta_x} \simeq \text{c-Ind}_K^G(\xi_x) \otimes_{\mathcal{H}_{\xi_x}(G,K),\zeta_x} \overline{\mathbb{Q}}_p =: H_{\xi_x,\zeta_x}.$$

Here we have changed notation $\xi_x := W_x$ to aid comparison with [Schneider and Teitelbaum 2006]. The algebra $\mathcal{H}_{\xi_x}(G, K)$ consists, by definition, of the G-endomorphisms of $\text{c-Ind}_K^G(\xi_x)$; or, more concretely, of compactly supported K-biequivariant functions $G \to \mathrm{End}(\xi_x)$ with convolution. However, since ξ is an irreducible representation of G (viewed as a representation of K), as on p. 639 in [Schneider and Teitelbaum 2006] one can identify the algebras

$$\mathcal{H}(G, K) \xrightarrow{\sim} \mathcal{H}_{\xi_x}(G, K), \quad h \mapsto (g \mapsto h(g)\xi_x(g)).$$

In the definition of H_{ξ_x,ζ_x} we view ζ_x as a character of $\mathcal{H}_{\xi_x}(G, K)$ via this isomorphism, as at the bottom of p. 670 in [Schneider and Teitelbaum 2006], where $H_{\xi,\zeta}$ is defined.

The representation H_{ξ_x,ζ_x} has a natural locally convex topology, being a quotient of $\text{c-Ind}_K^G(\xi_x)$, which has a supremum norm: Pick any norm $\|\cdot\|_{\xi_x}$ on ξ_x which is invariant under (the compact group) K. They are all equivalent since ξ_x is finite-dimensional (Proposition 4.13 in [Schneider 2002]). For $f \in \text{c-Ind}_K^G(\xi_x)$, we let

$$\|f\|_{\xi_x,\infty} = \sup_{g\in G(\mathbb{Q}_p)} \|f(g)\|_{\xi_x} < \infty$$

This defines an norm $\|\cdot\|_{\xi_x,\infty}$ on the compact induction, which is obviously invariant under $G(\mathbb{Q}_p)$, and it induces a quotient *seminorm* on the representation

$$H_{\xi_x,\zeta_x} = (\text{c-Ind}_K^G(\xi_x))/(\ker \zeta_x)(\text{c-Ind}_K^G(\xi_x)).$$

We will show below that in fact this is a *norm*, but this is far from clear a priori!

Following [Schneider and Teitelbaum 2006], on p. 671 where they define $B_{\xi,\zeta}$, we introduce the space

$$B_{\xi_x,\zeta_x} := \widehat{H}_{\xi_x,\zeta_x} = (H_{\xi_x,\zeta_x}/\overline{\{0\}})^\wedge,$$

the Hausdorff completion of H_{ξ_x,ζ_x}. (We refer to Proposition 7.5 in [Schneider 2002] for a general discussion of Hausdorff completions.) We have defined a Banach space B_{ξ_x,ζ_x} with a unitary $G(\mathbb{Q}_p)$-action. However, it is not clear at all that it is *nonzero*. This is in fact a fundamental problem! Conjecture 6.1 on p. 24 in [Breuil and Schneider 2007] says that $B_{\xi,\zeta} \neq 0$ whenever the Emerton condition is satisfied (the converse is known). This follows from our methods when the pair (ξ, ζ) comes from an eigenvariety, that is, when it is of the form (ξ_x, ζ_x) for an old irreducible point x. What we prove is a strengthening:

Theorem 21. *Let $x \in X_{\mathrm{reg}} \cap X_{\mathrm{irr}} \cap X_{\mathrm{old}}$ be a classical point such that $m(\pi_x) = 1$. Then H_{ξ_x,ζ_x} is Hausdorff and $B_{\xi_x,\zeta_x} \neq 0$ is its universal completion. Furthermore*:

(a) *There is a continuous map, with dense image, $B_{\xi_x,\zeta_x} \to \widetilde{\mathrm{BS}}(\rho_x)^\wedge$ (into the completion from Corollary 19) which restricts to an isomorphism $H_{\xi_x,\zeta_x} \xrightarrow{\sim} \widetilde{\mathrm{BS}}(\rho_x)$ onto the regular-algebraic vectors.*

(b) *There is a* nonzero *$G(\mathbb{Q}_p) \times \widehat{\mathcal{H}}(K^p)$-equivariant continuous map*

$$B_{\xi_x,\zeta_x} \otimes \left(\bigotimes_{v \nmid p} \pi_{x,v}^{K_v} \right) \to \overline{\mathcal{B}_x^{\mathrm{ralg}}}$$

with dense image.

Proof. From (2) of the previous section, we have a $G(\mathbb{Q}_p)$-embedding

$$H_{\xi_x,\zeta_x} \simeq \widetilde{\mathrm{BS}}(\rho_x) \hookrightarrow \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}(M, \mathcal{B}_x),$$

where we keep writing $M = \bigotimes_{v \nmid p} \pi_{x,v}^{K_v}$. We claim this map is automatically continuous when we equip the $\mathcal{L}$-space with the transformation norm and H_{ξ_x,ζ_x} with the quotient seminorm induced by $\|\cdot\|_{\xi_x,\infty}$. Since H_{ξ_x,ζ_x} gets the quotient topology, we just have to check continuity of the inflated map

$$\text{c-Ind}_K^G(\xi_x) \twoheadrightarrow H_{\xi_x,\zeta_x} \hookrightarrow \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}(M, \mathcal{B}_x).$$

This is simply Frobenius reciprocity made explicit. In particular, the seminorm on H_{ξ_x,ζ_x} is actually a norm (as the kernel of the above map is closed). Therefore, H_{ξ_x,ζ_x} is Hausdorff, and B_{ξ_x,ζ_x} is its universal completion. That is, there is an isometry with dense image,

$$H_{\xi_x,\zeta_x} \hookrightarrow B_{\xi_x,\zeta_x}$$

(so B_{ξ_x,ζ_x} is nonzero). By continuity of the initial map, it has a unique extension

$$B_{\xi_x,\zeta_x} \to \mathcal{L}_{\widehat{\mathcal{H}}(K^p)}(M, \mathcal{B}_x),$$

which is continuous (but not necessarily injective) and maps into the completion $\widetilde{\mathrm{BS}}(\rho_x)^\wedge$ from Corollary 19, with dense image (but not necessarily onto). □

Remark. This fits perfectly with the picture suggested in the papers [Schneider and Teitelbaum 2006; Breuil and Schneider 2007]. If there is a local p-adic Langlands correspondence $\rho \mapsto \mathfrak{B}(\rho)$, these references speculate that $B_{\xi,\zeta}$ maps to each $\mathfrak{B}(\rho)$, with dense image, for all crystalline representations ρ of type (ξ, ζ).

7. Zariski density of crystalline points

In general, it is not expected that $\mathcal{B}_x^{\mathrm{alg}}$ is dense in $\mathcal{B}_x$. In this section, we will adapt (and elaborate on) an argument from Sections 5.3 and 5.4 in [Emerton 2011], which builds on ideas of Katz — and which shows the density of crystalline points. This is not needed for this paper, but we include it here for future reference.

7.1. ***Injectivity of certain modules.*** We fix a finite extension $L|\mathbb{Q}_p$, and we will write $\mathcal{O} = \mathcal{O}_L$ and $\varpi = \varpi_L$, and so on. We will look at locally constant functions $f : Y(K^p) \to A$ taking values in various finite $\mathcal{O}$-modules $A = \mathcal{O}/\varpi^s\mathcal{O}$, where s is a positive integer. These functions form a (discrete) torsion $\mathcal{O}$-module, denoted $H^0(K^p, A)$, carrying a natural action of the Hecke algebra $\mathcal{H}_{\mathcal{O}}(K^p)$, and a commuting smooth $G(\mathbb{Q}_p)$-action, which is *admissible* in the following sense: for every compact open subgroup of $G(\mathbb{Q}_p)$, its invariants form a finite $\mathcal{O}$-module (torsion and finitely generated means finite cardinality, since A is a finite ring).

Lemma 22. *Suppose K^p is sufficiently small (for example, it suffices that K_v has no p-torsion for some $v \nmid p$). Then, for any compact open subgroup $K_p \subset G(\mathbb{Q}_p)$,*

$$H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O}) \textit{ is an } \text{injective} \textit{ smooth } (\mathcal{O}/\varpi^s\mathcal{O})[K_p]\textit{-module for all } s \geq 1.$$

Consequently, every direct summand of $H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O})$ is an injective[4] *module.*

Proof. We have to show the exactness of the functor sending a module M to

$$\mathrm{Hom}_{\mathcal{O}[K_p]}(M, H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O})).$$

Here M is an $\mathcal{O}[K_p]$-module with $\varpi^s M = 0$. Therefore, it has Pontryagin dual

$$M^\vee = \mathrm{Hom}_{\mathcal{O}}(M, L/\mathcal{O}) = \mathrm{Hom}_{\mathcal{O}}(M, \varpi^{-s}\mathcal{O}/\mathcal{O}) \simeq \mathrm{Hom}_{\mathcal{O}/\varpi^s\mathcal{O}}(M, \mathcal{O}/\varpi^s\mathcal{O}).$$

(Here M is smooth, so we equip it with the discrete topology.) The initial module above can then be identified with that consisting of all functions

$$f : Y(K^p) \to M^\vee, \quad f(gk) = k^{-1}f(g)$$

for $k \in K_p$. Choosing representatives $g_i \in G(\mathbb{A}_f)$ for the finite set $Y(K_pK^p)$, and mapping f to the tuple of all $f(g_i)$, then identifies the latter with the direct sum $\bigoplus_i (M^\vee)^{\Gamma_i}$, where the Γ_i are certain finite subgroups of K_p having prime-to-p order by assumption. This ensures that $(\cdot)^{\Gamma_i}$ is exact, by averaging. Also, taking the Pontryagin dual is exact ($L/\mathcal{O}$ is divisible). Finally, as is well known (and easy to check) every summand of an injective module is itself injective. □

Examples. Let us first introduce certain finite-type Hecke algebras. For each K_p, we let $\mathbb{T}(K_pK^p)$ denote the image of $\mathfrak{h}^\circ = \mathcal{H}_{\mathcal{O}}(K^p)^{\mathrm{sph}}$ in the endomorphism algebra $\mathrm{End}_{\mathcal{O}} H^0(Y(K_pK^p), \mathcal{O})$. Thus $\mathbb{T}(K_pK^p)$ is finite free over (the PID) $\mathcal{O}$, and we endow it with the ϖ-adic topology. If we have a subgroup $K_p' \subset K_p$, there

[4] The injectivity (of modules) addressed here should not be confused with the injectivity of the maps discussed in the remark on p. 209.

is a natural restriction map $\mathbb{T}(K_p'K^p) \to \mathbb{T}(K_pK^p)$, and we take the limit

$$\mathbb{T}(K^p) := \varprojlim_{K_p} \mathbb{T}(K_pK^p) \subset \operatorname{End}_{\mathcal{O}} \tilde{H}^0(K^p)^\circ,$$

the closure of the image of $\mathcal{H}_{\mathcal{O}}(K^p)^{\mathrm{sph}}$. This defines a reduced, commutative, complete, topological $\mathcal{O}$-algebra. Moreover, $\mathbb{T}(K^p)$ has only *finitely* many[5] maximal ideals: they correspond to the maximal ideals of $\mathbb{T}(K^p) \otimes \mathbb{F}$, which is the image of $\mathfrak{h}^\circ$ in $\operatorname{End}_{\mathbb{F}} H^0(K^p, \mathbb{F})$. Hence, the maximal ideals are in bijection with the (Galois conjugacy classes of) eigensystems $\mathfrak{h}^\circ \to \mathbb{F}$ which occur in $H^0(K^p, \mathbb{F})$. If K_p is any pro-p group, they all must occur in $H^0(Y(K_pK^p), \mathbb{F})$, which is finite-dimensional. Therefore, since $\mathcal{O}$ is complete, we have

$$\mathbb{T}(K^p) \xrightarrow{\sim} \prod_{\mathfrak{m}} \mathbb{T}(K^p)_{\mathfrak{m}},$$

where the product extends over the finitely many maximal ideals $\mathfrak{m} \subset \mathbb{T}(K^p)$, and $\mathbb{T}(K^p)_{\mathfrak{m}}$ denotes the corresponding localization, a complete local $\mathcal{O}$-algebra. (We refer to Chapter 4 of [Darmon et al. 1997] for a discussion of the commutative algebra needed.) We will use this product decomposition as follows: Obviously,

$$\tilde{H}^0(K^p)^\circ/\varpi^s\tilde{H}^0(K^p)^\circ \simeq H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O})$$

carries an action of $\mathbb{T}(K^p)$. This gives rise to a *direct sum* decomposition,

$$H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O}) \xrightarrow{\sim} \bigoplus_{\mathfrak{m}} H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O})_{\mathfrak{m}},$$

into localized smooth admissible $G(\mathbb{Q}_p)$-submodules over $\mathcal{O}/\varpi^s\mathcal{O}$

$$H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O})_{\mathfrak{m}} := H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O}) \otimes_{\mathbb{T}(K^p)} \mathbb{T}(K^p)_{\mathfrak{m}},$$

which are then *injective* $(\mathcal{O}/\varpi^s\mathcal{O})[K_p]$-modules for every compact open K_p.

To connect this to the previous discussion, one could take the maximal ideal $\mathfrak{m}_x = \ker(\bar{\lambda}_x)$ for a point $x \in \mathbb{X}(L)$. A priori, this is a maximal ideal in $\mathfrak{h}^\circ$, but it is the pull-back of an ideal $\mathfrak{m} \subset \mathbb{T}(K^p)$, since $\bar{\lambda}_x$ occurs in tame level K^p.

7.2. ***Projective modules over certain Iwasawa algebras.*** To simplify notation, we write $A = \mathcal{O}/\varpi^s\mathcal{O}$ in this section, where $s > 0$ is fixed for the moment. We will briefly recall known facts about the Pontryagin duality functor $M \mapsto M^\vee$, which sends a discrete $A[K_p]$-module M to the compact

$$M^\vee = \operatorname{Hom}_{\mathcal{O}}(M, L/\mathcal{O}) \simeq \operatorname{Hom}_A(M, A).$$

[5] The finiteness of the number of cohomological mod p Hecke eigensystems has been proved in much greater generality by Ash and Stevens [1986].

If M is smooth, $M = \varinjlim_H M^H$, with H running over normal open subgroups of K_p, and therefore its dual $M^\vee = \varprojlim_H (M^H)^\vee$ becomes a module for the *Iwasawa algebra*

$$A[\![K_p]\!] := \varprojlim_H A[K_p/H].$$

Conversely, if X is an $A[\![K_p]\!]$-module, $X/I_H X$ becomes a module for $A[K_p/H]$, where I_H is the kernel of the natural projection $A[\![K_p]\!] \twoheadrightarrow A[K_p/H]$. It follows that $X^\vee$ is again a smooth $A[K_p]$-module, since

$$(X^\vee)^H \simeq (X/I_H X)^\vee.$$

Thus, duality sets up a one-to-one correspondence $M \leftrightarrow X$ between smooth discrete $A[K_p]$-modules and compact $A[\![K_p]\!]$-modules which reverses arrows.

Lemma 23. *Suppose M is a smooth $A[K_p]$-module, with Pontryagin dual $M^\vee$.*

(i) *M is admissible $\iff$ $M^\vee$ is finitely generated over $A[\![K_p]\!]$.*

(ii) *M is injective $\iff$ $M^\vee$ is a projective $A[\![K_p]\!]$-module.*

Proof. For part (i), if X is finitely generated over $A[\![K_p]\!]$, we deduce that $X/I_H X$ is finitely generated over $A[K_p/H]$, which is a ring of finite cardinality. Therefore its dual M^H is (physically) finite. For the converse, suppose M is admissible. Then, first of all, $M^\vee$ is profinite, so we may apply the "converse" (topological) Nakayama lemma discussed in depth in [Balister and Howson 1997] (specifically, their main theorem in Chapter 3, Section (1), and its corollary): to verify that $M^\vee$ is finitely generated over the compact ring $A[\![K_p]\!]$, it suffices to check that $X/I_H X$ is finitely generated over $A[K_p/H]$ for *some* H such that $I_H^n \to 0$ as $n \to \infty$. This limit holds for any pro-p-group H; see Lemma 3.2 in [Schneider and Teitelbaum 2002], for example. Finiteness of $X/I_H X$, or rather its dual M^H, is admissibility.

For part (ii), use that Pontryagin duality is exact (divisibility of $L/\mathcal{O}$). It follows that $\mathrm{Hom}_{A[K_p]}(-, M)$ is exact if and only if $\mathrm{Hom}_{A[\![K_p]\!]}(M^\vee, -)$ is exact. □

From the last two lemmas, we immediately conclude the following:

Proposition 24. *Suppose K^p is sufficiently small. Then, for any compact open subgroup $K_p \subset G(\mathbb{Q}_p)$, the dual $H^0(K^p, A)^\vee$ is a* projective *finitely generated module over $A[\![K_p]\!]$ for all $s \geq 1$. The same is true for any direct summand, such as the localized module $H^0(K^p, A)^\vee_{\mathfrak{m}}$ for any maximal ideal $\mathfrak{m}$.*

For later use, we will record the following fact here. Often, the Iwasawa algebra $A[\![K_p]\!]$ is viewed as a distribution algebra. Indeed, there is a natural pairing with the continuous (that is, locally constant) functions $\mathcal{C}(K_p, A)$.

Lemma 25. $A[\![K_p]\!] \xrightarrow{\sim} \mathcal{C}(K_p, A)^\vee$, *as modules over* $A[\![K_p]\!]$.

Proof. For any normal open subgroup H, there is a canonical integration pairing,

$$\mathcal{C}(K_p/H, A) \times A[K_p/H] \to A, \quad (f, \mu) \mapsto \sum\nolimits_{k \in K_p/H} f(k)\mu(k),$$

which is nondegenerate, and therefore defines an isomorphism

$$A[K_p/H] \xrightarrow{\sim} \mathcal{C}(K_p/H, A)^\vee, \quad \mu \mapsto (-, \mu).$$

This is easily checked to preserve the $A[K_p/H]$-module structures on both sides. Moreover, as H varies, these isomorphisms are compatible with the transition maps. Passing to the projective limit $\varprojlim_H$ gives the lemma. □

In other words, $\mathcal{C}(K_p, A) \leftrightarrow A[\![K_p]\!]$ under the correspondence discussed above.

7.3. ***Local Iwasawa algebras of pro-p-groups.*** A local ring is a (possibly noncommutative) ring R whose Jacobson radical $J(R)$ is a two-sided maximal ideal $\mathfrak{m}_R$. In other words, there is a unique maximal left ideal, and a unique maximal right ideal, and they coincide. Nakayama's lemma even holds for noncommutative local rings, as is easily checked. In particular, a finitely generated projective R-module is free, a key fact we will make use of below, by taking R to be the Iwasawa algebra of a pro-p-group, which turns out to be local. We first assemble the following well-known facts.

Lemma 26. *Let K_p be a pro-p-group, and let A be any p-ring* (*that is, its cardinality is a finite power of p, such as for $A = \mathcal{O}/\varpi^s\mathcal{O}$*). *Then*:

(1) *Let M be a left $A[\![K_p]\!]$-module, and $H \subset K_p$ an open normal subgroup. Then $M/I_H M$ has a nonzero K_p-invariant element if $M \neq I_H M$.*

(2) *K_p acts trivially on any simple left $A[\![K_p]\!]$-module.*

(3) *$I_{K_p} \subset J(A[\![K_p]\!])$.*

(4) *A local $\Longrightarrow$ $A[\![K_p]\!]$ local.* (*Furthermore, $J(A[\![K_p]\!]) = \mathfrak{m}_A + I_{K_p}$.*)

(*The same is true when left modules are replaced by right modules.*)

Proof. This is all standard. We cannot resist briefly outlining the argument. For (1) it is clearly enough to show that a p-group K_p fixes a nonzero element of any $A[K_p]$-module $M \neq 0$. This is basic group action theory; the fact that A is a p-ring allows us to count fixed points modulo p. For (2), if M is a simple left $A[\![K_p]\!]$-module, we must have $I_H M = M$ or $I_H M = 0$ for all H. There must be some H for which $I_H M \neq M$, since $M \neq 0$ is the inverse limit of all quotients $M/I_H M$. Now (1) shows that $M^{K_p} \neq 0$. By simplicity, K_p acts trivially on M. For (3) just use that I_{K_p} is generated by elements $k - 1$ with $k \in K_p$. We see from (2) that I_{K_p} acts trivially on any simple left $A[\![K_p]\!]$-module, and therefore, by the very definition of the Jacobson radical, we have the inclusion as claimed. Now

(4) is immediate from (3). Indeed any maximal left ideal of $A[\![K_p]\!]$ must be the pull-back of $\mathfrak{m}_A$ under the augmentation map. □

Together with Proposition 24, we will apply this to $A[\![K_p]\!]$ with $A = \mathcal{O}/\varpi^s\mathcal{O}$.

Proposition 27. *Suppose K^p is sufficiently small, and let $K_p \subset G(\mathbb{Q}_p)$ be an open pro-p-group. Then there exists an integer $r > 0$ such that*

$$\widetilde{H}^0(K^p)^\circ \simeq \mathcal{C}(K_p, \mathcal{O})^r$$

as $\mathcal{O}[K_p]$-modules. Moreover, for any maximal ideal $\mathfrak{m} \subset \mathbb{T}(K^p)$, the localization $\widetilde{H}^0(K^p)^\circ_{\mathfrak{m}}$ sits as a topologically direct summand.

Proof. Since $A[\![K_p]\!]$ is local, Nakayama's lemma (and Proposition 24) tells us that $H^0(K^p, A)^\vee$ is a *free* $A[\![K_p]\!]$-module, of finite rank r_s, say. Taking the Pontryagin dual then yields an isomorphism of smooth $A[K_p]$-modules

$$H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O}) \simeq \mathcal{C}(K_p, \mathcal{O}/\varpi^s\mathcal{O})^{r_s}.$$

Now, we claim that r_s is in fact independent of $s > 0$ (and we will just write r instead of r_s). To see this, scale both sides of the isomorphism by ϖ, compare the corresponding quotients, take H-invariants for some H, and compare dimensions over $\mathbb{F}$. This shows that $r_s = r_1$. This allows us to take the inverse limit over s to obtain an isomorphism of modules over $\mathcal{O}[K_p]$

$$\widetilde{H}^0(K^p)^\circ = \varprojlim_s H^0(K^p, \mathcal{O}/\varpi^s\mathcal{O}) \simeq \mathcal{C}(K_p, \mathcal{O})^r,$$

or, in other words, an isometry $\widetilde{H}^0(K^p) \simeq \mathcal{C}(K_p, L)^r$ of Banach representations of K_p. Finally, we may localize at any maximal ideal $\mathfrak{m} \subset \mathbb{T}(K^p)$ and realize $\widetilde{H}^0(K^p)^\circ_{\mathfrak{m}}$ as a (topologically) direct summand of $\mathcal{C}(K_p, \mathcal{O})^r$. □

7.4. ***Mahler expansions and full level at p.*** Proposition 27 already shows that the algebraic vectors are dense in $\widetilde{H}^0(K^p)$ (by employing Mahler expansions, as below). In fact, this is even true for the unit ball $\widetilde{H}^0(K^p)^\circ$. However, we can be more precise, and prove density of the smaller set of $G(\mathbb{Z}_p)$-locally algebraic vectors, those $f \in \widetilde{H}^0(K^p)$ such that $\langle G(\mathbb{Z}_p)f\rangle$ is an algebraic representation of $G(\mathbb{Z}_p)$:

Proposition 28. *$\widetilde{H}^0(K^p)^{G(\mathbb{Z}_p)\text{-alg}}$ is dense in $\widetilde{H}^0(K^p)$ (similarly for $\widetilde{H}^0(K^p)_{\mathfrak{m}}$).*

Proof. Pick an open, normal pro-p-subgroup $K_p \subset G(\mathbb{Z}_p)$. From Proposition 27, we have an isometry $\widetilde{H}^0(K^p) \simeq \mathcal{C}(K_p, L)^r$ of Banach space representations of K_p. We take the topological dual space $\mathcal{L}(-, L)$ on both sides, and get

$$\widetilde{H}^0(K^p)^\vee \simeq L[\![K_p]\!]^r, \quad L[\![K_p]\!] := L \otimes_{\mathcal{O}} \mathcal{O}[\![K_p]\!].$$

Here $L[\![K_p]\!]$ is identified with the distribution algebra $\mathcal{C}(K_p, L)^\vee$ (equipped with the bounded-weak topology) as in [Schneider and Teitelbaum 2002], so $\widetilde{H}^0(K^p)^\vee$ is a free $L[\![K_p]\!]$-module of rank r. It follows that $\widetilde{H}^0(K^p)^\vee$ is projective over $L[\![G(\mathbb{Z}_p)]\!]$, as

$$\mathrm{Hom}_{L[\![G(\mathbb{Z}_p)]\!]}(\widetilde{H}^0(K^p)^\vee, -) = \mathrm{Hom}_{L[\![K_p]\!]}(\widetilde{H}^0(K^p)^\vee, -)^{G(\mathbb{Z}_p)/K_p}$$

is exact: $\widetilde{H}^0(K^p)^\vee$ is projective over $L[\![K_p]\!]$, and taking invariants under the finite group $G(\mathbb{Z}_p)/K_p$ is exact, by averaging (we are in characteristic zero). Being projective, $\widetilde{H}^0(K^p)^\vee$ is a direct summand of a free module (of finite rank by finite generation). That is, there is an $s > 0$, and a submodule Z, such that

$$\widetilde{H}^0(K^p)^\vee \oplus Z \simeq L[\![G(\mathbb{Z}_p)]\!]^s.$$

Again, undoing the dual, and invoking Corollary 2.2 and Theorem 3.5 in [Schneider and Teitelbaum 2002],

$$\widetilde{H}^0(K^p) \oplus Z^\vee \simeq \mathcal{C}(G(\mathbb{Z}_p), L)^s.$$

Comparing the $G(\mathbb{Z}_p)$-algebraic vectors on both sides, we see that it suffices to show that they are dense in $\mathcal{C}(G(\mathbb{Z}_p), L)$. Now, topologically, we identify $G(\mathbb{Z}_p) \simeq \prod_{v|p} \mathrm{GL}_n(\mathcal{O}_{\tilde{v}})$ with a closed–open subset of $\prod_{v|p} \mathcal{O}_{\tilde{v}}^{n^2} \simeq \mathbb{Z}_p^t$, where we have introduced $t = [F : \mathbb{Q}]n^2$. Any continuous function on $G(\mathbb{Z}_p)$ therefore extends (nonuniquely) to a continuous function on $\mathbb{Z}_p^t$, which has a (multivariable) Mahler power series expansion [1958], which shows that the polynomials are dense in $\mathcal{C}(\mathbb{Z}_p^t, L)$. Finally, observe that polynomials obviously restrict to $G(\mathbb{Z}_p)$-algebraic functions in $\mathcal{C}(G(\mathbb{Z}_p), L)$. At last, localize at $\mathfrak{m}$. □

7.5. ***Density and locally algebraic vectors.*** Following [Emerton 2011, Section 5.4], we deduce from the previous proposition that "crystalline points are dense".

Corollary 29. *The submodule $\bigoplus_{\lambda \in C} \widetilde{H}^0(K^p)^{\mathrm{alg}}[\lambda]$ is dense in $\widetilde{H}^0(K^p)$, where C denotes the collection of Hecke eigensystems $\lambda : \mathcal{H}(K^p)^{\mathrm{sph}} \to \overline{\mathbb{Q}}_p$ associated with an automorphic π, which is unramified at p (and of tame level K^p). Thus, the set of points* $\ker(\lambda)$, *with $\lambda \in C$, are Zariski dense in* $\mathrm{Spec}\,\mathbb{T}(K^p)[1/p]$.

Proof. First off, recall from Section 3.2 that we have a decomposition

$$\widetilde{H}^0(K^p)^{\mathrm{alg}} = \bigoplus_W W \otimes H^0(K^p, \mathcal{V}_{\check{W}}) = \bigoplus_W \bigoplus_{\pi : \pi_\infty \simeq W} m_G(\pi)(W \otimes \pi_f^{K^p}).$$

In particular,

$$\widetilde{H}^0(K^p)^{G(\mathbb{Z}_p)\text{-alg}} = \bigoplus_W \bigoplus_{\pi : \pi_\infty \simeq W} m_G(\pi)(W \otimes \pi_p^{G(\mathbb{Z}_p)} \otimes (\pi_f^p)^{K^p}),$$

which is dense in $\widetilde{H}^0(K^p)$. A fortiori, so is the $G(\mathbb{Q}_p)$-submodule it generates,

$$\langle \widetilde{H}^0(K^p)^{G(\mathbb{Z}_p)\text{-alg}}\rangle_{G(\mathbb{Q}_p)} = \bigoplus_W \bigoplus_{\substack{\pi:\pi_\infty \simeq W,\\ \pi_p^{G(\mathbb{Z}_p)} \neq 0}} m_G(\pi)(W \otimes \pi_f^{K^p}).$$

We decompose the latter into eigenspaces for the action $\mathcal{H}(K^p)^{\text{sph}}$. That is, as

$$\langle \widetilde{H}^0(K^p)^{G(\mathbb{Z}_p)\text{-alg}}\rangle_{G(\mathbb{Q}_p)} = \bigoplus_\lambda \widetilde{H}^0(K^p)^{\text{alg}}[\lambda]$$

where $\lambda : \mathcal{H}(K^p)^{\text{sph}} \to \overline{\mathbb{Q}}_p$ runs over all eigensystems of the form $\lambda = \lambda_\pi$ for some automorphic π, of tame level K^p, which is *unramified* at p (and of some weight W). Thus, elements of $\bigcap_{\lambda \in C} \ker(\lambda)$ act trivially on $\widetilde{H}^0(K^p)$. □

Acknowledgements

I wish to acknowledge the impact of the visions of Breuil and Emerton on this work. I would like to thank both of them for discreetly pointing out some embarrassing misconceptions of mine. Conversations and correspondence with Chojecki, Dospinescu, Helm, Herzig, Newton, Ramakrishnan, and Shin have been a great help and a source of inspiration. I am grateful to Mínguez for clarifying Corollary 4.2 of [Mínguez 2011] to me, and for pointing to the reference [Clozel et al. 2011]. Thanks are due to Caraiani, for making a preliminary version of [Caraiani 2012] available to me before making it public. Finally, I would like to convey my sincere thanks to the referee for a meticulous reading and thoughtful feedback.

References

[Ash and Stevens 1986] A. Ash and G. Stevens, "Cohomology of arithmetic groups and congruences between systems of Hecke eigenvalues", *J. Reine Angew. Math.* **365** (1986), 192–220. MR 87i:11069 Zbl 0596.10026

[Balister and Howson 1997] P. N. Balister and S. Howson, "Note on Nakayama's lemma for compact Λ-modules", *Asian J. Math.* **1**:2 (1997), 224–229. MR 99f:16047 Zbl 0904.16019

[Barnet-Lamb et al. 2014] T. Barnet-Lamb, T. Gee, D. Geraghty, and R. Taylor, "Local–global compatibility for $l = p$, II", *Ann. Sci. École Norm. Sup.* (4) **47**:1 (2014), 165–179. MR 3205603 Zbl 06324769

[Bellaïche and Chenevier 2009] J. Bellaïche and G. Chenevier, *Families of Galois representations and Selmer groups*, Astérisque **324**, Société Mathématique de France, Paris, 2009. MR 2011m:11105 Zbl 1192.11035

[Berger 2011] L. Berger, "La correspondance de Langlands locale p-adique pour $\mathrm{GL}_2(\mathbb{Q}_p)$", pp. 157–180 in *Séminaire Bourbaki: Vol. 2009/2010, Exposés 1012–1026*, Astérisque **339**, Société Mathématique de France, Paris, 2011. MR 2012k:11186

[Bergeron and Clozel 2005] N. Bergeron and L. Clozel, *Spectre automorphe des variétés hyperboliques et applications topologiques*, Astérisque **303**, Société Mathématique de France, Paris, 2005. MR 2007j:22031 Zbl 1098.11035

[Breuil 2010] C. Breuil, "The emerging p-adic Langlands programme", pp. 203–230 in *Proceedings of the International Congress of Mathematicians* (Hyderabad, 2010), vol. 2, edited by R. Bhatia et al., Hindustan Book Agency, New Delhi, 2010. MR 2012k:22024 Zbl 05971096

[Breuil and Schneider 2007] C. Breuil and P. Schneider, "First steps towards p-adic Langlands functoriality", *J. Reine Angew. Math.* **610** (2007), 149–180. MR 2009f:11147 Zbl 1180.11036

[Caraiani 2012] A. Caraiani, "Local–global compatibility and the action of monodromy on nearby cycles", *Duke Math. J.* **161**:12 (2012), 2311–2413. MR 2972460 Zbl 06095601

[Caraiani et al. 2014] A. Caraiani, M. Emerton, T. Gee, D. Geraghty, V. Paskunas, and S. W. Shin, "Patching and the p-adic local Langlands correspondence", preprint, 2014. arXiv 1310.0831

[Chenevier 2004] G. Chenevier, "Familles p-adiques de formes automorphes pour GL_n", *J. Reine Angew. Math.* **570** (2004), 143–217. MR 2006b:11046 Zbl 1093.11036

[Chenevier 2009] G. Chenevier, "Une application des variétés de Hecke des groupes unitaires", preprint, 2009, http://gaetan.chenevier.perso.math.cnrs.fr/articles/famgal.pdf.

[Clozel et al. 2011] L. Clozel, M. Harris, and J.-P. Labesse, "Endoscopic transfer", pp. 475–496 in *On the stabilization of the trace formula*, edited by L. Clozel et al., Stabilization of the Trace Formula, Shimura Varieties, and Arithmetic Applications **1**, International Press, Somerville, MA, 2011. MR 2856382

[Colmez and Fontaine 2000] P. Colmez and J.-M. Fontaine, "Construction des représentations p-adiques semi-stables", *Invent. Math.* **140**:1 (2000), 1–43. MR 2001g:11184 Zbl 1010.14004

[Darmon et al. 1997] H. Darmon, F. Diamond, and R. Taylor, "Fermat's last theorem", pp. 2–140 in *Elliptic curves, modular forms & Fermat's last theorem*, 2nd ed. (Hong Kong, 1993), edited by J. H. Coates and S.-T. Yau, International Press, Cambridge, MA, 1997. MR 99d:11067b Zbl 0877.11035

[Emerton 2006a] M. Emerton, "Jacquet modules of locally analytic representations of p-adic reductive groups, I: Construction and first properties", *Ann. Sci. École Norm. Sup.* (4) **39**:5 (2006), 775–839. MR 2008c:22013 Zbl 1117.22008

[Emerton 2006b] M. Emerton, "On the interpolation of systems of eigenvalues attached to automorphic Hecke eigenforms", *Invent. Math.* **164**:1 (2006), 1–84. MR 2007k:22018 Zbl 1090.22008

[Emerton 2007] M. Emerton, "Jacquet modules of locally analytic representations of p-adic reductive groups, II: The relation to parabolic induction", preprint, 2007, http://www.math.uchicago.edu/~emerton/pdffiles/jacquet-two.pdf. To appear in *J. Inst. Math. Jussieu.*

[Emerton 2011] M. Emerton, "Local–global compatibility in the p-adic Langlands programme for $\mathrm{GL}_{2/\mathbb{Q}}$", preprint, 2011, http://www.math.uchicago.edu/~emerton/pdffiles/lg.pdf.

[Emerton and Gee 2013] M. Emerton and T. Gee, "p-adic Hodge-theoretic properties of étale cohomology with mod p coefficients, and the cohomology of Shimura varieties", preprint, 2013. arXiv 1203.4963

[Fontaine and Mazur 1995] J.-M. Fontaine and B. Mazur, "Geometric Galois representations", pp. 41–78 in *Elliptic curves, modular forms, & Fermat's last theorem* (Hong Kong, 1993), edited by J. H. Coates and S.-T. Yau, Series in Number Theory **1**, International Press, Cambridge, MA, 1995. MR 96h:11049 Zbl 0839.14011

[Fontaine and Rapoport 2005] J.-M. Fontaine and M. Rapoport, "Existence de filtrations admissibles sur des isocristaux", *Bull. Soc. Math. France* **133**:1 (2005), 73–86. MR 2005m:14032 Zbl 1073.14025

[Harris and Taylor 2001] M. Harris and R. Taylor, *The geometry and cohomology of some simple Shimura varieties*, Annals of Mathematics Studies **151**, Princeton University Press, 2001. MR 2002m:11050 Zbl 1036.11027

[Hu 2009] Y. Hu, "Normes invariantes et existence de filtrations admissibles", *J. Reine Angew. Math.* **634** (2009), 107–141. MR 2011c:11177 Zbl 1214.11132

[Kaletha et al. 2014] T. Kaletha, A. Mínguez, S. W. Shin, and P.-J. White, “Endoscopic classification of representations: inner forms of unitary groups”, preprint, 2014. arXiv 1409.3731

[Katz 1978] N. M. Katz, “p-adic L-functions for CM fields”, *Invent. Math.* **49**:3 (1978), 199–297. MR 80h:10039 Zbl 0417.12003

[Lazarus 2000] X. Lazarus, *Module universel non ramifié pour un groupe réductif p-adique*, Ph.D. thesis, Université Paris-Sud, 2000.

[Loeffler 2011] D. Loeffler, “Overconvergent algebraic automorphic forms”, *Proc. Lond. Math. Soc.* (3) **102**:2 (2011), 193–228. MR 2012f:11098 Zbl 1232.11056

[Mahler 1958] K. Mahler, “An interpolation series for continuous functions of a p-adic variable”, *J. Reine Angew. Math.* **199** (1958), 23–34. MR 20 #2321 Zbl 0080.03504

[Mínguez 2011] A. Mínguez, “Unramified representations of unitary groups”, pp. 389–410 in *On the stabilization of the trace formula*, edited by L. Clozel et al., Stabilization of the Trace Formula, Shimura Varieties, and Arithmetic Applications **1**, International Press, Somerville, MA, 2011. MR 2856377

[Mok 2014] C. P. Mok, *Endoscopic classification of representations of quasi-split unitary groups*, Memoirs of the American Mathematical Society **235**:1108, American Mathematical Society, Providence, RI, online publication October 2014. In press.

[Schneider 2002] P. Schneider, *Nonarchimedean functional analysis*, Springer, Berlin, 2002. MR 2003a:46106 Zbl 0998.46044

[Schneider and Teitelbaum 2001] P. Schneider and J. Teitelbaum, “$U(\mathfrak{g})$-finite locally analytic representations”, *Represent. Theory* **5** (2001), 111–128. MR 2002e:22023 Zbl 1028.17007

[Schneider and Teitelbaum 2002] P. Schneider and J. Teitelbaum, “Banach space representations and Iwasawa theory”, *Israel J. Math.* **127** (2002), 359–380. MR 2003c:22026 Zbl 1006.46053

[Schneider and Teitelbaum 2006] P. Schneider and J. Teitelbaum, “Banach–Hecke algebras and p-adic Galois representations”, *Doc. Math.* Extra Volume (2006), 631–684. MR 2008b:11126 Zbl 1140.11026

[Sorensen 2013] C. M. Sorensen, “A proof of the Breuil–Schneider conjecture in the indecomposable case”, *Ann. Math.* (2) **177**:1 (2013), 367–382. MR 2999043 Zbl 06146422

[Taylor 1991] R. Taylor, “Galois representations associated to Siegel modular forms of low weight”, *Duke Math. J.* **63**:2 (1991), 281–332. MR 92j:11044 Zbl 0810.11033

[Vignéras 2004] M.-F. Vignéras, “On highest Whittaker models and integral structures”, pp. 773–801 in *Contributions to automorphic forms, geometry, and number theory*, edited by H. Hida et al., Johns Hopkins University Press, Baltimore, MD, 2004. MR 2006b:11048 Zbl 1084.11023

[Vignéras 2010] M.-F. Vignéras, “Banach l-adic representations of p-adic groups”, pp. 1–11 in *Représentations p-adiques de groupes p-adiques, II: Représentations de* $\mathrm{GL}_2(\mathbb{Q}_p)$ *et* (φ, Γ)*-modules*, edited by L. Berger et al., Astérisque **330**, Société Mathématique de France, Paris, 2010. MR 2011g:11102 Zbl 1232.11062

[White 2012] P.-J. White, “Tempered automorphic representations of the unitary group”, preprint, 2012. arXiv 1106.1127

Received May 7, 2014.

CLAUS M. SORENSEN
DEPARTMENT OF MATHEMATICS
UCSD
9500 GILMAN DR.
LA JOLLA, CA 92093
UNITED STATES
csorensen@ucsd.edu

PACIFIC JOURNAL OF MATHEMATICS
Vol. 275, No. 1, 2015
dx.doi.org/10.2140/pjm.2015.275.231

THE HEEGAARD DISTANCES COVER ALL NONNEGATIVE INTEGERS

RUIFENG QIU, YANQING ZOU AND QILONG GUO

We prove two main results: (1) For any integers $n \geq 1$ and $g \geq 2$, there is a closed 3-manifold M_g^n admitting a distance-n, genus-g Heegaard splitting, unless $(g, n) = (2, 1)$. Furthermore, M_g^n can be chosen to be hyperbolic unless $(g, n) = (3, 1)$. (2) For any integers $g \geq 2$ and $n \geq 4$, there are infinitely many nonhomeomorphic closed 3-manifolds admitting distance-n, genus-g Heegaard splittings.

1. Introduction

Let S be a compact surface with $\chi(S) \leq -2$ but not a 4-punctured sphere. Harvey [1981] defined the curve complex $\mathcal{C}(S)$ as follows: The vertices of $\mathcal{C}(S)$ are the isotopy classes of essential simple closed curves on S, and $k+1$ distinct vertices $x_0, x_1, \ldots, x_k$ determine a k-simplex of $\mathcal{C}(S)$ if and only if they are represented by pairwise disjoint simple closed curves. For two vertices x and y of $\mathcal{C}(S)$, the distance of x and y, denoted by $d_{\mathcal{C}(S)}(x, y)$, is defined to be the minimal number of 1-simplexes in a simplicial path joining x to y. In other words, $d_{\mathcal{C}(S)}(x, y)$ is the smallest integer $n \geq 0$ such that there is a sequence of vertices $x_0 = x, \ldots, x_n = y$, such that x_{i-1} and x_i are represented by two disjoint essential simple closed curves on S for each $1 \leq i \leq n$. For two sets of vertices in $\mathcal{C}(S)$, say X and Y, $d_{\mathcal{C}(S)}(X, Y)$ is defined to be $\min\{d_{\mathcal{C}(S)}(x, y) \mid x \in X, y \in Y\}$. Now let S be a torus or a once-punctured torus. In this case, the curve complex $\mathcal{C}(S)$ is defined as follows: The vertices of $\mathcal{C}(S)$ are the isotopy classes of essential simple closed curves on S, and $k+1$ distinct vertices $x_0, x_1, \ldots, x_k$ determine a k-simplex of $\mathcal{C}(S)$ if and only if x_i and x_j are represented by two simple closed curves c_i and c_j on S, such that c_i intersects c_j in just one point for each $0 \leq i \neq j \leq k$.

Let M be a compact orientable 3-manifold. If there is a closed surface S which cuts M into two compression bodies V and W such that $S = \partial_+ V = \partial_+ W$, then we say M has a Heegaard splitting, denoted by $M = V \cup_S W$, where $\partial_+ V$ (resp. $\partial_+ W$) is the positive boundary of V (resp. W). Let $\mathcal{D}(V)$ (resp. $\mathcal{D}(W)$) be the set

Qiu is supported by NSFC 11171108, Zou and Guo are supported by NSFC 11271058.
MSC2010: primary 57M27; secondary 57M50.
Keywords: attaching handlebody, Heegaard distance, subsurface projection.

of vertices in $\mathcal{C}(S)$ such that each element of $\mathcal{D}(V)$ (resp. $\mathcal{D}(W)$) represents the boundary of an essential disk in V (resp. W). Then the distance of the Heegaard splitting $V \cup_S W$, denoted by $d_{\mathcal{C}(S)}(V, W)$, is defined to be $d_{\mathcal{C}(S)}(\mathcal{D}(V), \mathcal{D}(W))$; see [Hempel 2001].

It is well known that a 3-manifold admitting a high distance Heegaard splitting has good topological and geometric properties. For example, Hartshorn [2002] and Scharlemann [2006] showed that a 3-manifold admitting a high distance Heegaard splitting contains no essential surface with small Euler characteristic number; Scharlemann and Tomova [2006] showed that a high distance Heegaard splitting is the unique minimal Heegaard splitting up to isotopy. By Geometrization theorem and Hempel's work [2001] in Heegaard splittings of Seifert manifolds, a 3-manifold M admitting a distance at least three Heegaard splitting is hyperbolic. From this point of view, Heegaard distance is an active topic in Heegaard splitting. Here we give a brief survey on the existences of high distance Heegaard splittings. Hempel [ibid.] showed that for any integers $g \geq 2$, and $n \geq 2$, there is a 3-manifold that admits a distance at least n Heegaard splitting of genus g. Similar results were obtained using different methods in [Evans 2006; Campisi and Rathbun 2012]. Minsky, Moriah and Schleimer [Minsky et al. 2007] proved the same result for knot complements, and Li [2013] constructed the non-Haken manifolds admitting high distance Heegaard splittings. In general, generic Heegaard splittings have Heegaard distances at least n for any $n \geq 2$; see [Lustig and Moriah 2009; 2010; 2012]. By studying Dehn filling, Ma, Qiu and Zou announced that they had proved that distances of genus-two Heegaard splittings cover all nonnegative integers except one. Recently, Ido, Jang and Kobayashi [Ido et al. 2014] proved that, for any $n > 1$ and $g > 1$, there is a compact 3-manifold with two boundary components which admits a distance-n Heegaard splitting of genus g; Johnson informed us that he had proved that there is always a closed 3-manifold admitting a distance-n (≥ 5), genus-g Heegaard splitting and a genus larger strongly irreducible Heegaard splitting.

The main result of this paper is the following:

Theorem 1.1. *For any integers $n \geq 1$ and $g \geq 2$, there is a closed 3-manifold M_g^n which admits a distance-n Heegaard splitting of genus g unless $(g, n) = (2, 1)$. Furthermore, M_g^n can be chosen to be hyperbolic unless $(g, n) = (3, 1)$.*

Remark 1.2. (1) It is well known that there is no distance-one Heegaard splitting of genus two.

(2) Hempel [2001] showed that any Heegaard splitting of a Seifert 3-manifold has distance at most two. Now a natural question is: For any integer $g \geq 2$, is there a closed hyperbolic 3-manifold admitting a distance-2 Heegaard splitting of genus g?

When $g = 2$, Eudave-Muñoz [1999] proved that there is a hyperbolic (1, 1)-knot in 3-sphere, say K. In this case, the complement of K, say M_K, admits a distance-2 Heegaard splitting of genus two. By the main results in [Scharlemann 2006; Kobayashi and Qiu 2008; Agol 2010], there is an essential simple closed curve r on ∂M_K such that the manifold obtained by doing a Dehn filling on M_K along r, say M_K^r, is still hyperbolic. Hence M_K^r admits a distance-2 Heegaard splitting of genus two. Maybe the answer to this question has been well known for $g \geq 3$, but we find no published paper or book related to it.

(3) If M admits a distance-1 Heegaard splitting of genus three, then M contains an essential torus. Hence M is not hyperbolic.

(4) The proof of Theorem 1.1 implies the following fact: Let n be a positive integer, let $\{F_1, \dots, F_n\}$ be a collection of closed orientable surfaces, and let I and $J = \{1, \dots, n\} \backslash I$ be two subsets of $\{1, \dots, n\}$. Then, for any integers

$$g \geq \max\left\{\sum_{i \in I} g(F_i),\ \sum_{j \in J} g(F_j)\right\}$$

and $m \geq 2$, there is a compact 3-manifold M admitting a distance-m Heegaard splitting of genus g, denoted by $M = V \cup_S W$, such that $F_i \subset \partial_- V$ for $i \in I$, $F_j \subset \partial_- W$ for $j \in J$. We omit the proof.

By the arguments in Theorem 1.1, we have:

Theorem 1.3. *For any integers $g \geq 2$ and $n \geq 4$, there are infinitely many nonhomeomorphic closed 3-manifolds admitting distance-n Heegaard splittings of genus g.*

We organize this paper as follows. In Section 2, we introduce some results on curve complex. Then we will prove Theorem 1.1 for $n \neq 2$ in Section 3, for $n = 2$ in Section 5 and Theorem 1.3 in Section 4.

2. Preliminaries of curve complex

Let S be a compact surface of genus at least one and $\mathcal{C}(S)$ the curve complex of S. We say that a simple closed curve c in S is essential if c bounds no disk in S and is not parallel to ∂S. Hence each vertex of $\mathcal{C}(S)$ is represented by the isotopy class of an essential simple closed curve in S. For simplicity, we do not distinguish the essential simple closed curve c and its isotopy class c.

Lemma 2.1 [Minsky 1996; Masur and Minsky 1999; 2000]. *$\mathcal{C}(S)$ is connected, and the diameter of $\mathcal{C}(S)$ is infinite.*

We say that a collection $\mathcal{G} = \{a_0, a_1, \dots, a_n\}$ is a geodesic in $\mathcal{C}(S)$ if $a_i \subset \mathcal{C}^0(S)$ and $d_{\mathcal{C}(S)}(a_i, a_j) = |i - j|$, for any $0 \leq i, j \leq n$. And the length of $\mathcal{G}$, denoted by $\mathcal{L}(\mathcal{G})$, is defined to be n. By the connectedness of $\mathcal{C}^1(S)$, there is always a shortest path in $\mathcal{C}^1(S)$ connecting any two vertices of $\mathcal{C}(S)$. For any two vertices α, β with $d_S(\alpha,\beta)=n$, we say that a geodesic $\mathcal{G}$ connects α, β if $\mathcal{G}=\{a_0=\alpha,\dots,a_n=\beta\}$. Now for any two subsimplicial complex $X, Y\subset\mathcal{C}(S)$, we say that a geodesic $\mathcal{G}$ realizes the distance between X and Y if $\mathcal{G}$ connects a vertex $\alpha\in X$ and a vertex $\beta\in Y$ such that $\mathcal{L}(\mathcal{G})=d_{\mathcal{C}(S)}(X,Y)$.

Let F be a compact surface of genus at least one with nonempty boundary. Similar to the definition of the curve complex $\mathcal{C}(F)$, we define the arc and curve complex $\mathcal{AC}(F)$ as follows. Each vertex of $\mathcal{AC}(F)$ is the isotopy class of an essential simple closed curve or an essential properly embedded arc in F, and a set of vertices forms a simplex of $\mathcal{AC}(F)$ if these vertices are represented by pairwise disjoint arcs or curves in F. For any two vertices which are realized by disjoint curves or arcs, we place an edge between them. All the vertices and edges form the 1-skeleton of $\mathcal{AC}(F)$, denoted by $\mathcal{AC}^1(F)$. For each edge, we assign it length one. Thus for any two vertices α and β in $\mathcal{AC}^1(F)$, the distance $d_{\mathcal{AC}(F)}(\alpha,\beta)$ is defined to be the minimal length of paths in $\mathcal{AC}^1(F)$ connecting α and β. Similarly, we can define the geodesic in $\mathcal{AC}(F)$.

When F is a subsurface of S, we say that F is essential in S if the induced map of the inclusion from $\pi_1(F)$ to $\pi_1(S)$ is injective. Furthermore, we say that F is a proper essential subsurface of S if F is essential in S and at least one boundary component of F is essential in S. For more details, see [Masur and Minsky 2000].

If F is an essential subsurface of S, there is some connection between $\mathcal{AC}(F)$ and $\mathcal{C}(S)$. For any $\alpha\in\mathcal{C}^0(S)$, there is an essential simple closed curve α_{geo} representing α such that the geometric intersection number $i(\alpha_{\text{geo}},\partial F)$ is minimal. Hence each component of $\alpha_{\text{geo}}\cap F$ is essential in F. Now for $\alpha\in\mathcal{C}(S)$, let $\kappa_F(\alpha)$ be the collection of isotopy classes of the essential components of $\alpha_{\text{geo}}\cap F$.

For any $\gamma\in\mathcal{C}(F)$, we define the set $\sigma_F(\gamma)$ as follows: $\gamma'\in\sigma_F(\gamma)$ if and only if γ' is the essential boundary component of a closed regular neighborhood of $\gamma\cup\partial F$. Set $\sigma_F(\varnothing)=\varnothing$. Now let $\pi_F=\sigma_F\circ\kappa_F$. Then the map π_F links $\mathcal{C}(F)$ and $\mathcal{C}(S)$, which is the subsurface projection map in [ibid.].

We say $\alpha\in\mathcal{C}^0(S)$ cuts F if $\pi_F(\alpha)\neq\varnothing$. If $\alpha,\beta\in\mathcal{C}^0(S)$ both cut F, we denote $d_{\mathcal{C}(F)}(\alpha,\beta)=\operatorname{diam}_{\mathcal{C}(F)}(\pi_F(\alpha),\pi_F(\beta))$. And if $d_{\mathcal{C}(S)}(\alpha,\beta)=1$, then

$$d_{\mathcal{AC}(F)}(\alpha,\beta)\le 1,$$
$$d_{\mathcal{C}(F)}(\alpha,\beta)\le 2,$$

observed by H. Masur and Y. N. Minsky. When the two vertices α and β have distance k in $\mathcal{C}(S)$, we have a direct consequence of the above observation:

Lemma 2.2. *Let F and S be as above, $\mathcal{G}=\{\alpha_0,\dots,\alpha_k\}$ be a geodesic in $\mathcal{C}(S)$ such that α_i cuts F for each $0\le i\le k$. Then $d_{\mathcal{C}(F)}(\alpha_0,\alpha_k)\le 2k$.*

Moreover, Masur and Minsky [ibid.] proved:

Lemma 2.3 (bounded geodesic image theorem). *Let F be an essential proper subsurface of S, and let γ be a geodesic segment in $\mathcal{C}(S)$, so that $\pi_F(v) \neq \varnothing$ for every vertex v of γ. Then there is a constant $\mathcal{M}$ depending only on S so that* $\operatorname{diam}_{\mathcal{C}(F)}(\pi_F(\gamma)) \leq \mathcal{M}$.

When S is closed with $g(S) \geq 2$, there is always a compact 3-manifold M with S as its compressible boundary. Let $\mathcal{D}(M, S)$, called the disk complex for S, be the subset of vertices of $\mathcal{C}(S)$, where each element bounds a disk in M. For an essential simple closed curve on S, say c, we say that it is disk-busting if $S-c$ is incompressible in M.

Now let's consider the subsurface projection of disk complex. The following disk image theorem is proved by Li [2012], Masur and Schleimer [2013] independently.

For any I-bundle J over a bounded compact surface P, $\partial J = \partial_v J \cup \partial_h J$, where the vertical boundary $\partial_v J$ is the I-bundle related to ∂P, and the horizontal boundary $\partial_h J$ is the portion of ∂J transverse to the I-fibers.

Lemma 2.4. *Let M be a compact orientable and irreducible* 3*-manifold. S is a boundary component of M. Suppose $\partial M - S$ is incompressible. Let $\mathcal{D}$ be the disk complex of S, and let $F \subset S$ be an essential subsurface. Assume each component of ∂F is disk-busting. Then either*

(1) *M is an I-bundle over some compact surface, F is a horizontal boundary of the I-bundle and the vertical boundary of this I-bundle is a single annulus. Or,*

(2) *The image of this complex, $\kappa_F(\mathcal{D})$, lies in a ball of radius three in $\mathcal{AC}(F)$. In particular, $\kappa_F(\mathcal{D})$ has diameter six in $\mathcal{AC}(F)$. Moreover, $\pi_F(\mathcal{D})$ has diameter at most twelve in $\mathcal{C}(F)$.*

Hempel introduced a full simplex X on S which is a dimension $3g(S)-4$ simplex in $\mathcal{C}(S)$. Then after attaching 2-handles and 3-handles along the vertices of X on the same side of S, there is a handlebody H_X with $\partial H_X = S$.

Lemma 2.5 [Hempel 2001]. *Let S be a closed, orientable surface of genus at least two. For any positive number d and any full simplex X of $\mathcal{C}(S)$, there is another full simplex Y of $\mathcal{C}(S)$ such that $d_{\mathcal{C}(S)}(\mathcal{D}(H_X), \mathcal{D}(H_Y)) \geq d$.*

Through subsurface projection, the bounded geodesic image theorem links the geodesic in the curve complex of the entire surface to the curve complex of a proper subsurface. Since the diameter of the curve complex is infinite, we can construct a geodesic of any given length in the curve complex. Furthermore, we require that the constructed geodesic satisfies that both the first and last vertices are represented by separating essential simple closed curves.

We organize our results:

Lemma 2.6. *Let g, n, m, s, t be integers such that $g, m, n \geq 2$, $1 \leq t, s \leq g-1$. Let S_g be a closed surface of genus g. Then there are two essential separating curves α*

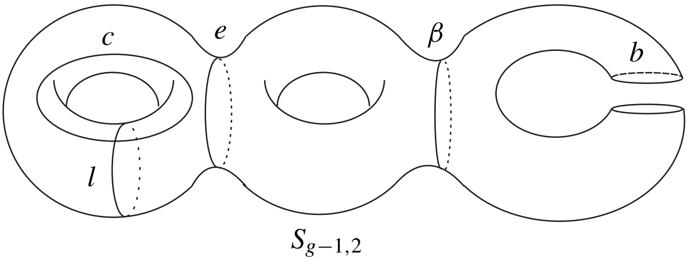


Figure 1. Self-banding.

and β in S_g such that $d_{\mathcal{C}(S_g)}(\alpha, \beta) = n$; one component of $S_g - \alpha$ has genus t; one component of $S_g - \beta$ has genus s. Furthermore, there is a geodesic

$$\mathcal{G} = \{a_0 = \alpha, a_1, \dots, a_{n-1}, a_n = \beta\}$$

in $\mathcal{C}(S_g)$ such that

(1) *a_i is nonseparating in S_g for $1 \le i \le n-1$, and*

(2) *$m\mathcal{M} + 2 \le d_{\mathcal{C}(S^{a_i})}(a_{i-1}, a_{i+1}) \le m\mathcal{M} + 6$, where S^{a_i} is the surface $S - N(a_i)$ for $1 \le i \le n-1$ and $\mathcal{M}$ is the constant in Lemma 2.3.*

Proof. Let α be an essential separating curve in S such that one component of $S_g - \alpha$, say S_1, has genus t.

Suppose first that $n = 2$. Let b be a nonseparating curve in S_g which is disjoint from α. Let S^b be the surface $S_g - N(b)$, where $N(b)$ is an open regular neighborhood of b in S_g. Then S^b is a genus-$(g-1)$ surface with two boundary components. Furthermore, α is an essential separating simple closed curve in S^b.

By Lemma 2.1, $\mathcal{C}^1(S^b)$ is connected and its diameter is infinite. Hence there is an essential simple closed curve c in S^b with $d_{\mathcal{C}(S^b)}(\alpha, c) = m\mathcal{M} + 4$. Note that $g - 1 \ge 1$. If c is separating in S^b, then there is a nonseparating essential simple closed curve c^* in S^b such that $c \cap c^* = \varnothing$. Hence $d_{\mathcal{C}(S^b)}(c, c^*) = 1$, and

$$m\mathcal{M} + 3 \le d_{\mathcal{C}(S^b)}(\alpha, c^*) \le m\mathcal{M} + 5.$$

So there is a nonseparating essential simple closed curve c in S^b such that

$$m\mathcal{M} + 3 \le d_{\mathcal{C}(S^b)}(\alpha, c) \le m\mathcal{M} + 5.$$

Let l be a nonseparating simple closed curve in S^b such that l intersects c in one point, and let e be the boundary of the closed regular neighborhood of $c \cup l$ in S^b. Then e bounds a once-punctured torus T containing l and c. Since $s \le g - 1$, there is an essential separating simple closed curve β in S^b such that β bounds a once-punctured surface of genus s containing T as a subsurface, see Figure 1.

So β is also separating in S_g. Now we prove that

$$d_{\mathcal{C}(S_g)}(\alpha, \beta) = 2 \quad \text{and} \quad d_{\mathcal{C}(S_g)}(\alpha, c) = 2.$$

Since $\alpha \cap b = \varnothing$, $\beta \cap b = \varnothing$ and $c \cap b = \varnothing$, $d_{\mathcal{C}(S_g)}(\alpha, \beta) \leq 2$ and $d_{\mathcal{C}(S_g)}(\alpha, c) \leq 2$. Since $c \cap \beta = \varnothing$, by the assumption on $d_{\mathcal{C}(S^b)}(\alpha, c)$,

$$m\mathcal{M} + 2 \leq d_{\mathcal{C}(S^b)}(\alpha, \beta) \leq m\mathcal{M} + 6.$$

So $d_{\mathcal{C}(S_g)}(\beta, \alpha) = 2$. For if $d_{\mathcal{C}(S_g)}(\alpha, \beta) \leq 1$, then, by Lemma 2.3, $d_{\mathcal{C}(S^b)}(\alpha, \beta) \leq \mathcal{M}$, a contradiction. Similarly, $d_{\mathcal{C}(S_g)}(\alpha, c) = 2$. And

$$\mathcal{G} = \{a_0 = \alpha, a_1 = b, a_2 = \beta\} \quad \text{and} \quad \mathcal{G}^* = \{a_0 = \alpha, a_1 = b, a_2 = c\}$$

are two geodesics of $\mathcal{C}(S_g)$. Furthermore, $\mathcal{G}$ satisfies the conclusion of Lemma 2.6.

Now we prove this lemma by induction on n.

Assumption. Let $k \geq 2$. Suppose that there are two essential separating simple closed curves α and β, and a nonseparating simple closed curve c in S_g such that

$$d_{\mathcal{C}(S_g)}(\alpha, \beta) = k,$$
$$d_{\mathcal{C}(S_g)}(\alpha, c) = k,$$

and one component of $S_g - \alpha$ has genus t while one component of $S_g - \beta$ has genus s. Furthermore, there is a geodesic $\mathcal{G}^* = \{\alpha, a_1, \ldots, a_{k-1}, a_k = c\}$ where a_i is nonseparating in S_g for each $1 \leq i \leq k$, satisfying

$$m\mathcal{M} + 3 \leq d_{\mathcal{C}(S^{a_i})}(a_{i-1}, a_{i+1}) \leq m\mathcal{M} + 5 \quad \text{for any } 1 \leq i \leq k-2,$$
$$m\mathcal{M} + 3 \leq d_{\mathcal{C}(S^{a_{k-1}})}(a_{k-2}, c) \leq m\mathcal{M} + 5,$$

and a geodesic $\mathcal{G} = \{\alpha = a_0, a_1, \ldots, a_{k-1}, \beta\}$ satisfying the conclusions (1) and (2) of Lemma 2.6.

Let S^c be the surface $S_g - N(c)$, where $N(c)$ is an open regular neighborhood of c in S_g. Since c is nonseparating in S_g, S^c is a genus-$(g-1)$ surface with two boundary components. Since $\mathcal{G}^* = \{\alpha, a_1, \ldots, a_{k-1}, c\}$ is also a geodesic connecting α to c, a_{k-1} is an essential nonseparating simple closed curve in S^c. By the above argument, there is an essential nonseparating curve h and an essential separating curve e in S^c such that

(1) e bounds an once-punctured torus T^* containing h;

(2) $m\mathcal{M} + 3 \leq d_{\mathcal{C}(S^c)}(h, a_{k-1}) \leq m\mathcal{M} + 5$;

(3) $m\mathcal{M} + 2 \leq d_{\mathcal{C}(S^c)}(e, a_{k-1}) \leq m\mathcal{M} + 6$.

And there is also an essential separating simple closed curve γ which bounds a genus-s subsurface of S^c containing T^* as a subsurface, while γ is also separating

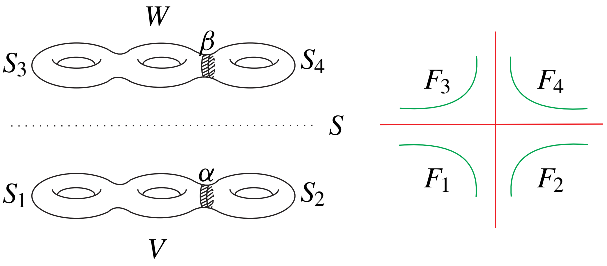


Figure 2. Heegaard splitting I.

in S_g. Since h is disjoint from γ,

$$m\mathcal{M}+2 \leq d_{\mathcal{C}(S^c)}(\gamma, a_{k-1}) \leq m\mathcal{M}+6.$$

Now we prove that $d_{\mathcal{C}(S_g)}(\alpha, h) = k+1$, $d_{\mathcal{C}(S_g)}(\alpha, \gamma) = k+1$.

Suppose, on the contrary, that $d_{\mathcal{C}(S_g)}(\alpha, h) = x \leq k$. Then there exists a geodesic $\mathcal{G}_1 = \{\alpha = b_0, \dots, b_x = h\}$. Note that each of α and h is not isotopic to c and the length is less than or equal to k. Since $d_{\mathcal{C}(S_g)}(\alpha, c) = k$, b_j is not isotopic to c for $1 \leq j \leq x-1$. This means b_j cuts S^c for each $0 \leq j \leq x$. By Lemma 2.3, $d_{\mathcal{C}(S^c)}(\alpha, h) \leq \mathcal{M}$. Since $d_{\mathcal{C}(S_g)}(\alpha, c) = k$, a_j is not isotopic to c for $0 \leq j \leq k-1$. By using Lemma 2.3 again, $d_{S^c}(\alpha, a_{k-1}) \leq \mathcal{M}$. Then $d_{\mathcal{C}(S^c)}(a_{k-1}, h) \leq 2\mathcal{M}$. It contradicts the choice of h.

Now $\mathcal{G}' = \{a_0 = \alpha, a_1, \dots, a_{k-1}, c, \gamma\}$ and $\mathcal{G}'' = \{a_0 = \alpha, a_1, \dots, a_{k-1}, c, h\}$ are two geodesics satisfying the conclusion. □

3. Proof of Theorem 1.1 ($n \neq 2$)

In this section, we will prove:

Proposition 3.1. *For any positive integers $n \neq 2$ and $g \geq 2$, there is a closed 3-manifold which admits a distance-n Heegaard splitting of genus g unless $(g, n) = (2, 1)$. Furthermore, M_g^n can be chosen to be hyperbolic unless $(g, n) = (3, 1)$.*

Proof. We first suppose that $n \geq 3$.

Let S be a closed surface of genus g. By Lemma 2.6, there are two separating essential simple closed curves α and β such that $d_{\mathcal{C}(S)}(\alpha, \beta) = n$ for $n \geq 3$. Let V be the compression body obtained by attaching a 2-handle to $S \times [0, 1]$ along $\alpha \times \{1\}$, and let W be the compression body obtained by attaching a 2-handle to $S \times [-1, 0]$ along $\beta \times \{-1\}$. Then $V \cup_S W$ is a Heegaard splitting where S is the surface $S \times \{0\}$; see Figure 2.

Since V contains only one essential disk B with $\partial B = \alpha$ up to isotopy and W contains only one essential disk D with $\partial D = \beta$ up to isotopy, $d_{\mathcal{C}(S)}(V, W) = n$.

Let F_1 and F_2 be the components of $\partial_- V$, and S_1 and S_2 the two components of $S - \alpha$. Similarly, let F_3 and F_4 be the components of $\partial_- W$, and S_3 and S_4 the

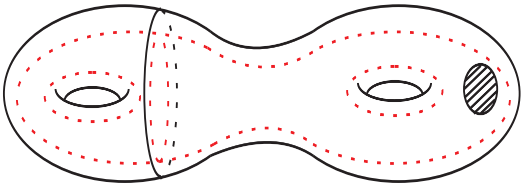


Figure 3. A spanning annulus.

two components of $S-\beta$. Now B cuts V into two manifolds $F_1 \times I$ and $F_2 \times I$, and D cuts W into two manifolds $F_3 \times I$ and $F_4 \times I$; see Figure 2. By Lemma 2.6, we assume that S_3 is a once-punctured torus.

We first consider the compression body V. We assume that $F_i = F_i \times \{0\}$, $S_i \cup B = F_i \times \{1\}$ for $1 \leq i \leq 2$. Let $f_{F_i} : S_i \cup B \to F_i$ be the natural homeomorphism such that $f_{F_i}(x \times \{1\}) = x \times \{0\}$ for $i = 1, 2$. And f_{F_i} is well defined. Then, for any two essential simple closed curves $\zeta, \theta \subset S_i \cup B$,

$$d_{\mathcal{C}(F_i)}(f_{F_i}(\zeta), f(\theta)) = d_{\mathcal{C}(S_i \cup B)}(\zeta, \theta) \quad \text{for } i = 1, 2;$$

see Figure 3. Hence f_{F_i} induces an isomorphism from $\mathcal{C}(S_i \cup B)$ to $\mathcal{C}(F_i)$, for any $i = 1, 2$. Denote the isomorphism by f_{F_i} too. Note that the shaded disk in Figure 3 is B.

Let $\iota : S_i \to S_i \cup B$ be the inclusion map for $i = 1, 2$. Note that ∂S_i contains only one component. If c is an essential simple closed curve in S_i, $\iota(c)$ is also essential in $S_i \cup B$. So, for any two essential simple closed curves $\zeta, \theta \subset S_i$,

$$d_{\mathcal{C}(S_i \cup B)}(\iota(\zeta), \iota(\theta)) \leq d_{S_i}(\zeta, \theta) \quad \text{for } i = 1, 2.$$

Hence ι induces a distance nonincreasing map from $\mathcal{C}(S_i)$ to $\mathcal{C}(S_i \cup B)$, for any $i = 1, 2$. Denote the inclusion map by ι too. Then we define

$$\psi_{F_i} = f_{F_i} \circ \iota \circ \pi_{S_i}.$$

Since $d_{\mathcal{C}(S)}(\alpha, \beta) = n \geq 2$, $\alpha \cap \beta \neq \varnothing$. By the argument in Section 2,

$$\operatorname{diam}_{\mathcal{C}(S_i)}(\pi_{S_i}(\beta)) \leq 2.$$

Hence,

$$\operatorname{diam}_{\mathcal{C}(F_i)}(\psi_{F_i}(\beta)) \leq 2.$$

We start to attach a handlebody to V along F_1. Then we have two cases:

(a) F_1 is a torus. By Lemma 2.1, there is an essential simple closed curve r in F_1 such that

$$d_{\mathcal{C}(F_1)}(\psi_{F_1}(\beta), r) \geq \mathcal{M} + 1. \tag{1}$$

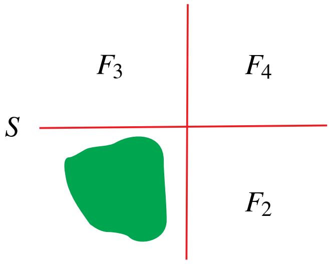


Figure 4. Heegaard splitting II.

Let J_r be a solid torus such that $\partial J_r = F_1$, and r bounds an essential disk in J_r. In this case, J_r contains only one essential disk up to isotopy. Let V_{F_1} be the manifold $V \cup J_r$.

(b) $g(F_1) \geq 2$. By Lemma 2.5, there is a full simplex X of $\mathcal{C}(F_1)$ such that

$$d_{\mathcal{C}(F_1)}(\mathcal{D}(H_X), \psi_{F_1}(\beta)) \geq \mathcal{M} + 1,$$

where H_X is the handlebody obtained by attaching 2-handles to F_1 along X then 3-handles to cap off the possible 2-spheres. In this case, we denote the manifold $V \cup H_X$ by V_{F_1}.

In a word, V_{F_1} is a compression body with only one negative boundary component F_2, where $\partial_+ V_{F_1} = \partial_+ W$; see Figure 4. Hence $V_{F_1} \cup_S W$ is a Heegaard splitting.

Claim 3.2. *The Heegaard distance $d_{\mathcal{C}(S)}(V_{F_1}, W)$ is n.*

Proof. Suppose, otherwise, that $d_{\mathcal{C}(S)}(V_{F_1}, W) = k < n$. Since W contains only one essential disk D up to isotopy where $\partial D = \beta$, there is an essential disk B_1 in V_{F_1} such that $d_{\mathcal{C}(S)}(\partial B_1, \beta) = k \leq n-1$, i.e, there is a geodesic $\mathcal{G} = \{a_0 = \beta, \ldots, a_k = \partial B_1\}$, where $k \leq n-1$. □

Claim 3.3. *$a_j \cap S_1 \neq \varnothing$, for any $0 \leq j \leq k$.*

Proof. Suppose that $a_j \cap S_1 = \varnothing$ for some $0 \leq j \leq k$. If $a_k \cap S_1 = \varnothing$, then $B_1 \subset F_2 \times I$ and B_1 is inessential in V_{F_1}. So $j \neq k$. Since $a_0 = \beta$, $j \neq 0$. Hence there is a geodesic $\mathcal{G}^* = \{\beta = a_0, \ldots, a_j, \alpha\}$. It means that $d_{\mathcal{C}(S)}(\alpha, \beta) \leq k < n$, a contradiction. □

By Lemma 2.3, $d_{\mathcal{C}(S_1 \cup B)}(\partial B_1, \beta) \leq \mathcal{M}$ and $d_{\mathcal{C}(F_1)}(\psi_{F_1}(\partial B_1), \psi_{F_1}(\beta)) \leq \mathcal{M}$. Depending on the intersection between B_1 and B, there are two cases:

(a) $B_1 \cap B = \varnothing$. Since B_1 is not isotopic to B, $\psi_{F_1}(\partial B_1)$ bounds an essential disk in H_X or J_r depending on $g(F_1)$, where H_X and J_r are constructed as above. Then

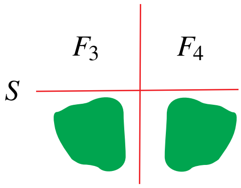


Figure 5. Heegaard splitting III.

by Lemma 2.3,

$$\begin{aligned} d_{\mathcal{C}(F_1)}(\psi_{F_1}(\partial B_1), \psi_{F_1}(\beta)) &\le \mathcal{M}, \\ d_{\mathcal{C}(F_1)}(r, \psi_{F_1}(\beta)) &\le \mathcal{M} \quad \text{if } g(F_1) = 1, \\ d_{\mathcal{C}(F_1)}(\mathcal{D}(H_X), \psi_{F_1}(\beta)) &\le \mathcal{M} \quad \text{if } g(F_1) \ge 2. \end{aligned}$$

It contradicts the choice of X or r.

(b) $B_1 \cap B \neq \varnothing$. Let a be an outermost arc of $B_1 \cap B$ on B_1. It means that a, together with a subarc $\gamma \subset \partial B_1$, bounds a disk B_γ such that $B_\gamma \cap B = a$. Since B cuts V_{F_1} into a handlebody H which contains F_1 and an I-bundle $F_2 \times I$, $B_\gamma \subset H$. Hence a curve in $\psi_{F_1}(\partial B_1)$ bounds an essential disk in H_X or J_r. By the argument in (a), it is impossible.

Now V_{F_1} is a compression body which has only one minus boundary component F_2. Since $d_{\mathcal{C}(S)}(\alpha, \beta) = n \ge 3$, $\beta \cap S_2 \neq \varnothing$. By Lemmas 2.1 and 2.5, there is always a simplex Y on F_2 such that $d_{\mathcal{C}(F_2)}(\mathcal{D}(H_Y), \psi_{F_2}(\beta)) \ge \mathcal{M} + 1$, where H_Y is the handlebody or the solid torus obtained by attaching 2-handles to F_2 along Y and 3-handles to cap off the possible 2-spheres. Let V_{F_1,F_2} be the manifold obtained by attaching H_Y to V_{F_1} along F_2; see Figure 5. Then V_{F_1,F_2} is a handlebody where $\partial_+ V_{F_1,F_2} = \partial_+ W$. Hence $V_{F_1,F_2} \cup_S W$ is also a Heegaard splitting.

Claim 3.4. *The Heegaard distance $d_{\mathcal{C}(S)}(V_{F_1,F_2}, W)$ is n.*

Proof. Suppose, on the contrary, that $d_{\mathcal{C}(S)}(V_{F_1,F_2}, W) = k < n$. Since W contains only one essential disk D up to isotopy such that $\partial D = \beta$, there is an essential disk B_2 in V_{F_1,F_2} such that $d_{\mathcal{C}(S)}(\partial B_2, \beta) = k$, i.e., there is a geodesic $\mathcal{G} = \{a_0 = \beta, \ldots, a_k = \partial B_2\}$, where $k \le n - 1$. By the definition of Heegaard distance, $a_j \cap \partial S_2 \neq \varnothing$ for $0 \le j \le k - 1$ when $k \ge 1$.

Note that $\partial B = \alpha$. Depending on the way of intersection between B_2 and B, there are two cases:

(a) $B_2 \cap B = \varnothing$. Since $d_{\mathcal{C}(S)}(\alpha, \beta) = n > k$, B_2 is not isotopic to B. By the proof of Claim 3.2, ∂B_2 does not lie in S_1. Hence $\partial B_2 \subset S_2$. It implies that $\psi_{F_2}(\partial B_2)$ bounds an essential disk in H_Y. By Lemma 2.3, $d_{\mathcal{C}(S_2)}(\partial B_2, \beta) \le \mathcal{M}$. Hence

$$d_{\mathcal{C}(F_2)}(\psi_{F_2}(\partial B_2), \psi_{F_2}(\beta)) \le \mathcal{M}, \quad d_{\mathcal{C}(F_2)}(\mathcal{D}(H_Y), \psi_{F_2}(\beta)) \le \mathcal{M}.$$

It contradicts the choice of Y.

(b) $B_2 \cap B \neq \varnothing$. Let a^* be an outermost arc of $B_2 \cap B$ on B_2. This means that a^*, together with a subarc $\gamma^* \subset \partial B_2$, bounds a disk B_{γ^*} such that $B_{\gamma^*} \cap B = a^*$. By the proof of Claim 3.2, $\gamma^* \subset S_2$. Thus $\psi_{F_2}(\partial B_2)$ bounds an essential disk in H_Y. By the same argument in Claim 3.2 again, it is impossible. □

Until now, we get a distance-n genus-g Heegaard splitting $V_{F_1,F_2} \cup_S W$. In this case, V_{F_1,F_2} is a handlebody, and W contains only one essential disk D such that $\partial D = \beta$. Furthermore, we cut S along β into two components S_3 and S_4, and cut W along D into two manifolds $F_3 \times I$ and $F_4 \times I$ such that $F_i = F_i \times \{0\}$, and $S_i \cup D = F_i \times \{1\}$ for $i = 3, 4$. Now the shaded disk in Figure 3 is D. Let $f_{F_i} : S_i \cup D \to F_i$ be the natural homeomorphism such that $f_{F_i}(x \times \{1\}) = x \times \{0\}$ for $i = 3, 4$. Then, for any two essential simple closed curves $\zeta, \theta \subset S_i \cup D$,

$$d_{\mathcal{C}(F_i)}(f_{F_i}(\zeta), f_{F_i}(\theta)) = d_{\mathcal{C}(S_i \cup D)}(\zeta, \theta) \quad \text{for } i = 3, 4;$$

see Figure 3. Hence f_{F_i} induces an isomorphism from $\mathcal{C}(S_i \cup D)$ to $\mathcal{C}(F_i)$, for any $i = 3, 4$. Denote the isomorphism by f_{F_i} too.

Let $\iota : S_i \to S_i \cup D$ be the inclusion map for $i = 3, 4$. Note that ∂S_i contains only one component. If c is an essential simple closed curve in S_i, $\iota(c)$ is also essential in $S_i \cup D$. Now, for any two essential simple closed curves $\zeta, \theta \subset S_i$,

$$d_{\mathcal{C}(S_i \cup D)}(\iota(\zeta), \iota(\theta)) \leq d_{S_i}(\zeta, \theta) \quad \text{for } i = 3, 4.$$

Hence ι induces a distance nonincreasing map from $\mathcal{C}(S_i)$ to $\mathcal{C}(S_i \cup D)$, for any $i = 3, 4$. Denote the inclusion map by ι too. Then we define

$$\psi_{F_i} = f_{F_i} \circ \iota \circ \pi_{S_i}.$$

Since $V_{F_1,F_2} \cup_S W$ is a distance-n (≥ 3) Heegaard splitting of genus g, and W contains only one essential disk D up to isotopy, S_3 and S_4 are incompressible in V_{F_1,F_2}. Hence $\beta = \partial S_3 = \partial S_4$ is disk-busting in V_{F_1,F_2}. Since the Heegaard distance $n \geq 3$ and $g(S_3) = 1$, V_{F_1,F_2} is not an I-bundle over some compact surface with S_i a horizontal boundary of the I-bundle while the vertical boundary of this I-bundle a single annulus for $i = 3, 4$. By Lemma 2.4, $\operatorname{diam}_{S_i}(\mathcal{D}(V_{F_1,F_2})) \leq 12$ for $i = 3, 4$. Hence $\operatorname{diam}_{F_i}(\psi_{F_i}(\mathcal{D}(V_{F_1,F_2}))) \leq 12$.

Since F_3 is a torus and $\operatorname{diam}_{F_3}(\psi_{F_3}(\mathcal{D}(V_{F_1,F_2}))) \leq 12$, by Lemma 2.1, there is an essential simple closed curve δ in F_3 such that $d_{\mathcal{C}(F_3)}(\psi_{F_3}(\mathcal{D}(V_{F_1,F_2})), \delta) \geq \mathcal{M} + 1$. Let W_{F_3} be the manifold obtained attaching a solid J_δ to W along F_3 so that δ bounds a disk in J_δ. Then W_{F_3} is a compression body.

Since $\operatorname{diam}_{F_4}(\psi_{F_4}(\mathcal{D}(V_{F_1,F_2}))) \leq 12$, by Lemmas 2.1 and 2.5, there is a simplex Z of $\mathcal{C}(F_4)$ such that

$$d_{\mathcal{C}(F_4)}(\mathcal{D}(H_Z), \psi_{F_4}(\mathcal{D}(V_{F_1,F_2}))) \geq \mathcal{M} + 1,$$

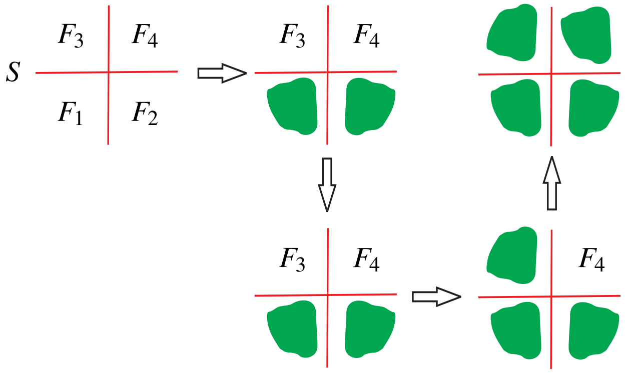


Figure 6. Heegaard splitting IV.

where H_Z is the handlebody or the solid torus obtained by attaching 2-handles to F_4 along Z then 3-handles to cap off the possible 2-spheres. In this case, let W_{F_3,F_4} be the handlebody $W_{F_3} \cup H_Z$ where $\partial_+ W_{F_3,F_4} = \partial_+ V_{F_1,F_2}$. Now $V_{F_1,F_2} \cup_S W_{F_3,F_4}$ is a Heegaard splitting of a closed 3-manifold; see Figure 6.

Claim 3.5. *The Heegaard distance $d_{\mathcal{C}(S)}(V_{F_1,F_2}, W_{F_3,F_4})$ is n.*

Proof. Let D be the essential disk in W_{F_3,F_4} bounded by β. Suppose, on the contrary, that the Heegaard distance is $k < n$. Then there is a geodesic

$$\mathcal{G} = \{a_0 = \partial B_1, \dots, a_k = \partial D_1\},$$

where $k \leq n-1$, B_1 is an essential disk in V_{F_1,F_2}, and D_1 is an essential disk in W_{F_3,F_3}. $\alpha_i \cap \beta \neq \varnothing$, for any $0 \leq i \leq k-1$. If not, the distance of $V_{F_1,F_2} \cup_S W$ would be at most $k < n$. Similarly, D_1 is not isotopic to D.

Then we have two cases:

(a) $D_1 \cap D = \varnothing$. Then ∂D_1 lies in one of S_3 and S_4. We assume that ∂D_1 lies in S_3. The other case is similar. Hence $\psi_{F_3}(\partial D_1) = \delta$. By Lemma 2.3, $\operatorname{diam}_{S_3}(\mathcal{D}(\mathcal{G})) \leq \mathcal{M}$. Since $\pi_{S_3}(\partial B_1) \in \pi_{S_3}(\mathcal{D}(V_{F_1,F_2}))$, we have

$$d_{\mathcal{C}(S_3)}(\pi_{S_3}(\mathcal{D}(V_{F_1,F_2})), \partial D_1) \leq \mathcal{M}.$$

Hence,

$$d_{\mathcal{C}(F_3)}(\psi_{F_3}(\mathcal{D}(V_{F_1,F_2})), \psi_{F_3}(\partial D_1) = \delta) \leq \mathcal{M},$$

a contradiction.

(b) $D_1 \cap D \neq \varnothing$. Let c be an outermost arc of $D_1 \cap D$ on D_1. This means that c, together with a subarc $\delta^* \subset \partial D_1$, bounds a disk D_c such that $D_c \cap D = c$. We assume that $\partial D_c \subset S_4$. The other case is similar. By Lemma 2.3, $\operatorname{diam}_{S_4}(\mathcal{G}) \leq \mathcal{M}$. Hence

$$d_{\mathcal{C}(F_4)}(\psi_{F_4}(\mathcal{D}(V_{F_1,F_2})), \psi_{F_4}(\partial D_1)) \leq \mathcal{M}.$$

Note that $\psi_{F_4}(\partial B_1) \in \mathcal{D}(H_Z)$. Then by the same argument in (a), it is impossible. □

Now we prove the proposition for $n = 1$. It is known that if a Heegaard splitting has distance 1, there are on the Heegaard surface two disjoint nonisotopic essential simple closed curves that bound essential disks in different compression bodies. That is to say, a distance-1 Heegaard splitting is always weakly reducible. For a reducible Heegaard splitting, since there is an essential simple closed curve in the Heegaard surface bounding essential disks in both of these two compression bodies, it has distance zero. Hence it is only needed to prove the proposition for weakly reducible and irreducible Heegaard splittings.

Let M_1 and M_2 be two 3-manifolds with homeomorphic connected boundary. For any homeomorphism f from ∂M_1 to ∂M_2, let M_f be the manifold obtained by gluing M_1 and M_2 along f. Suppose M_i has a Heegaard splitting $V_i \cup_{S_i} W_i$ for $i = 1, 2$. In this case, M_f has a natural Heegaard splitting called the amalgamation of $V_1 \cup_{S_1} W_1$ and $V_2 \cup_{S_2} W_2$. The following facts are well known:

(1) If the gluing map f is complicated enough, then the amalgamation of $V_1 \cup_{S_1} W_1$ and $V_2 \cup_{S_2} W_2$ is unstabilized; see [Lackenby 2004; Bachman et al. 2006; Li 2010].

(2) If both $V_1 \cup_{S_1} W_1$ and $V_2 \cup_{S_2} W_2$ have high distance, then the amalgamation of $V_1 \cup_{S_1} W_1$ and $V_2 \cup_{S_2} W_2$ is unstabilized and irreducible; see [Kobayashi and Qiu 2008; Yang and Lei 2009].

Now let $M_i = V_i \cup_{S_i} W_i$ be a Heegaard splitting of genus two such that ∂M_i is a torus, and $d(S_i) > 8$ for $i = 1, 2$, then, by the main result in [Kobayashi and Qiu 2008], the amalgamation of $V_1 \cup_{S_1} W_1$ and $V_2 \cup_{S_2} W_2$, say $V \cup_S W$, is unstabilized.

Suppose that $g \geq 4$. By the above argument, there exist a Heegaard splitting $M_1 = V_1 \cup_{S_1} W_1$ of genus $g-1$ such that $g(\partial M_1) = 2$ and $d(S_1) \geq 2g$, and a Heegaard splitting $V_2 \cup_{S_2} W_2$ of genus three such that $g(\partial M_2) = 2$ and $d(S_2) \geq 2g$. Hence both M_1 and M_2 are hyperbolic. By the main result in [ibid.], the amalgamation of $V_1 \cup_{S_1} W_1$ and $V_2 \cup_{S_2} W_2$, say $M = V \cup_S W$, is unstabilized and weakly reducible. Furthermore, $g(S) = g$. By Thurston's theorem, both M_1 and M_2 have hyperbolic structures with totally geodesic boundaries. Hence M is hyperbolic. □

Remark 3.6. The strongly irreducible Heegaard splitting $V \cup_S W$ where both V and W contain only one essential separating disk up to isotopy independently is always a minimal Heegaard splitting of $M = V \cup_S W$. Li [2010] defined a subcomplex $\mathcal{U}(F_1)$, for $F_1 \subset \partial_- V$ and proved that for any handlebody H attached to M along F_1, if $d_{\mathcal{C}(F_1)}(\mathcal{U}(F_1), \mathcal{D}(H))$ is larger than a constant $\mathcal{K}$ which depends on M and H, then the new generated Heegaard splitting $V_{F_1} \cup_S W$ is still the minimal Heegaard splitting of $M^{F_1} = V_{F_1} \cup_S W$. Similar to the other boundaries of M. Now in our construction of distance-n (≥ 2) strongly irreducible Heegaard

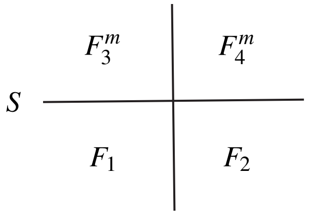


Figure 7. Heegaard splitting V.

splitting (for $n = 2$, see Section 5), we can choose a full simplex X in F_1 such that $d_{\mathcal{C}(F_1)}(\psi_{F_1}(\mathcal{D}(W)), \mathcal{D}(H_X))$ is large enough and $d_{\mathcal{C}(F_1)}(\mathcal{U}(F_1), \mathcal{D}(H_X))$ is larger than $\mathcal{K}$. Then the new Heegaard splitting $V_{F_1} \cup_S W$ is still the minimal Heegaard splitting of $M^{F_1} = V_{F_1} \cup_S W$ and has the same distance.

4. Proof of Theorem 1.3

Proposition 4.1. *For any integers $g \geq 2$ and $n \geq 4$, there are infinitely many nonhomeomorphic closed 3-manifolds which admit distance-n, genus-g Heegaard splittings.*

Proof. Let S_g be a closed surface of genus g. By Lemma 2.6, for each $m \geq 2$, there is a geodesic $\mathcal{G}^m = \{\alpha = a_0^m, a_1^m, \ldots, a_{n-1}^m, a_n^m = \beta^m\}$ in $\mathcal{C}(S_g)$ such that

(1) a_i^m is nonseparating in S_g for $1 \leq i \leq n-1$, α and β^m are two essential separating simple closed curves on S_g,

(2) $m\mathcal{M} + 2 \leq d_{\mathcal{C}(S^{a_i^m})}(a_{i-1}^m, a_{i+1}^m) \leq m\mathcal{M} + 6$, where S^{a_i} is the surface $S - N(a_i)$ for $1 \leq i \leq n-1$, and

(3) one component of $S_g - \beta^m$ has genus one.

Without loss of generality, we assume that $\mathcal{M} \geq 6$. Let M_m be the manifold obtained by attaching two 2-handles to $S_g \times [-1, 1]$ along $\alpha \times \{-1\}$ and $\beta^m \times \{1\}$. We also use S_g representing the surface $S_g \times \{0\}$. Now M_m has a Heegaard splitting as $V_m \cup_{S_g} W_m$, where V_m is the compression body obtained by attaching a 2-handle to $S \times [-1, 0]$ along $\alpha \times \{-1\}$, and W_m is the manifold obtained by attaching a 2-handle to $S \times [0, 1]$ along $\beta^m \times \{1\}$. Then $\partial_- V_m$ contains two components F_1 and F_2, and $\partial_- W_m$ contains two components F_3^m and F_4^m; see Figure 7.

By the proof of Theorem 1.1 ($n \neq 2$), there is a closed 3-manifold M_m^* which admits a distance-n Heegaard splitting $V_m^* \cup_{S_g} W_m^*$, where V_m^* is obtained by attaching handlebodies H_{X_1} and H_{X_2} to V_m along F_1 and F_2, and W_m^* is obtained by attaching handlebodies H_{Y_1} and H_{Y_2} to W_m along F_3^m and F_4^m such that

$$d_{\mathcal{C}(F_i)}(\psi_{F_i}(\beta^m), \mathcal{D}(H_{X_i})) \geq \mathcal{M} + 15 \quad \text{for } i = 1, 2,$$
$$d_{\mathcal{C}(F_i)}(\psi_{F_i}(\alpha), \mathcal{D}(H_{Y_i})) \geq \mathcal{M} + 15 \quad \text{for } i = 3, 4.$$

Replace M_m^*, V_m^* and W_m^* by M_m, V_m and W_m. Now

$$\mathcal{G}^m = \{\alpha = a_0^m, a_1^m, \dots, a_{n-1}^m, a_n^m = \beta^m\}$$

is a geodesic of $\mathcal{C}(S_g)$ realizing the distance of $M_m = V_m \cup_{S_g} W_m$.

Claim 4.2. *Let*

$$\mathcal{G} = \{b_0, \dots, b_n\}$$

be a geodesic of $\mathcal{C}(S_g)$ realizing the distance of $V_m \cup_{S_g} W_m$. Then

$$b_i = a_i^m$$

for any $1 \le i \le n-1$.

Proof. Let S_1 and S_2 be the two components of $S_g - \alpha$. We assume that b_0 bounds a disk B_0 in V_m, and b_n bounds a disk D_n in W_m. We first prove that α (resp. β^m) is disjoint from b_1 (resp. b_{n-1}).

Let B be the essential disk bounded by α in V_m. Suppose, on the contrary, that $\alpha \cap b_1 \neq \varnothing$. Hence b_0 is not isotopic to $a_0^m = \alpha$. Then there are two cases:

(a) $B_0 \cap B \neq \varnothing$. Let a be an outermost arc of $B_0 \cap B$ on B_0. It means that a, together with a subarc of $\gamma \subset \partial B_0$, bounds a disk B_γ such that $B_\gamma \cap B = a$. We assume that $\gamma \subset S_1$. The other case is similar. By the argument in Section 3, $\psi_{F_1}(\partial B_0)$ bounds an essential disk in H_{X_1}. But with $b_1 \cap \partial S_1 \neq \varnothing$, it implies that $d_{\mathcal{C}(S_1)}(b_0, b_n) \le \mathcal{M}$. Hence $d_{\mathcal{C}(F_1)}(\psi_{F_1}(b_n), \mathcal{D}(H_{X_1})) \le \mathcal{M}$.

(b) $B_0 \cap B = \varnothing$. Since $b_1 \cap \alpha \neq \varnothing$, B_0 is not isotopic to B. Then ∂B_0 is essential in S_1 or S_2. We assume that $\partial B_0 \subset S_1$. The other case is similar. Hence by the arguments in the previous case, $d_{\mathcal{C}(F_1)}(\psi_{F_1}(b_n), \mathcal{D}(H_{X_1})) \le \mathcal{M}$.

However, since the Heegaard distance is at least four and $\alpha = \partial S_1 = \partial S_2$ bounds an essential disk in V^m, the curve α is disk-busting for W^m and W^m can not be the I-bundle over S_1 or S_2. Then by Lemma 2.4,

$$\operatorname{diam}_{\mathcal{C}(S_1)}(\mathcal{D}(W^m)) \le 12$$

and

$$\operatorname{diam}_{\mathcal{C}(S_2)}(\mathcal{D}(W^m)) \le 12.$$

Hence $\operatorname{diam}_{\mathcal{C}(F_1)}(\psi_{F_1}(\mathcal{D}(W^m))) \le 12$ and $\operatorname{diam}_{\mathcal{C}(F_2)}(\psi_{F_2}(\mathcal{D}(W^m))) \le 12$. Together with (a) and (b), by the triangle inequality, we have

$$d_{\mathcal{C}(F_1)}(\psi_{F_1}(\beta^m), \mathcal{D}(H_{X_1})) \le \mathcal{M} + 12.$$

It contradicts the choice of X_1 in F_1.

Let $\mathcal{G}^* = \{\alpha = a_0^m, b_1, \dots, b_{n-1}, a_n^m\}$ be a new geodesic realizing the distance of $V_m \cup_{S_g} W_m$. Now we prove that b_1 is isotopic to a_1^m. The other case is similar.

Suppose, otherwise, that b_1 is not isotopic to a_1^m. Note that b_i is not isotopic to a_1^m. Otherwise, the distance of $V_m \cup_{S_g} W_m$ would be at most $n-1$. Let $S^{a_1^m}$ be the surface $S_g - N(a_1^m)$, where $N(a_1^m)$ is an open regular neighborhood of a_1^m on S_g. By Lemma 2.3,

$$d_{\mathcal{C}(S^{a_1^m})}(\pi_{S^{a_1^m}}(a_0^m), \pi_{S^{a_1^m}}(a_n^m)) \le \mathcal{M}.$$

Now let's consider the shorter geodesic

$$\mathcal{G}^{**} = \{a_2^m, \dots, a_{n-1}^m, a_n^m = \beta^m\},$$

which is a subgeodesic of

$$\mathcal{G}^m = \{\alpha = a_0^m, a_1^m, \dots, a_{n-1}^m, a_n^m = \beta^m\}.$$

By the definition of geodesic in the curve complex, a_i^m is not isotopic to a_1^m for any $i \ge 2$. By Lemma 2.3 again,

$$d_{\mathcal{C}(S^{a_1^m})}(\pi_{S^{a_1^m}}(a_2^m), \pi_{S^{a_1^m}}(a_n^m)) \le \mathcal{M}.$$

Hence

$$d_{\mathcal{C}(S^{a_1^m})}(\pi_{S^{a_1^m}}(a_0^m), \pi_{S^{a_1^m}}(a_2^m)) \le 2\mathcal{M}.$$

This contradicts our assumption on $m\mathcal{M} \le d_{\mathcal{C}(S^{a_1^m})}(\pi_{S^{a_1^m}}(a_0^m), \pi_{S^{a_1^m}}(a_2^m))$ and $m \ge 2$. Hence b_1 is isotopic to a_1^m. □

Replace $M_m = V_m \cup_{S_g} W_m$ by $M_m = V_m \cup_{S_g^m} W_m$.

The following claim reveals the connection between geodesics in the curve complex and closed 3-manifolds:

Claim 4.3. *For any t, s such that $2 \le t \ne s \in N$, either*

(1) *$M_t = V_t \cup_{S_g^t} W_t$ and $M_s = V_s \cup_{S_g^s} W_s$ are two different 3-manifolds up to homeomorphism, or,*

(2) *M_t is homeomorphic to M_s, but $V_t \cup_{S_g^t} W_t$ and $V_s \cup_{S_g^s} W_s$ are two different Heegaard splittings of M_t up to homeomorphic equivalence.*

Proof. Suppose that M_t is homeomorphic to M_s for some $t, s \in N$ where $2 \le t, s$ and $t \ne s$. If (2) fails, then $V_t \cup_{S_g^t} W_t$ and $V_s \cup_{S_g^s} W_s$ are homeomorphic. It means that there is a homeomorphism f from M_t to M_s such that $f((S_g^t; V_t, W_t)) = (S_g^s; V_s, W_s)$. We assume that $f(V_t) = V_s$ and $f(W_t) = W_s$. The other case is similar. It is well known that f induces an isomorphism from $\mathcal{C}(S_g^t)$ to $\mathcal{C}(S_g^s)$, still denoted by f. Then for the geodesic

$$\mathcal{G}^t = \{\alpha = a_0^t, a_1^t, \dots, a_{n-1}^t, a_n^t = \beta^t\}$$

which realizes the distance of $V_t \cup_{S_g^t} W_t$, $f(\mathcal{G})$ is also a geodesic in $\mathcal{C}(S_g^s)$ realizing the distance of $V_s \cup_{S_g^s} W_s$. By Claim 4.2, $f(a_j^t)$ is isotopic to a_j^s for $1 \le j \le n-1$.

Since $f(a_2^t)$ is isotopic to a_2^s, we can perform an isotopy on S_g^s such that the composition of f with the isotopy gives an homeomorphism $f^\star$ from S_g^t to S_g^t and $f^\star(a_2^t) = a_2^s$, $f^\star(V_t) = V_s$, $f^\star(W_t) = W_s$. It's also true that $f^\star$ induces an automorphism from $\mathcal{C}(S_g^t)$ to $\mathcal{C}(S_g^s)$, denoted by $f^\star$ too. Thus $f^\star(\mathcal{G}^t)$ is also a geodesic realizing the distance of $V_s \cup_{S_g^s} W_s$. By Claim 4.2 again, for any $1 \le j \le n-1$, $f^\star(a_j^t)$ is still isotopic to a_j^s. Hence $f^\star(a_1^t)$ (resp. $f^\star(a_3^t)$) is isotopic to a_1^s (resp. a_3^t).

Let $S^{a_2^t}$ be the surface $S_g^t - N(a_2^t)$, where $N(a_2^t)$ is an open regular neighborhood of a_2^t on S_g^t, and let $S^{a_2^s}$ be the surface of $S_g^s - N(a_2^s)$. Then $f^\star(S^{a_2^t}) = S^{a_2^s}$ and $f^\star|_{S^{a_2^t}}$ is a homeomorphism. Hence $f^\star$ also induces an isomorphism from $\mathcal{C}(S^{a_2^t})$ to $\mathcal{C}(S^{a_2^s})$, still denoted by $f^\star$. Now we also assume $a_1^t \cap a_2^t = \varnothing$ and $a_3^t \cap a_2^t = \varnothing$. Thus $f^\star(a_1^t) \cap (f^\star(a_2^t) = a_2^s) = \varnothing$ and $f^\star(a_3^t) \cap (f^\star(a_2^t) = a_2^s) = \varnothing$. Then $d_{\mathcal{C}(S^{a_2^t})}(a_1^t, a_3^t) = d_{\mathcal{C}(S^{a_2^s})}(f^\star(a_1^t), f^\star(a_3^t))$. On the other hand, $f^\star(a_1^t)$ (resp. $f^\star(a_3^s)$) must be isotopic to a_1^s (resp. a_3^s) in $S^{a_2^s}$. For if not, then after removing possible bigon capped by them, they bound no annuli in $S^{a_2^s}$, and thus they bound no annuli and bigon in S_g^s. By bigon criterion [Farb and Margalit 2012, Proposition 1.7], they realize the geometry intersection number. Since they are isotopic in S_g^s, they must bound an annulus in S_g^s. So

$$d_{\mathcal{C}(S^{a_2^t})}(a_1^t, a_3^t) = d_{\mathcal{C}(S^{a_2^s})}(f^\star(a_1^t), f^\star(a_3^t)),$$
$$d_{\mathcal{C}(S^{a_2^s})}(f^\star(a_1^t), f^\star(a_3^t)) = d_{\mathcal{C}(S^{a_2^s})}(a_1^s, a_3^s).$$

It means that

$$d_{\mathcal{C}(S^{a_2^t})}(a_1^t, a_3^t) = d_{\mathcal{C}(S^{a_2^s})}(a_1^s, a_3^s).$$

However, by the assumption,

$$t\mathcal{M} + 2 \le d_{\mathcal{C}(S^{a_2^t})}(a_1^t, a_3^t) \le t\mathcal{M} + 6,$$
$$s\mathcal{M} + 2 \le d_{\mathcal{C}(S^{a_2^s})}(a_1^t, a_3^t) \le s\mathcal{M} + 6,$$
$$\mathcal{M} \ge 6,$$

we have

$$d_{\mathcal{C}(S^{a_2^t})}(a_1^t, a_3^t) \ne d_{\mathcal{C}(S^{a_2^s})}(a_1^s, a_3^s),$$

a contradiction. □

The Waldhausen conjecture proved by Johanson [1990; 1995] and Li [2006; 2007] implies that, for any positive integer g, an atoroidal closed 3-manifold M admits only finitely many Heegaard splittings of genus g up to homeomorphism. Since M_t admits a Heegaard splitting with distance at least four, it is atoroidal for any $t \ge 2$; see [Hartshorn 2002; Scharlemann 2006]. Now Theorem 1.3 immediately follows from Claim 4.3 and the Waldhausen conjecture. □

5. Proof of Theorem 1.1 ($n = 2$)

We rewrite the second part of Theorem 1.1:

Proposition 5.1. *For any integer $g \geq 2$, there is a closed hyperbolic 3-manifold which admits a distance-2 Heegaard splitting of genus g.*

Proof. By Remark 1.2(2), there is a hyperbolic closed 3-manifold which admits a distance-2 Heegaard splitting of genus two. So we only need to prove it for $g \geq 3$.

Assumption 1. Let S be a closed surface of genus g. By Lemma 2.6, there are two separating essential simple closed curves α and γ such that

(1) $d_{\mathcal{C}(S)}(\alpha, \gamma) = 2$,

(2) one component of $S - \alpha$, say S_1, has genus one while the component of $S - \alpha$, say S_2, has genus $g - 1$,

(3) one component of $S - \gamma$, say S_3, has genus one, while the component of $S - \gamma$, say S_4, has genus $g - 1$,

(4) there is a nonseparating slope β on S such that α and γ are disjoint from β, and $d_{\mathcal{C}(S^\beta)}(\alpha, \gamma) > 4$, where S^β is the surface $S - \eta(\beta)$, and

(5) $\beta \subset S_2 \cap S_4$.

Let V be the compression body obtained by attaching a 2-handle to $S \times [0, 1]$ along a separating curve $\alpha \times \{1\}$, and let W be the compression body obtained by attaching a 2-handle to $S \times [-1, 0]$ along a separating curve $\gamma \times \{-1\}$. Denote $S \times \{0\}$ by S too. Then $V \cup_S W$ is a Heegaard splitting. Since V contains only one essential disk B with $\partial B = \alpha$ up to isotopy, and W contains only one essential disk D with $\partial D = \gamma$ up to isotopy, $d_{\mathcal{C}(S)}(V, W) = 2$.

Let F_1 and F_2 be the components of $\partial_- V$, such that F_i is homeomorphic to $S_i \cup B$ for $i = 1, 2$. Similarly, let F_3 and F_4 be the components of $\partial_- W$ such that F_i is homeomorphic to $S_i \cup D$ for $i = 3, 4$. Then both S_1 and S_3 are once-punctured tori, and F_1 and F_3 are two tori; see Figure 2. Furthermore, both F_3 and F_4 have genus at least two. Now B cuts V into two manifolds $F_1 \times I$ and $F_2 \times I$, and D cuts W into two manifolds $F_3 \times I$ and $F_4 \times I$.

Since $d_{\mathcal{C}(S)}(V, W) = 2$, $\gamma \cap S_i \neq \varnothing$ for $i = 1, 2$, and $\alpha \cap S_i \neq \varnothing$ for $i = 3, 4$. Hence $\psi_{F_i}(\gamma) \neq \varnothing$ for $i = 1, 2$, and $\psi_{F_i}(\alpha) \neq \varnothing$ for $i = 3, 4$, where ψ is defined in Section 3.

Assumption 2. (1) Let δ be an essential simple closed curve on the torus F_1 such that $d_{\mathcal{C}(F_2)}(\psi_{F_2}(\gamma), \delta) \geq 5$.

(2) Let X be a full complex of $\mathcal{C}(F_2)$ such that $d_{\mathcal{C}(F_2)}(\psi_{F_2}(\gamma), \mathcal{D}(H_X)) \geq 24$, where H_X is the handlebody obtained by attaching 2-handles to F_2 along the vertices of X then 3-handles to cap off the spherical boundary components.

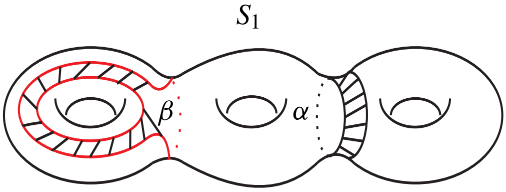


Figure 8. Essential annulus.

Let $V_{F_2} = V \cup H_X$, and let V_{F_1,F_2} be the handlebody obtained by doing a surgery on V_{F_2} along the slope δ on F_1. By Assumption 1, $g(S_3) = 1$, $g(S_4) \geq 2$, V_{F_1,F_2} is not an I-bundle over S_i for $i = 3, 4$. By Lemma 2.4, $\operatorname{diam}_{\mathcal{C}(S_i)}(\pi_{S_i}(\mathcal{D}(V_{F_1,F_2}))) \leq 12$ for $i = 3, 4$.

Assumption 3. (1) Let r be an essential simple closed curve on the torus F_3 such that $d_{\mathcal{C}(F_3)}(\psi_{F_3}(\mathcal{D}(V_{F_1,F_2})), r) \geq 24$.

(2) Let Y be a full complex of $\mathcal{C}(F_4)$ such that $d_{\mathcal{C}(F_4)}(\psi_{F_4}(\mathcal{D}(V_{F_1,F_2})), \mathcal{D}(H_Y)) \geq 24$, where H_Y is the handlebody obtained by attaching 2-handles to F_4 along the vertices of Y then 3-handles to cap off the spherical boundary components.

Let $W_{F_4} = W \cup H_Y$, and let W_{F_3,F_4} be the handlebody obtained by doing a surgery on W_{F_4} along the slope r on F_3. Now both $M^* = V_{F_2} \cup_S W_{F_4}$ and $V_{F_1,F_2} \cup_S W_{F_3,F_4}$ are Heegaard splittings. Furthermore, we can prove that these two Heegaard splittings have distance two by arguments in the proof of Proposition 3.1.

Now we consider $M^* = V_{F_2} \cup_S W_{F_4}$. Note that M^* has only two toroidal boundary components. Since the distance of $V_{F_2} \cup_S W_{F_4}$ is two, M^* is irreducible and ∂-irreducible.

Claim 5.2. *M^* is atoroidal.*

Proof. Suppose, on the contrary, that M^* contains an essential torus T. Since the distance of $V_{F_2} \cup_S W_{F_4}$ is two, $V_{F_2} \cup_S W_{F_4}$ is strongly irreducible. By Schultens' lemma [Schultens 1993], we may assume that each component of $T \cap S$ is essential on both T and S. Hence each component of $T \cap V_{F_2}$ and $T \cap W_{F_4}$ is an incompressible annulus in V_{F_2} or W_{F_4}.

Let A_0 be one component of $T \cap V_{F_2}$. We first prove that there is one component of ∂A_0, say a_0, not isotopic to β.

Now V_{F_2} contains a ∂-compressing disk B^* of A_0. Note that A_0 has a ∂-compression disk B^* in V_{F_2}. By doing a surgery on A_0 along B^*, we get a disk B_0 in V_{F_2}. Since A_0 is essential, B_0 is essential. Suppose that the two components of ∂A_0 are isotopic to β. Since β is nonseparating on S, ∂B_0 bounds a once-punctured torus containing β; see Figure 8.

By Assumption 1, $\beta \subset S_2$. Since S_2 has genus $g - 1 \geq 2$, ∂B_0 is not isotopic to $\alpha = \partial S_2$. By standard outermost disk argument, $\psi_{F_2}(\partial B_0)$ bounds an essential disk in H_X. Therefore $d_{\mathcal{C}(F_2)}(\mathcal{D}(H_X), \psi_{F_2}(\beta)) \leq 1$. Since $\gamma \cap \beta = \varnothing$,

$d_{\mathcal{C}(F_2)}(\psi_{F_2}(\beta), \psi_{F_2}(\gamma)) \leq 2$. Hence $d_{\mathcal{C}(F_2)}(\mathcal{D}(H_X), \psi_{F_2}(\gamma)) \leq 3$. It contradicts Assumption 2.

Let A_1 be a component of $T \cap W_{F_4}$ which is incident to A_0 at a_0. This means that a_0 is one component of ∂A_1. We consider two cases:

Case 1. $a_0 \cap \alpha = \varnothing$ and $a_0 \cap \gamma = \varnothing$.

Recall the definition of the surface S^β. Since a_0 is not isotopic to β, $a_0 \cap S^\beta \neq \varnothing$. Since $\alpha, \gamma \subset S^\beta$,

$$d_{\mathcal{C}(S_\beta)}(\pi_{S^\beta}(a_0), \alpha) \leq 1,$$
$$d_{\mathcal{C}(S^\beta)}(\gamma, \pi_{S^\beta}(a_0)) \leq 1.$$

Hence $d_{\mathcal{C}(S^\beta)}(\alpha, \gamma) \leq 2$. This contradicts Assumption 1.

Case 2. $a_0 \cap (\alpha \cup \gamma) \neq \varnothing$.

We assume that $a_0 \cap \alpha \neq \varnothing$. By the above argument, B_0 is an essential disk in V_{F_2} such that ∂B_0 is disjoint from a_0. Furthermore, ∂B_0 is not isotopic to α. Since B cuts V_{F_2} into $F_1 \times I$ and a handlebody H such that $S_2 \cup B = \partial H$, $\partial B_0 \cap S_2 \neq \varnothing$. Furthermore, all outermost disks of $B_0 \cap B$ on B_0 lie in H. Hence a curve in $\pi_{S_2}(\partial B_0)$ bounds an essential disk in H. This means a curve in $\psi_{F_2}(\partial B_0)$ bounds an essential disk in H_X.

If $a_0 \cap \gamma = \varnothing$, then

$$d_{\mathcal{C}(F_2)}(\psi_{F_2}(\partial B_0), \psi_{F_2}(\gamma)) \leq d_{\mathcal{C}(F_2)}(\psi_{F_2}(\partial B_0), \psi_{F_2}((a_0)) + d_{\mathcal{C}(F_2)}(\psi_{F_2}(a_0), \psi_{F_2}(\gamma)) \leq 4.$$

It contradicts Assumption 2. Hence $a_0 \cap \gamma \neq \varnothing$, and $\psi_{F_4}(a_0) \neq \varnothing$.

Since A_1 is an essential annulus in W_{F_4}, there is an essential disk D_0 obtained by doing boundary compression on A_1 in W_{F_4}. Furthermore $\partial D_0 \cap a_0 = \varnothing$. Since D cuts W_{F_4} into $F_3 \times I$ and a handlebody H^* containing H_Y, all outermost disks of $D_0 \cap D$ in D_0 lie in H^*. Hence $\psi_{F_4}(\partial D_0)$ bounds an essential disk in H_Y. Hence a curve in $\pi_{S_4}(\partial D_0) \neq \varnothing$. Since $\partial D_0 \cap a_0 = \varnothing$, by Lemma 2.2, $d_{\mathcal{C}(S_4)}(\pi_{S_4}(\partial D_0), \pi_{S_4}(a_0)) \leq 2$. According to the definition of ψ_{F_4}, $d_{\mathcal{C}(F_4)}(\psi_{F_4}(\partial D_0), \psi_{F_4}(a_0)) \leq 2$.

Recall that the essential disk B_0 is obtained by doing a boundary compression on A_0 in V_{F_2}. Since the distance of $V_{F_2} \cup_S W_{F_4}$ is two, $\partial B_0 \cap \gamma \neq \varnothing$. Since $g(S_3) = 1$ and $g(S_4) \geq 2$, V_{F_2} is not an I-bundle over S_4. By Lemma 2.4, $d_{\mathcal{C}(S_4)}(\pi_{S_4}(\partial B_0), \pi_{S_4}(\alpha)) \leq 12$. Hence

$$d_{\mathcal{C}(F_4)}(\psi_{F_4}(\partial B_0), \psi_{F_4}(\alpha)) \leq 12.$$

Since $\partial B_0 \cap a_0 = \varnothing$,

$$d_{\mathcal{C}(F_4)}(\psi_{F_4}(\partial B_0), \psi_{F_4}(a_0)) \leq 2.$$

The above inequalities implies that

$$\begin{aligned}
d_{\mathcal{C}(F_4)}&(\psi_{F_4}(\partial D_0),\, \psi_{F_4}(\alpha)) \\
&\le d_{\mathcal{C}(F_4)}(\psi_{F_4}(\partial D_0),\, \psi_{F_4}(a_0)) + d_{\mathcal{C}(F_4)}(\psi_{F_4}(\partial B_0),\, \psi_{F_4}(a_0)) \\
&\qquad\qquad + d_{\mathcal{C}(F_4)}(\psi_{F_4}(\partial B_0),\, \psi_{F_4}(\alpha)) \\
&\le 16.
\end{aligned}$$

It contradicts Assumption 3. □

Claim 5.3. *M^* is anannular.*

Proof. Since the distance of $M^* = V_{F_2} \cup_S W_{F_4}$ is two, $M^* = V_{F_2} \cup_S W_{F_4}$ is strongly irreducible and boundary irreducible. Suppose, on the contrary, that M^* contains an essential annulus A. Then there are two cases:

(a) ∂A lies in the same boundary component of M^*. Without assumption, we assume that $\partial A \subset F_2$. Hence the boundary of closed regular neighborhood of $F_2 \cup A$ consists of three tori, denoted by F_2, T_1 and T_2. By Claim 5.2, both T_1 and T_2 are inessential in M^*. Since the boundary of M^* is not connected, one of T_1 and T_2, says T_1, is compressible and the other one is boundary parallel. This means that M^* is a Seifert manifold, whose orbifold is an annulus with at most one cone point. By [Moriah and Schultens 1998], each irreducible Heegaard splitting of M^* is vertical or horizontal. Hence each irreducible Heegaard splitting of M^* has genus two. So each genus at least three Heegaard splitting of M^* is stabilized and reducible. A contradiction.

(b) ∂A lies in different boundary components of M^*. Then the boundary of $A \cup \partial M^*$ consists of three tori, denoted by T, F_2 and F_4. By Claim 5.2, T is inessential in M^*. It is not hard to see that T is not boundary parallel to F_2 or F_4. Then T is compressible in M^*. So M^* is a Seifert manifold, whose orbifold is an annulus with at most one cone point. By [ibid.] again, each irreducible Heegaard splitting of M^* is vertical or horizontal. Hence each irreducible Heegaard splitting of M^* has genus two. So each genus at least three Heegaard splitting of M^* is stabilized and reducible. A contradiction. □

Now M^* is a hyperbolic 3-manifold, $M^* = V_{F_2} \cup_S W_{F_4}$ is a distance-2 Heegaard splitting of genus g. Furthermore, M^* contains two toral boundary components F_1 and F_3. By the main results in [Agol 2010; Lackenby and Meyerhoff 2013], there are at most ten slopes δ on F_1 such that the manifold $M^*(\delta)$ obtained by doing Dehn filling on M^* along δ is nonhyperbolic. By Assumption 2, there are infinitely many slopes δ so that $M^*(\delta)$ has a distance-2 Heegaard splitting of genus g. Hence there is at least one slope δ on F_1 such that $M^*(\delta)$ is hyperbolic and $M^*(\delta)$ admits a distance-2 Heegaard splitting of genus g. Similarly, by Assumption 3, there is a hyperbolic closed manifold which admits a distance-2 Heegaard splitting of genus g. □

Acknowledgements

The authors thank Mario Eudave-Muñoz and Jiming Ma for some helpful discussions. The authors also thank the referee for the careful reading and pointing out a shorter proof of Claim 5.3.

References

[Agol 2010] I. Agol, "Bounds on exceptional Dehn filling II", *Geom. Topol.* **14** (2010), 1921–1940.

[Bachman et al. 2006] D. Bachman, S. Schleimer, and E. Sedgwick, "Sweepouts of amalgamated 3-manifolds", *Algebr. Geom. Topol.* **6** (2006), 171–194. MR 2006k:57057 Zbl 1099.57016

[Campisi and Rathbun 2012] M. M. Campisi and M. Rathbun, "High distance knots in closed 3-manifolds", *J. Knot Theory Ramifications* **21**:2 (2012), Article ID #1250017. MR 2885479 Zbl 1250.57031

[Eudave-Muñoz 1999] M. Eudave-Muñoz, "Incompressible surfaces in tunnel number one knot complements", *Topology Appl.* **98**:1-3 (1999), 167–189. MR 2000h:57010 Zbl 0934.57012

[Evans 2006] T. Evans, "High distance Heegaard splittings of 3-manifolds", *Topology Appl.* **153**:14 (2006), 2631–2647. MR 2007j:57020 Zbl 1107.57011

[Farb and Margalit 2012] B. Farb and D. Margalit, *A primer on mapping class groups*, Princeton Mathematical Series **49**, Princeton University Press, 2012. MR 2012h:57032 Zbl 1245.57002

[Hartshorn 2002] K. Hartshorn, "Heegaard splittings of Haken manifolds have bounded distance", *Pacific J. Math.* **204**:1 (2002), 61–75. MR 2003a:57037 Zbl 1065.57021

[Harvey 1981] W. J. Harvey, "Boundary structure of the modular group", pp. 245–251 in *Riemann surfaces and related topics* (Stony Brook, NY, 1978), edited by I. Kra and B. Maskit, Ann. of Math. Stud. **97**, Princeton University Press, 1981. MR 83d:32022 Zbl 0461.30036

[Hempel 2001] J. Hempel, "3-manifolds as viewed from the curve complex", *Topology* **40**:3 (2001), 631–657. MR 2002f:57044 Zbl 0985.57014

[Ido et al. 2014] A. Ido, Y. Jang, and T. Kobayashi, "Heegaard splittings of distance exactly n", *Algebr. Geom. Topol.* **14**:3 (2014), 1395–1411. MR 3190598 Zbl 1297.57029

[Johannson 1990] K. Johannson, "Heegaard surfaces in Haken 3-manifolds", *Bull. Amer. Math. Soc. (N.S.)* **23**:1 (1990), 91–98. MR 91d:57010 Zbl 0715.57006

[Johannson 1995] K. Johannson, *Topology and combinatorics of* 3*-manifolds*, Lecture Notes in Math. **1599**, Springer, Berlin, 1995. MR 98c:57014 Zbl 0820.57001

[Kobayashi and Qiu 2008] T. Kobayashi and R. Qiu, "The amalgamation of high distance Heegaard splittings is always efficient", *Math. Ann.* **341**:3 (2008), 707–715. MR 2009c:57013 Zbl 1140.57012

[Lackenby 2004] M. Lackenby, "The Heegaard genus of amalgamated 3-manifolds", *Geom. Dedicata* **109** (2004), 139–145. MR 2005i:57021 Zbl 1081.57018

[Lackenby and Meyerhoff 2013] M. Lackenby and R. Meyerhoff, "The maximal number of exceptional Dehn surgeries", *Invent. Math.* **191** (2013), 241–382.

[Li 2006] T. Li, "Heegaard surfaces and measured laminations, II: Non-Haken 3-manifolds", *J. Amer. Math. Soc.* **19**:3 (2006), 625–657. MR 2007g:57036 Zbl 1108.57015

[Li 2007] T. Li, "Heegaard surfaces and measured laminations, I: The Waldhausen conjecture", *Invent. Math.* **167**:1 (2007), 135–177. MR 2008h:57033 Zbl 1109.57012

[Li 2010] T. Li, "Heegaard surfaces and the distance of amalgamation", *Geom. Topol.* **14**:4 (2010), 1871–1919. MR 2011j:57027 Zbl 1207.57031

[Li 2012] T. Li, "Images of the disk complex", *Geom. Dedicata* **158** (2012), 121–136. MR 2922707 Zbl 1244.57039

[Li 2013] T. Li, "Small 3-manifolds with large Heegaard distance", *Math. Proc. Cambridge Philos. Soc.* **155**:3 (2013), 431–441. MR 3118411 Zbl 06266602

[Lustig and Moriah 2009] M. Lustig and Y. Moriah, "High distance Heegaard splittings via fat train tracks", *Topology Appl.* **156**:6 (2009), 1118–1129. MR 2011b:57025 Zbl 1214.57021

[Lustig and Moriah 2010] M. Lustig and Y. Moriah, "Horizontal Dehn surgery and genericity in the curve complex", *J. Topol.* **3**:3 (2010), 691–712. MR 2011k:57028 Zbl 1245.57022

[Lustig and Moriah 2012] M. Lustig and Y. Moriah, "Are large distance Heegaard splittings generic?", *J. Reine Angew. Math.* **670** (2012), 93–119. MR 2982693 Zbl 1248.57012

[Masur and Minsky 1999] H. A. Masur and Y. N. Minsky, "Geometry of the complex of curves, I: Hyperbolicity", *Invent. Math.* **138**:1 (1999), 103–149. MR 2000i:57027 Zbl 0941.32012

[Masur and Minsky 2000] H. A. Masur and Y. N. Minsky, "Geometry of the complex of curves, II: Hierarchical structure", *Geom. Funct. Anal.* **10**:4 (2000), 902–974. MR 2001k:57020 Zbl 0972.32011

[Masur and Schleimer 2013] H. A. Masur and S. Schleimer, "The geometry of the disk complex", *J. Amer. Math. Soc.* **26**:1 (2013), 1–62. MR 2983005 Zbl 1272.57015

[Minsky 1996] Y. N. Minsky, "A geometric approach to the complex of curves on a surface", pp. 149–158 in *Topology and Teichmüller spaces* (Katinkulta, 1995), edited by S. Kojima et al., World Scientific, River Edge, NJ, 1996. MR 2000g:32016 Zbl 0937.30027

[Minsky et al. 2007] Y. N. Minsky, Y. Moriah, and S. Schleimer, "High distance knots", *Algebr. Geom. Topol.* **7** (2007), 1471–1483. MR 2008k:57016 Zbl 1167.57002

[Moriah and Schultens 1998] Y. Moriah and J. Schultens, "Irreducible Heegaard splittings of Seifert fibered spaces are either vertical or horizontal", *Topology* **37**:5 (1998), 1089–1112. MR 99g:57021 Zbl 0926.57016

[Scharlemann 2006] M. Scharlemann, "Proximity in the curve complex: boundary reduction and bicompressible surfaces", *Pacific J. Math.* **228**:2 (2006), 325–348. MR 2008c:57035 Zbl 1127.57010

[Scharlemann and Tomova 2006] M. Scharlemann and M. Tomova, "Alternate Heegaard genus bounds distance", *Geom. Topol.* **10** (2006), 593–617. MR 2007b:57040 Zbl 1128.57022

[Schultens 1993] J. Schultens, "The classification of Heegaard splittings for (compact orientable surface) × S^1", *Proc. London Math. Soc.* (3) **67**:2 (1993), 425–448. MR 94d:57043 Zbl 0789.57012

[Yang and Lei 2009] G. Yang and F. Lei, "On amalgamations of Heegaard splittings with high distance", *Proc. Amer. Math. Soc.* **137**:2 (2009), 723–731. MR 2009h:57038 Zbl 1162.57013

Received June 18, 2014. Revised September 10, 2014.

RUIFENG QIU
DEPARTMENT OF MATHEMATICS
EAST CHINA NORMAL UNIVERSITY
DONGCHUAN ROAD 500
SHANGHAI, 200241
CHINA
rfqiu@math.ecnu.edu.cn

YANQING ZOU
DEPARTMENT OF MATHEMATICS
DALIAN NATIONALITIES UNIVERSITY
DALIAN, 116600
CHINA
yanqing@dlnu.edu.cn

QILONG GUO
SCHOOL OF MATHEMATICAL SCIENCES
PEKING UNIVERSITY
BEIJING, 100871
CHINA
guolong1999@yahoo.com.cn

Guidelines for Authors

Authors may submit articles at msp.org/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 or to Pacific Journal of Mathematics, University of California, Los Angeles, CA 90095–1555. Correspondence by email is requested for convenience and speed.

Manuscripts must be in English, French or German. A brief abstract of about 150 words or less in English must be included. The abstract should be self-contained and not make any reference to the bibliography. Also required are keywords and subject classification for the article, and, for each author, postal address, affiliation (if appropriate) and email address if available. A home-page URL is optional.

Authors are encouraged to use LaTeX, but papers in other varieties of TeX, and exceptionally in other formats, are acceptable. At submission time only a PDF file is required; follow the instructions at the web address above. Carefully preserve all relevant files, such as LaTeX sources and individual files for each figure; you will be asked to submit them upon acceptance of the paper.

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used but tags will be converted to the house format (see a current issue for examples).

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple or Mathematica. Many drawing tools such as Adobe Illustrator and Aldus FreeHand can produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, the authors should check with production by sending an email to pacific@math.berkeley.edu.

Each figure should be captioned and numbered, so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript will all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables, which should be used sparingly.

Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Page proofs will be made available to authors (or to the designated corresponding author) at a website in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

PACIFIC JOURNAL OF MATHEMATICS

Volume 275 No. 1 May 2015

0030-8730(201505)275:1;1-0