

Volume XVIII

Surveys in Differential Geometry 2013

Geometry and Topology

Lectures given at the Geometry and Topology conferences at Harvard University in 2011 and at Lehigh University in 2012

Huai-Dong Cao
Shing-Tung Yau
editors

 International Press

Preface

Each year the *Journal of Differential Geometry* (JDG) sponsors a conference on Geometry and Topology. The conference is held every third year at Harvard University, and other years at Lehigh University.

The current volume includes papers presented by several speakers at both the 2011 conference at Harvard and the 2012 conference at Lehigh. We have articles by Simon Brendle, on the Lagrangian minimal surface equation and related problems; by Sergio Cecotti and Cumrun Vafa, concerning classification of complete $N = 2$ supersymmetric theories in 4 dimensions; by F. Reese Harvey and H. Blaine Lawson Jr., on existence, uniqueness, and removable singularities for non-linear PDEs in geometry; by János Kollár, concerning links of complex analytic singularities; by Claude LeBrun, on Calabi energies of extremal toric surfaces; by Mu-Tao Wang, concerning mean curvature flows and isotopy problems; and by Steve Zelditch, on eigenfunctions and nodal sets.

We are grateful to the many distinguished geometers and topologists who presented invited talks at these two conferences, especially those who contributed articles to this volume of the *Surveys in Differential Geometry* book series.

Huai-Dong Cao
Lehigh University

Shing-Tung Yau
Harvard University

Contents

<i>Preface</i>	v
On the Lagrangian minimal surface equation and related problems	
SIMON BRENDLE	1
Classification of complete $\mathcal{N} = 2$ supersymmetric theories in 4 dimensions	
SERGIO CECOTTI AND CUMRUN VAFA	19
Existence, uniqueness and removable singularities for nonlinear partial differential equations in geometry	
F. REESE HARVEY AND H. BLAINE LAWSON, JR.	103
Links of complex analytic singularities	
JÁNOS KOLLÁR	157
Calabi energies of extremal toric surfaces	
CLAUDE LEBRUN	195
Mean curvature flows and isotopy problems	
MU-TAO WANG	227
Eigenfunctions and nodal sets	
STEVE ZELDITCH	237

On the Lagrangian minimal surface equation and related problems

Simon Brendle

ABSTRACT. We give a survey of various existence results for minimal Lagrangian graphs. We also discuss the mean curvature flow for Lagrangian graphs.

1. Background on minimal Lagrangian geometry

Minimal submanifolds are among the central objects in differential geometry. There is an important subclass of minimal submanifolds which was introduced by Harvey and Lawson [6] in 1982. Given a Riemannian manifold (M, g) , a calibrating form Ω is a closed m -form on M with the property that $\Omega(e_1, \dots, e_m) \leq 1$ for each point $p \in M$ and every orthonormal k -frame $\{e_1, \dots, e_m\} \subset T_p M$. An oriented m -dimensional submanifold $\Sigma \subset M$ is said to be calibrated by Ω if $\Omega(e_1, \dots, e_m) = 1$ for every point $p \in \Sigma$ and every positively oriented orthonormal basis $\{e_1, \dots, e_m\}$ of $T_p \Sigma$. Using Stokes theorem, Harvey and Lawson showed that every calibrated submanifold is necessarily minimal:

THEOREM 1.1 (R. Harvey, H.B. Lawson [6]). *Let (M, g) be a Riemannian manifold. Moreover, let Ω be a calibrating k -form and let Σ be a k -dimensional submanifold calibrated by Ω . Then Σ minimizes volume in its homology class.*

In the following, we consider the special case when (M, g) is the Euclidean space \mathbb{R}^{2n} . We denote by $(x_1, \dots, x_n, y_1, \dots, y_n)$ the standard coordinates on \mathbb{R}^{2n} . Moreover, we denote by $\omega = \sum_{k=1}^n dx_k \wedge dy_k$ the standard symplectic form. Let J be the associated complex structure, so that $J \frac{\partial}{\partial x_k} = \frac{\partial}{\partial y_k}$ and $J \frac{\partial}{\partial y_k} = -\frac{\partial}{\partial x_k}$. Finally, we define

$$\sigma = (dx_1 + i dy_1) \wedge \dots \wedge (dx_n + i dy_n).$$

The author was supported in part by the National Science Foundation under grant DMS-0905628.

Note that σ is a complex-valued n -form on \mathbb{R}^{2n} . Moreover, we have

$$\sigma(Jv_1, v_2, \dots, v_n) = i\sigma(v_1, v_2, \dots, v_n)$$

for all vectors $v_1, \dots, v_n \in \mathbb{R}^{2n}$.

Let now Σ be a submanifold of \mathbb{R}^{2n} of dimension n . Recall that Σ is said to be Lagrangian if $\omega|_{\Sigma} = 0$. If Σ is a Lagrangian submanifold, then it can be shown that $|\sigma(e_1, \dots, e_n)| = 1$, where $\{e_1, \dots, e_n\}$ is an orthonormal basis of $T_p\Sigma$. We may therefore write

$$(1) \quad \sigma(e_1, \dots, e_n) = e^{i\gamma}$$

for some function $\gamma : \Sigma \rightarrow \mathbb{R}/2\pi\mathbb{Z}$. The function γ is referred to as the Lagrangian angle of Σ .

The mean curvature vector of a Lagrangian submanifold Σ is given by $J\nabla^{\Sigma}\gamma$, where $\nabla^{\Sigma}\gamma \in T_p\Sigma$ denotes the gradient of the Lagrangian angle. In particular, this implies:

THEOREM 1.2 (R. Harvey, H.B. Lawson [6]). *If Σ is a Lagrangian submanifold with $H = 0$, then the Lagrangian angle must be constant. Conversely, if Σ is a Lagrangian and the Lagrangian angle is constant (so that $\gamma = c$), then Σ is calibrated by the n -form $\Omega = \operatorname{Re}(e^{-ic}\sigma)$.*

In particular, minimal Lagrangian submanifolds are special cases of calibrated submanifolds.

The first non-trivial examples of minimal Lagrangian submanifolds in \mathbb{R}^{2n} were constructed by Harvey and Lawson [6]. These examples are nearly flat and are constructed by means of the implicit function theorem.

2. Minimal Lagrangian graphs in \mathbb{R}^{2n}

We now assume that Σ is an n -dimensional submanifold of \mathbb{R}^{2n} which can be written as a graph over a Lagrangian plane in \mathbb{R}^{2n} . In other words, we write

$$\Sigma = \{(x_1, \dots, x_n, y_1, \dots, y_n) \in \mathbb{R}^{2n} : (y_1, \dots, y_n) = f(x_1, \dots, x_n)\}.$$

Here, the map f is defined on some domain in \mathbb{R}^n and takes values in \mathbb{R}^n .

The condition that Σ is Lagrangian is equivalent to the condition that $\partial_k f_l = \partial_l f_k$. Thus, Σ is Lagrangian if and only if the map f can locally be written as the gradient of some real-valued function u . In this case, the Lagrangian angle of Σ is given by

$$\gamma = \sum_{k=1}^n \arctan(\lambda_k),$$

where $\lambda_1, \dots, \lambda_k$ denote the eigenvalues of $Df(x) = D^2u(x)$. Therefore, Σ is a minimal Lagrangian submanifold if and only if u satisfies the Hessian equation

$$(2) \quad \sum_{k=1}^n \arctan(\lambda_k) = c.$$

A natural question is to classify all entire solutions of (2). In this direction Tsui and Wang proved the following result:

THEOREM 2.1 (M.P. Tsui, M.T. Wang [15]). *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a smooth map with the property that $\Sigma = \{(x, f(x)) : x \in \mathbb{R}^n\}$ is a minimal Lagrangian graph. Moreover, we assume that, for each point $x \in \mathbb{R}^n$, the eigenvalues of $Df(x)$ satisfy $\lambda_i \lambda_j \geq -1$ and $|\lambda_i| \leq K$. Then f is an affine function.*

A closely related Bernstein-type result was established independently in [23]:

THEOREM 2.2 (Y. Yuan [23]). *Let $u : \mathbb{R}^n \rightarrow \mathbb{R}$ be a smooth convex solution of (2). Then u is a quadratic polynomial.*

In order to study the equation (2) on a bounded domain in \mathbb{R}^n , one needs to impose a boundary condition. One possibility is to impose a Dirichlet boundary condition for the potential function u . This boundary value problem was studied in the fundamental work of Caffarelli, Nirenberg, and Spruck [4]. In particular, they obtained the following existence theorem:

THEOREM 2.3 (L. Caffarelli, L. Nirenberg, J. Spruck [4]). *Let Ω be a uniformly convex domain in \mathbb{R}^n , and let $\varphi : \partial\Omega \rightarrow \mathbb{R}$ be a smooth function. Then there exists a smooth function $u : \Omega \rightarrow \mathbb{R}$ satisfying*

$$\sum_{k=1}^n \arctan(\lambda_k) = \left[\frac{n-1}{2} \right] \pi$$

and $u|_{\partial\Omega} = \varphi$.

We now describe another natural boundary condition for (2). Instead of prescribing the boundary values of u , we prescribe the image of Ω under the map $f = \nabla u$. This choice of boundary condition has been studied before in connection with the Monge-Ampère equation (see [3], [17], [18]).

THEOREM 2.4 (S. Brendle, M. Warren [2]). *Let Ω and $\tilde{\Omega}$ be uniformly convex domains in \mathbb{R}^n . Then we can find a smooth function $u : \Omega \rightarrow \mathbb{R}$ and a real number c with the following properties:*

- (i) *The function u is uniformly convex.*
- (ii) *The function u solves the equation (2).*
- (iii) *The map $\nabla u : \Omega \rightarrow \mathbb{R}$ is a diffeomorphism from Ω to $\tilde{\Omega}$.*

Moreover, the pair (u, c) is unique.

Thus, we can draw the following conclusion:

COROLLARY 2.5 (S. Brendle, M. Warren [2]). *Let Ω and $\tilde{\Omega}$ be uniformly convex domains in \mathbb{R}^n with smooth boundary. Then there exists a diffeomorphism $f : \Omega \rightarrow \tilde{\Omega}$ such that the graph $\Sigma = \{(x, f(x)) : x \in \Omega\}$ is a minimal Lagrangian submanifold of \mathbb{R}^{2n} .*

In particular, the submanifold Σ satisfies $\partial\Sigma \subset \partial\Omega \times \partial\tilde{\Omega}$. Thus, the surface Σ satisfies a free boundary value problem.

We note that the potential function u is not a geometric quantity; on the other hand, the gradient $\nabla u = f$ does have geometric significance. From a geometric point of view, the second boundary value problem is more natural than the Dirichlet boundary condition.

We now describe the proof of Theorem 2.4. The uniqueness statement follows from a standard argument based on the maximum principle. In order to prove the existence statement, we use the continuity method. The idea is to deform Ω and $\tilde{\Omega}$ to the unit ball in \mathbb{R}^n . As usual, the central issue is to bound the Hessian of the potential function u . In geometric terms, this corresponds to a bound on the slope of Σ .

PROPOSITION 2.6 ([2]). *Let us fix two uniformly convex domains Ω and $\tilde{\Omega}$. Moreover, let u be a convex solution of (2) with the property that ∇u is a diffeomorphism from Ω to $\tilde{\Omega}$. Then $|D^2u(x)| \leq C$ for all points $x \in \Omega$ and all vectors $v \in \mathbb{R}^n$. Here, C is a positive constant, which depends only on Ω and $\tilde{\Omega}$.*

The proof of Proposition 2.6 is inspired by earlier work of Urbas on the Monge-Ampère equation. By assumption, we can find uniformly convex boundary defining functions $h : \Omega \rightarrow (-\infty, 0]$ and $\tilde{h} : \tilde{\Omega} \rightarrow (-\infty, 0]$, so that $h|_{\partial\Omega} = 0$ and $\tilde{h}|_{\partial\tilde{\Omega}} = 0$. Moreover, let us fix a constant $\theta > 0$ such that $D^2h(x) \geq \theta I$ for all points $x \in \Omega$ and $D^2\tilde{h}(y) \geq \theta I$ for all points $y \in \tilde{\Omega}$.

In the following, we sketch the main steps involved in the proof of Proposition 2.6.

Step 1: Let u be a convex solution of (2) with the property that ∇u is a diffeomorphism from Ω to $\tilde{\Omega}$. Differentiating the equation (2), we obtain

$$(3) \quad \sum_{i,j=1}^n a_{ij}(x) \partial_i \partial_j \partial_k u(x) = 0$$

for all $x \in \Omega$ and all $k \in \{1, \dots, n\}$. Here, the coefficients $a_{ij}(x)$ are defined as the components of the matrix $A(x) = (I + (D^2u(x))^2)^{-1}$.

We now define a function $H : \Omega \rightarrow \mathbb{R}$ by $H(x) = \tilde{h}(\nabla u(x))$. Using the identity (3), one can show that

$$\left| \sum_{i,j=1}^n a_{ij}(x) \partial_i \partial_j H(x) \right| \leq C$$

for some uniform constant C . Using the maximum principle, we conclude that $H(x) \geq Ch(x)$ for all points $x \in \Omega$. Here, C is a uniform constant which depends only on Ω and $\tilde{\Omega}$. This implies $\langle \nabla h(x), \nabla H(x) \rangle \leq C |\nabla h(x)|^2$ at each point $x \in \partial\Omega$. As a result, we can bound certain components of the Hessian of u along $\partial\Omega$.

Step 2: In the next step, we prove a uniform obliqueness estimate. To that end, we consider the function $\chi(x) = \langle \nabla h(x), \nabla \tilde{h}(\nabla u(x)) \rangle$. It is not difficult to show that $\chi(x) > 0$ for all $x \in \partial\Omega$. The goal is to obtain a uniform lower bound for $\inf_{x \in \partial\Omega} \chi(x)$. Using the relation (3), one can show that

$$\left| \sum_{i,j=1}^n a_{ij}(x) \partial_i \partial_j \chi(x) \right| \leq C$$

for some uniform constant C . We can therefore find a uniform constant K such that

$$\sum_{i,j=1}^n a_{ij}(x) \partial_i \partial_j (\chi(x) - K h(x)) \leq 0.$$

We now consider a point $x_0 \in \partial\Omega$, where the function $\chi(x) - K h(x)$ attains its global minimum. Then $\nabla \chi(x_0) = (K - \mu) \nabla h(x_0)$ for some real number $\mu \geq 0$. Hence, we obtain

$$\begin{aligned} (K - \mu) \chi(x_0) &= \langle \nabla \chi(x_0), \nabla \tilde{h}(\nabla u(x_0)) \rangle \\ &= \sum_{i,j=1}^n \partial_i \partial_j h(x_0) (\partial_i \tilde{h})(\nabla u(x_0)) (\partial_j \tilde{h})(\nabla u(x_0)) \\ &\quad + \sum_{i,j=1}^n (\partial_i \partial_j \tilde{h})(\nabla u(x_0)) \partial_i h(x_0) \partial_j H(x_0) \\ &\geq \theta |\nabla \tilde{h}(\nabla u(x_0))|^2 + \sum_{i,j=1}^n (\partial_i \partial_j \tilde{h})(\nabla u(x_0)) \partial_i h(x_0) \partial_j H(x_0). \end{aligned}$$

Since $\nabla H(x_0)$ is a positive multiple of $\nabla h(x_0)$, it follows that

$$K \chi(x_0) \geq \theta |\nabla \tilde{h}(\nabla u(x_0))|^2.$$

Since $\inf_{x \in \partial\Omega} \chi(x) = \chi(x_0)$, we obtain a uniform lower bound for $\inf_{x \in \partial\Omega} \chi(x)$.

Step 3: Having established the uniform obliqueness estimate, we next bound the tangential components of the Hessian $D^2u(x)$ for each point $x \in \partial\Omega$. To explain this, let

$$M = \sup \left\{ \sum_{k,l=1}^n \partial_k \partial_l u(x) z_k z_l : x \in \partial\Omega, z \in T_x(\partial\Omega), |z| = 1 \right\}.$$

Our goal is to establish an upper bound for M . To that end, we fix a point $x_0 \in \partial M$ and a vector $w \in T_{x_0}(\partial\Omega)$ such that $|w| = 1$ and

$$\sum_{k,l=1}^n \partial_k \partial_l u(x_0) w_k w_l = M.$$

We then consider the function

$$\psi(x) = \sum_{k,l=1}^n \partial_k \partial_l u(x) w_k w_l.$$

Differentiating the identity (2) twice, we obtain

$$\sum_{i,j=1}^n a_{ij}(x) \partial_i \partial_j \psi(x) \geq 0$$

for all $x \in \Omega$. Using the definition of M , it can be shown that

$$(4) \quad \begin{aligned} \psi(x) &\leq M \left| w - \frac{\langle \nabla h(x), w \rangle}{\langle \nabla h(x), \nabla \tilde{h}(\nabla u(x)) \rangle} \nabla \tilde{h}(\nabla u(x)) \right|^2 \\ &\quad + L \langle \nabla h(x), w \rangle^2 \end{aligned}$$

for all points $x \in \partial\Omega$. Here, L is fixed constant that depends only on Ω and $\tilde{\Omega}$.

Let ε be a positive real number such that $\inf_{x \in \partial\Omega} \chi(x) > \varepsilon$, and let $\eta : \mathbb{R} \rightarrow (0, \infty)$ be a smooth function satisfying $\eta(s) = s$ for all $s \geq \varepsilon$. Using (4) and the maximum principle, we obtain an estimate of the form

$$(5) \quad \begin{aligned} \psi(x) &\leq M \left| w - \frac{\langle \nabla h(x), w \rangle}{\eta(\langle \nabla h(x), \nabla \tilde{h}(\nabla u(x)) \rangle)} \nabla \tilde{h}(\nabla u(x)) \right|^2 \\ &\quad + L \langle \nabla h(x), w \rangle^2 - C h(x) \end{aligned}$$

for all $x \in \Omega$. Moreover, equality holds in (5) when $x = x_0$. Consequently, we obtain a lower bound for the normal derivative of ψ at the point x_0 . More precisely,

$$\langle \nabla \psi(x_0), \nabla \tilde{h}(\nabla u(x_0)) \rangle + C M + C \geq 0,$$

where C is a uniform constant that depends only on Ω and $\tilde{\Omega}$. On the other hand, we have

$$\begin{aligned} &\langle \nabla \psi(x_0), \nabla \tilde{h}(\nabla u(x_0)) \rangle + \theta M^2 \\ &\leq \sum_{i,k,l=1}^n (\partial_i \tilde{h})(\nabla u(x_0)) \partial_i \partial_k \partial_l u(x_0) w_k w_l \\ &\quad + \sum_{i,j,k,l=1}^n (\partial_i \partial_j \tilde{h})(\nabla u(x_0)) \partial_i \partial_k u(x_0) \partial_j \partial_l u(x_0) w_k w_l \\ &= \sum_{k,l=1}^n \partial_k \partial_l H(x_0) w_k w_l \\ &= -\langle \nabla H(x_0), II(w, w) \rangle, \end{aligned}$$

where II denotes the second fundamental form of $\partial\Omega$. Consequently,

$$\langle \nabla \psi(x_0), \nabla \tilde{h}(\nabla u(x_0)) \rangle + \theta M^2 \leq C.$$

Putting these facts together, we obtain an a-priori estimate for M .

Step 4: Once we have uniform bounds for the Hessian of u along the boundary, we can use the maximum principle to bound the Hessian of u in the interior of Ω . This step is by now standard, and follows ideas in [4].

3. Area-preserving minimal maps between surfaces

We now describe a different boundary problem value for minimal Lagrangian graphs. To that end, let M be a two-dimensional surface equipped with a Riemannian metric g and a complex structure J . We consider the product $M = N \times N$ equipped with the product metric. We define a complex structure on M by

$$J_{(p,q)}(w, \tilde{w}) = (J_p w, -J_q \tilde{w})$$

for all vectors $w \in T_p N$ and $\tilde{w} \in T_q N$.

Our goal is to construct minimal Lagrangian submanifolds in M . We will assume throughout this section that N is a surface with constant Gaussian curvature, so that M is a Kähler-Einstein manifold. (Otherwise, the minimal Lagrangian equation leads to an overdetermined system of PDEs).

In the special case when $N = \mathbb{R}^2$, the existence of minimal Lagrangian graphs can be reduced to the solvability of the second boundary value problem for the Monge-Ampère equation. To describe this, we consider two domains $\Omega, \tilde{\Omega} \subset \mathbb{R}^2$. Moreover, we consider a diffeomorphism $f : \Omega \rightarrow \tilde{\Omega}$, and let

$$\Sigma = \{(p, f(p)) : p \in \Omega\}.$$

The graph Σ is Lagrangian if and only if the map f is area-preserving and orientation-preserveing, so that $\det Df = 1$. Moreover, Σ has vanishing mean curvature if and only if the Lagrangian angle is constant; this means that

$$\cos \gamma (\partial_1 f_2 - \partial_2 f_1) = \sin \gamma (\partial_1 f_1 + \partial_2 f_2)$$

for some constant $\gamma \in \mathbb{R}$. Hence, we may locally write

$$\begin{aligned} f_1 &= \cos \gamma \partial_1 u - \sin \gamma \partial_2 u \\ f_2 &= \sin \gamma \partial_1 u + \cos \gamma \partial_2 u \end{aligned}$$

for some potential function u .

In other words, the map f can locally be expressed as the composition of a gradient mapping with a rotation in \mathbb{R}^2 . Since f is area-preserving, the potential function solves the Monge-Ampère equation $\det D^2 u = 1$.

It was shown by Delanoë [5] that the second boundary value problem for the Monge-Ampère equation is solvable, provided that Ω and $\tilde{\Omega}$ are uniformly convex and have the same area. This implies the following result:

THEOREM 3.1 (P. Delanoë [5]). *Let Ω and $\tilde{\Omega}$ be uniformly convex domains in \mathbb{R}^2 with smooth boundary. Assume that Ω and $\tilde{\Omega}$ have the same area. Then there exists a minimal Lagrangian diffeomorphism from Ω to $\tilde{\Omega}$.*

The assumption that Ω and $\tilde{\Omega}$ are uniformly convex cannot be removed. In fact, Urbas [19] constructed two domains in \mathbb{R}^2 such that the second

boundary value for the Monge-Ampère equation does not admit a smooth solution. In this example, the domain Ω is the unit disk; moreover, the geodesic curvature of $\partial\tilde{\Omega}$ is greater than $-\varepsilon$.

We next consider the case when N is a complete, simply connected surface with negative Gaussian curvature. In this case, we have the following result:

THEOREM 3.2 (S. Brendle [1]). *Let N be a complete, simply connected surface with constant negative Gaussian curvature, and let Ω and $\tilde{\Omega}$ be uniformly convex domains in N with smooth boundary. Assume that Ω and $\tilde{\Omega}$ have the same area. Given any point $\bar{p} \in \partial\Omega$ and any point $\bar{q} \in \partial\tilde{\Omega}$, there exists a unique minimal Lagrangian diffeomorphism from Ω to $\tilde{\Omega}$ that maps \bar{p} to \bar{q} .*

We note that the product M does not admit a parallel complex volume form. Therefore, we do not have a notion of Lagrangian angle in this setting. As a result, it is no longer possible to reduce the minimal Lagrangian equation to a PDE for a scalar function.

The proof of Theorem 3.2 uses the continuity method. To that end, we consider a continuous family of domains $\Omega_t, \tilde{\Omega}_t \subset N$ with the following properties:

- For each $t \in (0, 1]$, the domains Ω_t and $\tilde{\Omega}_t$ are uniformly convex, and $\text{area}(\Omega_t) = \text{area}(\tilde{\Omega}_t)$.
- $\Omega_1 = \Omega$ and $\tilde{\Omega}_1 = \tilde{\Omega}$.
- If $t \in (0, 1]$ is sufficiently small, then Ω_t and $\tilde{\Omega}_t$ are geodesic disks in N . Moreover, the radius converges to 0 as $t \rightarrow 0$.

In order to construct domains $\Omega_t, \tilde{\Omega}_t \subset N$ with these properties, we consider the sub-level sets of suitable boundary defining functions (see [1] for details). We then consider the following problem:

(\star_t) *Find all area-preserving minimal maps $f : \Omega_t \rightarrow \tilde{\Omega}_t$ that map a given point on the boundary of Ω_t to a given point on the boundary of $\tilde{\Omega}_t$.*

As $t \rightarrow 0$, the domains Ω_t and $\tilde{\Omega}_t$ converge to the unit disk $\mathbb{B}^2 \subset \mathbb{R}^2$ after rescaling. Hence, for $t \rightarrow 0$, the problem (\star_t) reduces to the problem of finding all area-preserving minimal maps from \mathbb{B}^2 to itself. This problem is well understood: in fact, an area-preserving map from \mathbb{B}^2 to itself is minimal if and only if it is a rotation.

In order to make the continuity argument work, it is necessary to establish a-priori estimates for solutions of (\star_t) . The key step is the bound the differential Df .

PROPOSITION 3.3 ([1]). *Let Ω and $\tilde{\Omega}$ be uniformly convex domains in N with smooth boundary. Suppose that $f : \Omega \rightarrow \tilde{\Omega}$ is an area-preserving minimal map. Then $|Df_p| \leq C$ for all points $p \in \Omega$, where C is a uniform constant that depends only on Ω and $\tilde{\Omega}$.*

We now sketch the main ideas involved in the proof of Proposition 3.3. Let $h : \Omega \rightarrow (-\infty, 0]$ and $\tilde{h} : \tilde{\Omega} \rightarrow (-\infty, 0]$ be uniformly convex boundary defining functions for Ω and $\tilde{\Omega}$. We may choose h and \tilde{h} such that $|\nabla h_p| = 1$ for all $p \in \partial\Omega$ and $|\nabla \tilde{h}_q| = 1$ for all $q \in \partial\tilde{\Omega}$.

Since h and \tilde{h} are uniformly convex, we have

$$(6) \quad \theta g \leq D^2 h \leq \frac{1}{\theta} g$$

and

$$(7) \quad \theta g \leq D^2 \tilde{h} \leq \frac{1}{\theta} g$$

for some positive constant θ .

Step 1: Let

$$\Sigma = \{(p, f(p)) : p \in \Omega\}$$

denote the graph of f . By assumption, Σ is a minimal submanifold of M . We next define two functions $H, \tilde{H} : \Sigma \rightarrow \mathbb{R}$ by $H(p, f(p)) = h(p)$ and $\tilde{H}(p, f(p)) = \tilde{h}(f(p))$. The relations (6) and (7) imply $\theta \leq \Delta_{\Sigma} H \leq \frac{1}{\theta}$ and $\theta \leq \Delta_{\Sigma} \tilde{H} \leq \frac{1}{\theta}$. Using the maximum principle, we obtain $\frac{1}{\theta^2} H \leq \tilde{H} \leq \theta^2 H$ at each point on Σ . In other words, we have

$$\frac{1}{\theta^2} h(p) \leq \tilde{h}(f(p)) \leq \theta^2 h(p)$$

for all points $p \in \Omega$. Consequently,

$$\theta^2 \leq \langle Df_p(\nabla h_p), \nabla \tilde{h}_{f(p)} \rangle \leq \frac{1}{\theta^2}$$

for all points $p \in \partial\Omega$.

Step 2: In the next step, we define a linear isometry $Q_p : T_p N \rightarrow T_{f(p)} N$ by

$$Q_p = Df_p [Df_p^* Df_p]^{-\frac{1}{2}}.$$

It is straightforward to verify that $J_{f(p)} Q_p = Q_p J_p$ for all $p \in \Omega$. We next define a bilinear form $\sigma : T_{(p,f(p))} M \times T_{(p,f(p))} M \rightarrow \mathbb{C}$ by

$$\begin{aligned} \sigma((w_1, \tilde{w}_1), (w_2, \tilde{w}_2)) &= i \langle Q_p(w_1), \tilde{w}_2 \rangle + \langle Q_p(J_p w_1), \tilde{w}_2 \rangle \\ &\quad - i \langle Q_p(w_2), \tilde{w}_1 \rangle - \langle Q_p(J_p w_2), \tilde{w}_1 \rangle \end{aligned}$$

for all vectors $w_1, w_2 \in T_p N$ and all vectors $\tilde{w}_1, \tilde{w}_2 \in T_{f(p)} N$. The bilinear form σ satisfies $\sigma(W_2, W_1) = -\sigma(W_1, W_2)$ and $\sigma(JW_1, W_2) = i\sigma(W_1, W_2)$ for all vectors $W_1, W_2 \in T_{(p,f(p))} M$. Moreover, if $\{e_1, e_2\}$ is an orthonormal basis of $T_{(p,f(p))} \Sigma$, then $\sigma(e_1, e_2) = \pm 1$.

The crucial observation is that σ is parallel with respect to the Levi-Civita connection on M . More precisely, suppose that W_1 and W_2 are vector fields on M . Then the expression $\sigma(W_1, W_2)$ defines a complex-valued function on Σ . The derivative of that function is given by

$$(8) \quad V(\sigma(W_1, W_2)) = \sigma(\nabla_V^M W_1, W_2) + \sigma(W_1, \nabla_V^M W_2).$$

The relation (8) is a consequence of the fact that Σ has zero mean curvature (see [1], Proposition 3.3, for details). Differentiating the identity (8), we obtain

$$\begin{aligned}
 \Delta_\Sigma(\sigma(W_1, W_2)) &= \sum_{k=1}^2 \sigma(\nabla_{e_k, e_k}^{M,2} W_1, W_2) \\
 (9) \quad &\quad + \sum_{k=1}^2 \sigma(W_1, \nabla_{e_k, e_k}^{M,2} W_2) \\
 &\quad + 2 \sum_{k=1}^2 \sigma(\nabla_{e_k}^M W_1, \nabla_{e_k}^M W_2).
 \end{aligned}$$

Step 3: We now define a function $\varphi : \Sigma \rightarrow \mathbb{R}$ by

$$\varphi(p, f(p)) = \langle Q_p(\nabla h_p), \nabla \tilde{h}_{f(p)} \rangle.$$

It is easy to see that $\varphi(p, f(p)) > 0$ for $p \in \partial\Omega$. Our goal is to establish a lower bound for $\inf_{p \in \partial\Omega} \varphi(p, f(p))$. This estimate can be viewed as a generalization of the uniform obliqueness estimate in [5].

To prove this estimate, we define vector fields W_1 and W_2 on M by $(W_1)_{(p,q)} = (\nabla h_p, 0)$ and $(W_2)_{(p,q)} = (0, \nabla \tilde{h}_q)$. Clearly, $\varphi = \operatorname{Re}(\sigma(W_1, W_2))$. Hence, the identity (9) implies $\Delta_\Sigma \varphi \leq L$, where L is a positive constant that depends only on Ω and $\tilde{\Omega}$. Hence, we obtain $\Delta_\Sigma(\varphi - \frac{L}{\theta} H) \leq 0$. Consequently, the function $\varphi - \frac{L}{\theta} H$ attains its maximum at some point $(p_0, f(p_0)) \in \partial\Sigma$. At the point $(p_0, f(p_0))$, we have

$$\nabla^\Sigma \varphi = \left(\frac{L}{\theta} - \mu \right) \nabla^\Sigma H$$

for some real number $\mu \geq 0$. Consequently, for every vector $v \in T_{p_0} N$, we have

$$\begin{aligned}
 \left(\frac{L}{\theta} - \mu \right) \langle \nabla h_{p_0}, v \rangle &= \left(\frac{L}{\theta} - \mu \right) \langle \nabla^\Sigma H, (v, Df_{p_0}(v)) \rangle \\
 &= \langle \nabla^\Sigma \varphi, (v, Df_{p_0}(v)) \rangle \\
 &= (D^2 h)_{p_0}(v, Q_{p_0}^*(\nabla \tilde{h}_{f(p_0)})) \\
 &\quad + (D^2 \tilde{h})_{f(p_0)}(Q_{p_0}(\nabla h_{p_0}), Df_{p_0}(v)).
 \end{aligned}$$

In particular, if we choose $v = Q_{p_0}^*(\nabla \tilde{h}_{f(p_0)})$, then we obtain

$$\begin{aligned}
 \left(\frac{L}{\theta} - \mu \right) \varphi(p_0, f(p_0)) &= (D^2 h)_{p_0}(Q_{p_0}^*(\nabla \tilde{h}_{f(p_0)}), Q_{p_0}^*(\nabla \tilde{h}_{f(p_0)})) \\
 &\quad + (D^2 \tilde{h})_{f(p_0)}(Q_{p_0}(\nabla h_{p_0}), Q_{p_0}(Df_{p_0}^*(\nabla \tilde{h}_{f(p_0)}))).
 \end{aligned}$$

By (6), we have

$$\begin{aligned}
 &(D^2 h)_{p_0}(Q_{p_0}^*(\nabla \tilde{h}_{f(p_0)}), Q_{p_0}^*(\nabla \tilde{h}_{f(p_0)})) \\
 &\geq \theta |Q_{p_0}^*(\nabla \tilde{h}_{f(p_0)})|^2 = \theta |\nabla \tilde{h}_{f(p_0)}|^2 = \theta.
 \end{aligned}$$

Moreover, the vector $Df_{p_0}^*(\nabla \tilde{h}_{f(p_0)})$ is a positive multiple of ∇h_{p_0} . Since \tilde{h} is convex, it follows that

$$(D^2\tilde{h})_{f(p_0)}(Q_{p_0}(\nabla h_{p_0}), Q_{p_0}(Df_{p_0}^*(\nabla \tilde{h}_{f(p_0)}))) \geq 0.$$

Putting these facts together yields

$$\left(\frac{L}{\theta} - \mu\right)\varphi(p_0, f(p_0)) \geq \theta,$$

hence

$$(10) \quad \inf_{p \in \partial\Omega} \varphi(p, f(p)) = \varphi(p_0, f(p_0)) \geq \frac{\theta^2}{L}.$$

Step 4: We next show that $|Df_p| \leq C$ for all points $p \in \partial\Omega$. To see this, let us define $v_1 = \nabla h_p$ and $v_2 = J\nabla h_p$. Similarly, we define $\tilde{v}_1 = \nabla \tilde{h}_{f(p)}$ and $\tilde{v}_2 = J\nabla \tilde{h}_{f(p)}$. Clearly, the vectors $\{v_1, v_2\}$ form an orthonormal basis of $T_p N$, and the vectors $\{\tilde{v}_1, \tilde{v}_2\}$ form an orthonormal basis of $T_{f(p)} N$. We now write

$$Df_p(v_1) = a\tilde{v}_1 + b\tilde{v}_2$$

and

$$Df_p(v_2) = c\tilde{v}_2$$

for suitable coefficients a, b, c . Note that $ac = 1$ since f is area-preserving. Using the inequality $\theta^2 \leq \langle Df_p(\nabla h_p), \nabla \tilde{h}_{f(p)} \rangle \leq \frac{1}{\theta}^2$, we conclude that $\theta^2 \leq a \leq \frac{1}{\theta^2}$ and $\theta^2 \leq c \leq \frac{1}{\theta^2}$. In order to bound b , we observe that

$$\begin{aligned} a\langle Q_p(v_2), \tilde{v}_1 \rangle + b\langle Q_p(v_2), \tilde{v}_2 \rangle &= \langle Q_p(v_2), Df_p(v_1) \rangle \\ &= \langle Q_p(v_1), Df_p(v_2) \rangle \\ &= c\langle Q_p(v_1), \tilde{v}_2 \rangle. \end{aligned}$$

Moreover, we have

$$\langle Q_p(v_2), \tilde{v}_2 \rangle = \langle Q_p(v_1), \tilde{v}_1 \rangle = \varphi(p, f(p)) \geq \frac{\theta^2}{L}$$

by (10). Putting these facts together, we conclude that $|b| \leq C$ for some uniform constant C .

Step 5: In the last step, we show that $|Df_p| \leq C$ for all points $p \in \Omega$. To that end, we define a function $\beta : \Sigma \rightarrow \mathbb{R}$ by

$$\beta(p, f(p)) = \frac{2}{\sqrt{\det(I + Df_p^* Df_p)}}.$$

The function β satisfies an inequality of the form

$$(11) \quad \Delta_\Sigma \beta + \kappa \beta (1 - \beta^2) \leq 0.$$

Here, $\kappa < 0$ denotes the Gaussian curvature of the two-dimensional surface N . Moreover, the restriction $\beta|_{\partial\Sigma}$ is uniformly bounded from below. Using (11) and the maximum principle, one obtains a uniform lower bound for $\inf_{p \in \Omega} \beta(p, f(p))$.

The inequality (11) was first discovered by Wang [21] in his study of the Lagrangian mean curvature flow. In the remainder of this section, we shall sketch the proof of (11). Given any point $(p, q) \in M$, we define a two-form $\rho : T_{(p,q)}M \times T_{(p,q)}M \rightarrow \mathbb{R}$ by

$$\rho((w_1, \tilde{w}_1), (w_2, \tilde{w}_2)) = \langle Jw_1, w_2 \rangle + \langle J\tilde{w}_1, \tilde{w}_2 \rangle$$

for all vectors $w_1, w_2 \in T_p N$ and $\tilde{w}_1, \tilde{w}_2 \in T_q N$. Clearly, ρ is parallel. Moreover, we may write $\beta = \rho(e_1, e_2)$, where $\{e_1, e_2\}$ is a local orthonormal frame for $T\Sigma$. Differentiating this identity, we obtain

$$V(\beta) = \rho(\Pi(e_1, V), e_2) + \rho(e_1, \Pi(e_2, V))$$

for every vector $V \in T\Sigma$. This implies

$$(12) \quad \begin{aligned} \Delta_\Sigma \beta &= \sum_{k=1}^2 \rho(\nabla_{e_k}^M \Pi(e_1, e_k), e_2) + \sum_{k=1}^2 \rho(e_1, \nabla_{e_k}^M \Pi(e_2, e_k)) \\ &\quad + 2 \sum_{k=1}^2 \rho(\Pi(e_1, e_k), \Pi(e_2, e_k)). \end{aligned}$$

Using the Codazzi equations (see e.g. [9], Chapter 4, Proposition 33) we obtain

$$(13) \quad \begin{aligned} &\sum_{k=1}^2 \rho(\nabla_{e_k}^M \Pi(e_1, e_k), e_2) + \sum_{k=1}^2 \rho(e_1, \nabla_{e_k}^M \Pi(e_2, e_k)) \\ &= \sum_{k=1}^2 \rho(\nabla_{e_k}^\perp \Pi(e_1, e_k), e_2) + \sum_{k=1}^2 \rho(e_1, \nabla_{e_k}^\perp \Pi(e_2, e_k)) \\ &\quad + \sum_{k=1}^2 \langle \nabla_{e_k}^M \Pi(e_1, e_k), e_1 \rangle \rho(e_1, e_2) + \sum_{k=1}^2 \langle \nabla_{e_k}^M \Pi(e_2, e_k), e_2 \rangle \rho(e_1, e_2) \\ &= \sum_{k=1}^2 \rho(\nabla_{e_1}^\perp \Pi(e_k, e_k), e_2) + \sum_{k=1}^2 \rho(e_1, \nabla_{e_2}^\perp \Pi(e_k, e_k)) \\ &\quad - \sum_{k=1}^2 |\Pi(e_1, e_k)|^2 \rho(e_1, e_2) - \sum_{k=1}^2 |\Pi(e_2, e_k)|^2 \rho(e_1, e_2) \\ &\quad - R_M(e_2, e_1, e_2, Je_1) \rho(Je_1, e_2) - R_M(e_2, e_1, e_2, Je_2) \rho(Je_2, e_2) \\ &\quad - R_M(e_1, e_2, e_1, Je_1) \rho(e_1, Je_1) - R_M(e_1, e_2, e_1, Je_2) \rho(e_1, Je_2). \end{aligned}$$

Here, ∇^\perp denotes the induced connection on the normal bundle of Σ . Since N has constant Gaussian curvature κ , we have

$$\begin{aligned} &R_M(e_2, e_1, e_2, Je_1) \rho(Je_1, e_2) + R_M(e_2, e_1, e_2, Je_2) \rho(Je_2, e_2) \\ &+ R_M(e_1, e_2, e_1, Je_1) \rho(e_1, Je_1) + R_M(e_1, e_2, e_1, Je_2) \rho(e_1, Je_2) \\ &= \kappa \beta (1 - \beta^2). \end{aligned}$$

Substituting this into (13) gives

$$(14) \quad \begin{aligned} & \sum_{k=1}^2 \rho(\nabla_{e_k}^M II(e_1, e_k), e_2) + \sum_{k=1}^2 \rho(e_1, \nabla_{e_k}^M II(e_2, e_k)) \\ & = -|II|^2 \beta - \kappa \beta (1 - \beta^2). \end{aligned}$$

Moreover, we have

$$(15) \quad \begin{aligned} & \sum_{k=1}^2 \rho(II(e_1, e_k), II(e_2, e_k)) \\ & = \sum_{k=1}^2 \langle II(e_1, e_k), Je_1 \rangle \langle II(e_2, e_k), Je_2 \rangle \rho(Je_1, Je_2) \\ & + \sum_{k=1}^2 \langle II(e_1, e_k), Je_2 \rangle \langle II(e_2, e_k), Je_1 \rangle \rho(Je_2, Je_1) \\ & = \sum_{k=1}^2 \langle II(e_1, e_1), Je_k \rangle \langle II(e_2, e_2), Je_k \rangle \beta \\ & - \sum_{k=1}^2 \langle II(e_1, e_2), Je_k \rangle \langle II(e_1, e_2), Je_k \rangle \beta \\ & = -\frac{1}{2} |II|^2 \beta. \end{aligned}$$

Combining (12), (14), and (15), we obtain

$$(16) \quad \Delta_\Sigma \beta = -2 |II|^2 \beta - \kappa \beta (1 - \beta^2).$$

From this, the inequality (11) follows.

4. The Lagrangian mean curvature flow

In this final section, we briefly discuss the flow approach to special Lagrangian geometry. To that end, we consider a Lagrangian submanifold of a Kähler manifold (M, g) , and evolve it by the mean curvature flow. It was shown by Smoczyk that a Lagrangian submanifold of a Kähler-Einstein manifold remains Lagrangian when evolved by the mean curvature flow:

THEOREM 4.1 (K. Smoczyk [11],[12]). *Let (M, g) be a Kähler-Einstein manifold, and let $\{\Sigma_t : t \in [0, T]\}$ be a family of closed submanifolds of (M, g) which evolve by the mean curvature flow. If Σ_0 is Lagrangian, then Σ_t is Lagrangian for all $t \in [0, T]$.*

It is a very interesting question to study the longtime behavior of the Lagrangian mean curvature flow. Thomas and Yau [14] conjectured that the flow exists for all time provided that the initial surface Σ_0 satisfies a certain stability condition. Examples of finite-time singularities were recently constructed by Neves [8].

In the following, we discuss some results about Lagrangian graphs evolving by mean curvature flow. The case of graphs is much better understood than the general case, and some strong results are known in this setting. Let us first consider the torus $\mathbb{T}^{2n} = \mathbb{R}^{2n}/\mathbb{Z}^{2n}$. We assume that \mathbb{R}^{2n} is equipped with its standard metric and complex structure, so that $J \frac{\partial}{\partial x_k} = \frac{\partial}{\partial y_k}$ and $J \frac{\partial}{\partial y_k} = -\frac{\partial}{\partial x_k}$. The torus \mathbb{T}^{2n} inherits a metric and complex structure in the standard way. We then consider submanifolds of the form

$$\Sigma = \{(p, f(p)) : p \in \mathbb{T}^n\},$$

where f is a smooth map from \mathbb{T}^n to itself. The submanifold Σ is Lagrangian if and only if the map f can locally be written in the form $f = \nabla u$ for some potential function u . Smoczyk and Wang were able to analyze the longtime behavior of the mean curvature flow in the special case when the potential function u is convex.

THEOREM 4.2 (K. Smoczyk, M.T. Wang [13]). *Let Σ_0 be a Lagrangian submanifold of \mathbb{T}^{2n} which can be written as the graph of a map $f_0 : \mathbb{T}^n \rightarrow \mathbb{T}^n$. Moreover, suppose that the eigenvalues of $(Df_0)_p$ are strictly positive for each point $p \in \mathbb{T}^n$. Finally, let $\{\Sigma_t : t \in [0, T)\}$ denote the unique maximal solution of the mean curvature flow with initial surface Σ_0 . Then $T = \infty$, and the surfaces Σ_t converge to a totally geodesic Lagrangian submanifold as $t \rightarrow \infty$.*

We next consider the Lagrangian mean curvature flow in a product manifold.

THEOREM 4.3 (M.T. Wang [21]). *Let N and \tilde{N} be compact Riemann surfaces with the same constant curvature c . Moreover, suppose that $f_0 : N \rightarrow \tilde{N}$ is an area-preserving diffeomorphism, and let*

$$\Sigma_0 = \{(p, f_0(p)) : p \in N\} \subset N \times \tilde{N}$$

denote the graph of f_0 . Finally, let $\{\Sigma_t : t \in [0, T)\}$ be the unique maximal solution of the mean curvature flow with initial surface Σ_0 . Then $T = \infty$, and each surface Σ_t is the graph of an area-preserving diffeomorphism $f_t : N \rightarrow \tilde{N}$. Finally, the maps f_t converge smoothly to an area-preserving minimal map as $t \rightarrow \infty$.

The same result was proved independently by Smoczyk [12] under an extra condition on the Lagrangian angle.

Theorem 4.3 gives a new proof of the existence of minimal maps between Riemann surfaces; the existence of such maps was established earlier by Schoen [10] using harmonic map techniques. A stronger result holds when $N = \tilde{N} = S^2$:

THEOREM 4.4 (M.T. Wang [21]). *Let f_0 be an area-preserving diffeomorphism from S^2 to itself, and let*

$$\Sigma_0 = \{(p, f_0(p)) : p \in S^2\} \subset S^2 \times S^2$$

denote the graph of f_0 . Moreover, let $\{\Sigma_t : t \in [0, T)\}$ be the unique maximal solution of the mean curvature flow with initial surface Σ_0 . Then $T = \infty$, and each surface Σ_t is the graph of an area-preserving diffeomorphism $f_t : S^2 \rightarrow S^2$. Finally, the maps f_t converge to an isometry of S^2 as $t \rightarrow \infty$.

The proofs of Theorems 4.2 – 4.4 rely on maximum principle arguments. These techniques also have important applications to the study of area-decreasing maps between spheres (cf. [16], [20]). A detailed discussion of the Lagrangian mean curvature flow can be found in [22].

In a remarkable paper, Medoš and Wang [7] generalized this result to higher dimensions. In higher dimensions, it is necessary to impose a pinching condition on the initial map f_0 :

THEOREM 4.5 (I. Medoš, M.T. Wang [7]). *Given any positive integer n , there exists a real number $\Lambda(n) > 1$ such that the following holds: Let $f_0 : \mathbb{CP}^n \rightarrow \mathbb{CP}^n$ be a symplectomorphism satisfying*

$$\frac{1}{\Lambda(n)} |v| \leq |Df_p(v)| \leq \Lambda(n) |v|$$

for all vectors $v \in T_p \mathbb{CP}^n$. Moreover, let

$$\Sigma_0 = \{(p, f(p)) : p \in \mathbb{CP}^n\} \subset \mathbb{CP}^n \times \mathbb{CP}^n$$

denote the graph of f_0 , and let $\{\Sigma_t : t \in [0, T)\}$ be the unique maximal solution of the mean curvature flow with initial surface Σ_0 . Then $T = \infty$, and each surface Σ_t is the graph of a symplectomorphism $f_t : \mathbb{CP}^n \rightarrow \mathbb{CP}^n$. Moreover, the maps f_t converge smoothly to a biholomorphic isometry of \mathbb{CP}^n as $t \rightarrow \infty$.

In the remainder of this section, we sketch the main ingredients involved in the proof of Theorem 4.5 (see [7] for details). For each $t \geq 0$, one defines a function $\beta_t : \Sigma_t \rightarrow \mathbb{R}$ by

$$\beta_t = \prod_{k=1}^{2n} \frac{1}{\sqrt{1 + \lambda_k^2}},$$

where $\lambda_1, \dots, \lambda_n$ denote the singular values of Df_t . Since f_t is a symplectomorphism, the singular values of Df_t occur in pairs of reciprocal numbers. We may therefore assume that $\lambda_i \lambda_{\tilde{i}} = 1$, where $\tilde{i} = i + (-1)^{i-1}$. Consequently, $\beta_t \leq 2^{-n}$, and equality holds if and only if $\lambda_1 = \dots = \lambda_n = 1$.

The function β_t satisfies an evolution equation of the form

$$\begin{aligned} \frac{\partial}{\partial t} \beta_t &= \Delta_{\Sigma_t} \beta_t + \frac{\beta_t}{2} \sum_{k=1}^{2n} \left(\frac{1 - \lambda_k^2}{1 + \lambda_k^2} \right)^2 \\ &\quad + \beta_t \sum_{i,j,k=1}^{2n} h_{ijk}^2 - 2\beta_t \sum_{k=1}^{2n} \sum_{i < j} (-1)^{i+j} \lambda_i \lambda_j (h_{i\tilde{i}k} h_{j\tilde{j}k} - h_{i\tilde{j}k} h_{j\tilde{i}k}) \end{aligned}$$

where $h_{ijk} = \langle \Pi(e_i, e_j), Je_k \rangle$ denote the components of the second fundamental form of Σ_t (cf. [7], Proposition 2). It is shown in [7] that

$$(17) \quad \sum_{i,j,k=1}^{2n} h_{ijk}^2 - 2 \sum_{k=1}^{2n} \sum_{i < j} (-1)^{i+j} \lambda_i \lambda_j (h_{i\tilde{k}} h_{j\tilde{j}k} - h_{i\tilde{j}k} h_{j\tilde{i}k}) \geq \delta \sum_{i,j,k=1}^{2n} h_{ijk}^2,$$

provided that the singular values $\lambda_1, \dots, \lambda_n$ are sufficiently close to 1. In order to verify this, Medoš and Wang consider the quadratic form

$$\mathcal{Q}(h) = \sum_{i,j,k=1}^{2n} h_{ijk}^2 - 2 \sum_{k=1}^{2n} \sum_{i < j} (-1)^{i+j} (h_{i\tilde{k}} h_{j\tilde{j}k} - h_{i\tilde{j}k} h_{j\tilde{i}k}).$$

The estimate (17) is then a consequence of the following result (cf. [7], Lemma 4):

PROPOSITION 4.6. *The quadratic form $\mathcal{Q}(h)$ satisfies*

$$(18) \quad \mathcal{Q}(h) \geq \frac{2}{9} \sum_{i,j,k=1}^{2n} h_{ijk}^2.$$

In order to prove the inequality (18), we observe that $\sum_{i=1}^{2n} (-1)^i h_{i\tilde{k}} = 0$ for each k . From this, we deduce that $\sum_{i,j=1}^{2n} (-1)^{i+j} h_{i\tilde{k}} h_{j\tilde{j}k} = 0$ for each k . Consequently, the quadratic form $\mathcal{Q}(h)$ can be rewritten as

$$\begin{aligned} \mathcal{Q}(h) &= \sum_{i,j,k=1}^{2n} h_{ijk}^2 - \sum_{i,j,k=1}^{2n} (-1)^{i+j} (h_{i\tilde{k}} h_{j\tilde{j}k} - h_{i\tilde{j}k} h_{j\tilde{i}k}) \\ &= \sum_{i,j,k=1}^{2n} h_{ijk}^2 + \sum_{i,j,k=1}^{2n} (-1)^{i+j} h_{i\tilde{j}k} h_{j\tilde{i}k} \\ &= \frac{1}{2} \sum_{i,j,k=1}^{2n} ((-1)^i h_{i\tilde{j}k} + (-1)^j h_{\tilde{i}jk})^2. \end{aligned}$$

On the other hand, the identity

$$\begin{aligned} 2h_{ijk} &= (-1)^i ((-1)^i h_{ijk} + (-1)^{\tilde{j}} h_{i\tilde{j}k}) \\ &\quad + (-1)^i ((-1)^i h_{ijk} + (-1)^{\tilde{k}} h_{ij\tilde{k}}) \\ &\quad + (-1)^{i+j+k} ((-1)^k h_{i\tilde{j}k} + (-1)^j h_{\tilde{i}jk}) \end{aligned}$$

implies

$$\begin{aligned} 4h_{ijk}^2 &\leq 3 ((-1)^i h_{ijk} + (-1)^{\tilde{j}} h_{i\tilde{j}k})^2 \\ &\quad + 3 ((-1)^i h_{ijk} + (-1)^{\tilde{k}} h_{ij\tilde{k}})^2 \\ &\quad + 3 ((-1)^k h_{i\tilde{j}k} + (-1)^j h_{\tilde{i}jk})^2. \end{aligned}$$

Summation over i, j, k yields

$$4 \sum_{i,j,k=1}^{2n} h_{ijk}^2 \leq 18 \mathcal{Q}(h),$$

as claimed.

References

- [1] S. Brendle, *Minimal Lagrangian diffeomorphisms between domains in the hyperbolic plane*, J. Diff. Geom. 80, 1–22 (2008)
- [2] S. Brendle and M. Warren, *A boundary value problem for minimal Lagrangian graphs*, J. Diff. Geom. 84, 267–287 (2010)
- [3] L. Caffarelli, *Boundary regularity of maps with convex potentials, II*, Ann. of Math. 144, 453–496 (1996)
- [4] L. Caffarelli, L. Nirenberg, and J. Spruck, *The Dirichlet problem for nonlinear second order elliptic equations, III: functions of the eigenvalues of the Hessian*, Acta Math. 155, 261–301 (1985)
- [5] P. Delanoë, *Classical solvability in dimension two of the second boundary-value problem associated with the Monge-Ampère operator*, Ann. Inst. H. Poincaré 8, 443–457 (1991)
- [6] R. Harvey and H.B. Lawson, Jr., *Calibrated geometries*, Acta Math. 148, 47–157 (1982)
- [7] I. Medoš and M.T. Wang, *Deforming symplectomorphisms of complex projective spaces by the mean curvature flow*, J. Diff. Geom. 87, 309–342 (2011)
- [8] A. Neves, *Finite time singularities for Lagrangian mean curvature flow*, arxiv:1009.1083
- [9] B. O’Neill, *Semi-Riemannian geometry*, Academic Press, New York (1983)
- [10] R. Schoen, *The role of harmonic mappings in rigidity and deformation problems*, Complex geometry, Proc. Osaka International Conference, Marcel Dekker, New York, 1993
- [11] K. Smoczyk, *Der Lagrangesche mittlere Krümmungsfluß*, Habilitationsschrift, Leipzig University (1999)
- [12] K. Smoczyk, *Angle theorems for the Lagrangian mean curvature flow*, Math. Z. 240, 849–883 (2002)
- [13] K. Smoczyk and M.T. Wang, *Mean curvature flows of Lagrangian submanifolds with convex potentials*, J. Diff. Geom. 62, 243–257 (2002)
- [14] R.P. Thomas and S.T. Yau, *Special Lagrangians, stable bundles, and mean curvature flow*, Comm. Anal. Geom. 10, 1075–1113 (2002)
- [15] M.P. Tsui and M.T. Wang, *A Bernstein type result for special Lagrangian submanifolds*, Math. Res. Lett. 9 529–535 (2002)
- [16] M.P. Tsui and M.T. Wang, *Mean curvature flows and isotopy of maps between spheres*, Comm. Pure Appl. Math. 57, 1110–1126 (2004)
- [17] J. Urbas, *On the second boundary value problem for equations of Monge-Ampère type*, J. Reine Angew. Math. 487, 115–124 (1997)
- [18] J. Urbas, *The second boundary value problem for a class of Hessian equations*, Comm. PDE 26, 859–882 (2001)
- [19] J. Urbas, *A remark on minimal Lagrangian diffeomorphisms and the Monge-Ampère equation*, Bull. Austral. Math. Soc. 76, 215–218 (2007)
- [20] M.T. Wang, *Mean curvature flow of surfaces in Einstein four-manifolds*, J. Diff. Geom. 57, 301–338 (2001)
- [21] M.T. Wang, *Deforming area-preserving diffeomorphisms of surfaces by mean curvature flow*, Math. Res. Letters 8, 651–662 (2001)

- [22] M.T. Wang, *Some recent developments in Lagrangian mean curvature flows*, Surveys in Differential Geometry vol. XII, pp. 333–347, International Press, Somerville MA (2008)
- [23] Y. Yuan, *A Bernstein problem for special Lagrangian equations*, Invent. Math. 150, 117–125 (2002)

DEPARTMENT OF MATHEMATICS, STANFORD UNIVERSITY, 450 SERRA MALL,
BLDG. 380, STANFORD, CA 94305, USA

Classification of complete $\mathcal{N} = 2$ supersymmetric theories in 4 dimensions

Sergio Cecotti and Cumrun Vafa

ABSTRACT. We define the notion of a complete $\mathcal{N} = 2$ supersymmetric theory in 4 dimensions as one which has a maximal allowed dimension for a UV complete moduli space for the coupling constants, masses and Coulomb branch parameters. We classify all such theories whose BPS spectrum can be obtained via a quiver diagram. This is done using the 4d/2d correspondence and by showing that such complete $\mathcal{N} = 2$ theories map to quivers of finite mutation type. The list of such theories is given by the Gaiotto theories consisting of two 5-branes wrapping Riemann surfaces with punctures, as well as 11 additional exceptional cases, which we identify.

CONTENTS

1. Introduction	19
2. BPS quivers	22
3. Definition of complete $\mathcal{N} = 2$ theories	31
4. 4d-2d correspondence reviewed	32
5. Complete $\mathcal{N} = 2$ theories and quivers of finite mutation type	34
6. Identification of a large class of quivers of finite mutation type as generalized Gaiotto theories	45
7. Identification of the exceptional theories	78
8. Conformal, complete theories	86
9. Physical properties of gauging $\mathcal{N} = 2$ D -sub-systems	88
10. Conclusions	92
Appendix A. Strong coupling spectra of affine quiver models	92
Appendix B. Details on some Landau–Ginzburg models	93
References	99

1. Introduction

Supersymmetric gauge theories with high enough number of supersymmetries are relatively rigid. For example $\mathcal{N} = 4$ supersymmetric theories

in 4 dimensions are completely classified by the choice of the gauge group. However, the ones with lower number of supersymmetries are more flexible. In particular $\mathcal{N} = 1$ theories in 4 dimensions are far from being classified. An interesting intermediate case in four dimensions arises for $\mathcal{N} = 2$ theories, which are in some ways partially rigid, but still not rigid enough to be trivially classified. A large class of these theories are constructed as gauge theories with matter field representations, consistent with asymptotic freedom. On the other hand it is known that there are additional $\mathcal{N} = 2$ theories, that can be obtained from string theory, but which are not easily obtained from gauge theories. These include $\mathcal{N} = 2$ theories with exceptional symmetry groups obtained from 3-brane probes of F-theory, as well as ones which arise from singularities of Calabi-Yau compactifications of type II strings. It is thus natural to ask to what extent we can classify all UV complete $\mathcal{N} = 2$ theories in 4 dimensions.

A similar question arises in 2 dimensional theories with $\mathcal{N} = 2$ supersymmetry. In that case a program for their classification was initiated in [1] based on their BPS soliton/kink spectra. For example it was shown that a theory with two vacua can have only 1 or 2 solitons connecting the two, and the two theories were identified with cubic LG theories and \mathbb{CP}^1 sigma models respectively. The data of the 2d kinks are universal, except that as one changes the parameters of the theory, there could be jumps in the number of BPS states, which are easily computable. This computable change of data of the BPS kinks in 2d will be called a ‘mutation’. Four dimensional theories with $\mathcal{N} = 2$ also have an interesting set of BPS states, which in a sense characterize the theory. Moreover for typical such theories, there is an associated supersymmetric quantum mechanical quiver (with 4 supercharges), whose ground states correspond to such BPS states. It was proposed in [2] that the classification problem for $\mathcal{N} = 2$ theories in 2d and 4d are linked. The basic idea is that $\mathcal{N} = 2$ theories in 4d can be engineered in terms of type II string theories. And the type II theories will have an associated 2d worldsheet theory with $\mathcal{N} = 2$ supersymmetry, which has, in addition to 2d Liouville field, a massive $\mathcal{N} = 2$ theory (for fixed value of Liouville field) with central charge $\hat{c} \leq 2$. Moreover the BPS quiver of the 4d theory was mapped to the vacua and soliton data of the 2d theory. In particular the nodes of the 4d BPS quiver were mapped to vacua of the 2d theory, and the bifundamentals of the quiver, were mapped to solitons connecting the pairs of vacua. Moreover the mutation of the 2d quiver gets mapped to the analogs of Seiberg-like dualities for the supersymmetrical quantum mechanics which gives the number of solitons in different chambers of the 4d theory. Even though this 4d/2d correspondence was not proven in general, it was checked in a number of non-trivial cases and in this paper we continue to assume this holds generally and use it to classify 4d theories with $\mathcal{N} = 2$ supersymmetry.

Classification of 2d theories with $\mathcal{N} = 2$ supersymmetry with $\hat{c} \leq 2$ is already very non-trivial. However, we can refine our classification, by

asking if a natural subclass can be defined from the 4d point of view that can be effectively classified using this correspondence. In this paper we find that there is one natural condition from 4d perspective that can be defined and be classified in this way: We define the notion of ‘complete’ $\mathcal{N} = 2$ supersymmetric theories, as those whose Coulomb branch allows arbitrary deformations compatible with its symmetries. If we have a $U(1)^r$ gauge symmetry at a generic point on Coulomb branch, and a rank f flavor symmetry, the BPS lattice is $2r + f$ dimensional, corresponding to (electric, magnetic, flavor) charges. The maximal allowed deformation we would imagine in this case is $2r + f$ complex dimensional, corresponding to arbitrary local variations of the central charges of the BPS lattice. This could come from r Coulomb branch parameters, f masses, and r coupling constants of the $U(1)^r$ theory. Note, however, that this is not always possible. For example for an $SU(r+1)$ gauge group, we have r Coulomb branch parameters, but only 1 coupling constant, and not r independent ones. On the other hand, the product of $SU(2)$ theories with asymptotically free matter representation is ‘complete’ in the above sense, because each $SU(2)$ can have its own coupling constant. We will argue that this criteria for ‘completeness’ maps to 2d theories with $\hat{c} \leq 1$. Moreover, the corresponding BPS quivers have a finite number of elements in the mutation orbit. In other words, they are of *finite mutation type*. Since the quivers of finite mutation types have been classified mathematically [3–5], we can identify the corresponding theories.

The quivers of finite mutation type turn out to come in two types: They are either associated to a Riemann surface with punctures (with extra data at the punctures), or they belong to one of the 11 exceptional cases. The ones associated to Riemann surfaces get mapped to (generalized) Gaiotto theories with two five branes wrapping the corresponding Riemann surfaces. The 2d version of them correspond to Landau-Ginzburg theories whose fields live on Riemmann surface, with a superpotential with specified poles. Nine of the eleven exceptional cases correspond to type IIB on certain local Calabi-Yau singularities (three of them can also be viewed as an M5 brane wrapping a specific singular curve). These again map to 2d Landau-Ginzburg theories with $\hat{c} = 1$ and their deformations, as well as the exceptional minimal $\mathcal{N} = 2$ LG models. The last two correspond to a massive deformation of the genus 2 Gaiotto theory without punctures, and a certain limit of it. The 2d version of these last two theories is not known. It is remarkable that all complete $\mathcal{N} = 2$ gauge theories that admit a quiver realization for their BPS states are classifiable, and even more surprisingly identifiable! This gives further motivation for an even more complete classification of $\mathcal{N} = 2$ theories by relaxing the completeness criteria.

The organization of the remainder of this paper is as follows: In section 2 we discuss the general notion of quivers relevant for finding the BPS states of 4d, $\mathcal{N} = 2$ theories. In section 3 we give a definition of complete $\mathcal{N} = 2$ theories. In section 4 we review the 4d/2d correspondence advanced in [2].

In section 5 we discuss why the complete $\mathcal{N} = 2$ theories map to quivers of finite mutation type and review the mathematical classification of quivers of finite mutation type. In section 6 we identify the class corresponding to Riemann surfaces with punctures. In section 7 we identify the exceptional ones. In section 8 we identify the conformal subset. In section 9 we discuss some physical properties of the $\mathcal{N} = 2$ models corresponding to the *ADE* affine quivers. Finally in section 10 we present our conclusions. Appendices A and B deal with certain technical computations.

2. BPS quivers

Quivers have been studied in the context of supersymmetric gauge theories in two different ways. In one context one uses them to describe gauge theories with products of $U(N_i)$, one factor group per node, with bifundamental matter being captured by links between nodes. In another approach, one uses quiver to describe BPS states of supersymmetric gauge theories. In this context [6, 7] one is considering a supersymmetric quantum mechanics, again with the $U(N_i)$ gauge groups at the nodes and bifundamental matter. In this latter sense, each node corresponds to an elementary BPS state and one considers all possible ranks N_i for the gauge groups. Then normalizable zero modes for the quantum mechanics signify BPS bound states with the quantum numbers of N_i copies of each elementary state. It is this second sense of quivers that would be of interest in the present paper. We shall call the quivers interpreted in this sense the *BPS quivers*.

Let us give examples of BPS quivers. Consider for example type IIA in the presence of A_{n-1} singularity. We model this by $\mathbb{C}^2/\mathbb{Z}_n$. As it is well known [6], if we consider BPS states for this geometry we end up with the A_{n-1} quiver, corresponding to a supersymmetric quantum mechanical problem with 8 supercharges. The bound states of this theory correspond to the roots of $SU(n)$. These are the BPS states which complete the $U(1)^{n-1}$ vector bosons to an $SU(n)$ vector multiplet. These BPS states correspond to D2 branes wrapped over the 2-cycles of this geometry. Other examples, more relevant for this paper, are the local Calabi-Yau threefolds. For example consider type IIA in the geometry of $\mathbb{C}^3/\mathbb{Z}_3$. Then the corresponding BPS states are given by the quiver consisting of 3 nodes with three directed arrows (see Fig.(2.1)):



This theory corresponds to a supersymmetric quantum mechanical problem with 4 supercharges (the same number as $\mathcal{N} = 1$ in 4d) which captures the BPS states of the $\mathcal{N} = 2$ theory in 4d. The presence of three nodes reflects the fact that this theory can have bound states of D0, D2 and D4 branes, and for each of them there is only one allowed topological class. Each

node corresponds to a linear combination of these three charges. Note that, for generic ranks at each node, the number of incoming and outgoing arrows at each node are not equal. Of course this is not a problem for the quantum mechanical system (unlike the 4d case, where the same quiver would lead to an anomalous gauge theory unless the rank of the three nodes are the same). In addition to the quiver, this theory also has a superpotential. In principle for each closed loop we can introduce a term in the superpotential, and this theory indeed does have a superpotential of the form

$$W = \epsilon_{ijk} \epsilon^{IJK} \text{Tr}(A_I^i A_J^j A_K^k)$$

Where the A_I^i label the 3×3 bifundamental matter. In addition the supersymmetric ground states of the quantum mechanics depend on the choice of the FI parameters for each node, which depend on the choice of moduli. Moreover as we change the moduli sometimes the BPS quiver undergoes Seiberg-like dualities, known as mutations. In this way, one of the nodes is replaced by a dual node (corresponding to reversing the charge of that node), reversing the direction of the arrows to that node, replacing the corresponding bifundamentals from node i , $q_i, \tilde{q}_i \rightarrow Q_i, \tilde{Q}_i$, and adding to the new dual theory all the meson fields which pass through the node M_{ij} . In addition one needs to add, a term to the superpotential given by

$$\delta W = Q_i M_{ij} \tilde{Q}_j.$$

The ground states of the new quiver may be different from that of the old one, related to it by a suitable wall-crossing formula, as in [8–12].

There is another general fact which follows from the geometry of the D-branes. As we noted, each node of the quiver corresponds to a BPS state, which one can imagine as a brane wrapped over a cycle. If we have two nodes, corresponding to two different BPS states, clearly there will be bifundamental strings at the intersections of the branes. Thus we expect the net number of bifundamentals between two nodes to be given by the inner product of the corresponding classes.

So far we have given examples of simple quivers which arise from orbifolding. However it is known that many other $\mathcal{N} = 2$ theories in 4d also have a BPS quiver. For example it is known that the BPS quiver for the pure $SU(2)$ gauge theory is given by the affine Dynkin diagram \hat{A}_1 [7]. In fact this can simply be deduced by the condition that one is looking for a basis of the BPS states which can generate all the other by *positive* linear combinations (up to overall conjugation). Inside the curve of marginal stability, we know that there are only two BPS states, given by a monopole with (electric, magnetic) charge given by

$$\alpha_0 = (0, 1)$$

and a dyon with inner product two with the monopole, given by

$$\alpha_1 = (2, -1)$$

Note that the electro-magnetic inner product given by

$$(e_1, m_1) \cdot (e_2, m_2) = e_1 m_2 - m_1 e_2$$

in this case yields

$$\alpha_1 \cdot \alpha_0 = 2$$

Thus we obtain the quiver of the $SU(2)$ theory as given by the (oriented) affine Dynkin diagram¹ \hat{A}_1 :



The two nodes of the quiver have FI-terms. The $U(1)$ part of the D-term for this quantum mechanical problem will involve

$$(|q_1|^2 + |q_2|^2 + (f_0 - f_1))^2$$

where q_i denote the two bifundamentals, and f_i denote the FI D-term for each of the two nodes. It is clear that for one sign of the FI term there is no ground state. This means that the only ground state arises when one of the two nodes has zero rank, and so we will not have any q_i fields. As we change the sign of FI-term we cross the curve of marginal stability, and now we can have a bound state.

The ground states of this theory have been studied by mathematicians [13–17] in relation with the representations of quivers. See refs. [7, 18–21] for discussions in the physical literature. For this case it was shown that the only allowed representations will have charges given by

$$(2.3) \quad \alpha_0 + n(\alpha_0 + \alpha_1) \text{ or } \alpha_0 + \alpha_1.$$

The first series corresponds to dyons in the weak coupling region and the latter correspond to the massive W boson [7, 21]. Physically this result is obtained by analyzing the D -term equation [7, 18–21]; we shall review the argument in a more general context in §. 2.2.

Encouraged by this example, and assuming there is a quiver description, we can come up with a unique possibility for each matter representation of $SU(2)$. For example consider adding a quark in the fundamental representation. Let us consider the regime given by large quark mass. In this limit the massive field decouples without affecting the bound state structure for the pure $SU(2)$. So we would still have the light degrees of freedom captured by the \hat{A}_1 . On the other hand we have in addition two massive fields which should now be read off from the quiver as well. These two have electric/magnetic charges given by $(1, 0), (-1, 0)$. In addition they both carry a charge $+1$ under the additional $U(1)$ flavor symmetry. We need to add one of these two to generate all the fields in terms of them. We note that since

¹ In the math literature the quiver corresponding to the affine \hat{A}_1 Dynkin diagram with both arrows in the same direction is called the *Kronecker quiver*.

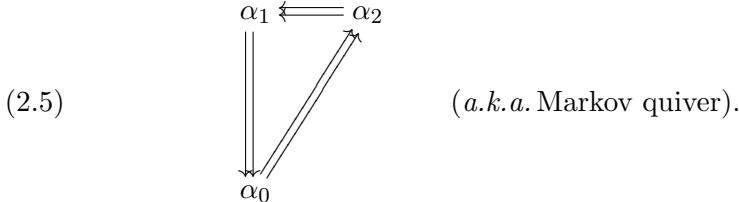
$\alpha_0 + \alpha_1 = (2, 0)$, adding the $(-1, 0)$ as a new node for the quiver, would allow us to obtain the $(1, 0)$ state using positive combination of the three nodes. Thus we end up with the proposed node charges for this theory given by

$$(0, 1), (2, -1), (-1, 0)$$

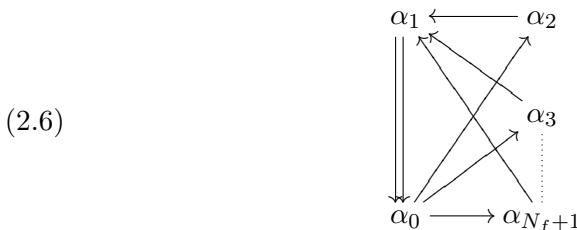
leading to the quiver



We will later present evidence that this quiver correctly reproduces the BPS states for $SU(2)$ with one fundamental field. If we change the matter representation to spin j , we get the same quiver except with $2j$ lines connecting the extra node with the first two nodes. This is because the additional node needed to generate all the BPS states is simply given by $(-2j, 0)$. In particular for the $\mathcal{N} = 2^*$ model, corresponding to mass deformations of the $SU(2)$ $\mathcal{N} = 4$ theory, we obtain:



Similarly for N_f fundamentals, by the same decoupling argument applied to N_f very massive quarks, we get the quiver obtained by adding N_f nodes each of which is connected to the original two nodes in the same way (*i.e.* by single arrows making oriented triangles together with the $SU(2)$ double arrow):



Since the BPS quiver captures the BPS degeneracies, it is natural to believe that the quiver completely captures the corresponding $\mathcal{N} = 2$ theory. In particular, it is natural to assume that, with generic enough superpotential for the quiver, the resulting ground states are universal and insensitive to the precise choice of the superpotential. Moreover changing

the FI-terms may result in wall-crossing phenomena, but should not be necessary to specify the $\mathcal{N} = 2$ theories if we are to study them up to moduli deformation.

The characterization of $4d, \mathcal{N} = 2$ theories using quivers is very powerful. This shifts the classification of $\mathcal{N} = 2$ theories to classification of allowed quivers up to mutations. But we have to first come up with a more precise criterion of what we mean by a BPS quiver, and also whether all $\mathcal{N} = 2$ theories admit such a description for their BPS states.

2.1. Generalities of quivers. Consider an $\mathcal{N} = 2$ theory in 4 dimensions which at a generic point on the Coulomb branch has an abelian rank r gauge symmetry $U(1)^r$. In addition we assume the theory has a rank f flavor symmetry group given by $U(1)^f$ for generic values of mass deformations. Then the total rank D of the charges determining the BPS mass of the $\mathcal{N} = 2$ theory is given by

$$D = 2r + f$$

given by r electric, r magnetic and f flavor charges. The set of BPS states should thus include at least $2r + f$ states. We say an $\mathcal{N} = 2$ supersymmetric gauge theory admits a BPS quiver, if the following conditions are satisfied:

1) There are $2r + f$ BPS hypermultiplets with charges $\alpha_i \in \Gamma^{2r+f}$ with spin 0, with their $\mathcal{N} = 2$ central charge lying on the same half plane, and such that all the BPS states are given by a *positive* linear combination of them, up to an overall conjugation. In other words, if there is a BPS particle of charge β , then

$$\beta = \pm \sum_{i=1}^{2r+f} n_i \alpha_i$$

where $n_i \geq 0$.

2) There is a quiver supersymmetric quantum mechanics with 4 supercharges, and $2r + f$ nodes, with unitary gauge groups on each node, such that as we vary the ranks of the unitary group, the ground states of the theory are in 1–1 correspondence with the BPS states. Moreover the nodes are in 1–1 correspondence with the BPS states with charges α_i , and the ground states corresponding to the supersymmetric quantum mechanics with gauge group $\prod U(n_i)$ corresponds to state(s) with charge $\beta = \sum_i n_i \alpha_i$.

3) The number of bi-fundamental between the nodes i, j is given by the electro-magnetic skew-symmetric inner product $\alpha_i \cdot \alpha_j$.

4) As we change the parameters of the theory, and in particular when one of the central charges $Z(\alpha_i)$ is about to exit the same half plane as the other α 's, we replace the corresponding BPS generator α_i with the conjugate state with charge $-\alpha_i$. Furthermore we replace all the other BPS states with charge α_j which have positive inner product n_{ij} with α_i with other BPS generators having charge

$$\alpha'_j = \alpha_j + n_{ij} \alpha_i$$

leading to a new quiver which is mutated (see sect. 5.2 for more details).

Above we have seen examples of $\mathcal{N} = 2$ theories for which there is a quiver description. Note that whether an $\mathcal{N} = 2$ theory admits a quiver description may and in fact does depend on which point on its moduli space we are. An example of this is the $\mathcal{N} = 2^*$ theory, say for the $SU(2)$ gauge group. As we have indicated for sufficiently large mass for the adjoint matter, there is a quiver description. However, if the mass is turned off we obtain an $\mathcal{N} = 4$ gauge theory. It is easy to see that for this value of moduli the $\mathcal{N} = 2^*$ theory cannot admit a quiver realization. The reason is that we would need to come up with three BPS states (since $r = 1, f = 1$) whose positive span contains all the BPS charges. On the other hand we know that the BPS states of $\mathcal{N} = 4$ are given by one hypermultiplet and one vector multiplet (in the $\mathcal{N} = 2$ counting) for each relatively prime p, q with electromagnetic charge (p, q) . Clearly this cannot be given by the positive span of three vectors which are in the same half-plane. In fact quite generally if we consider the phase of the central charge of $\mathcal{N} = 2$ BPS states, the condition that they be spanned by a finite number of BPS states implies that the phases of BPS central charges do not form a dense subset of the circle, which is not the case for this theory. Thus we have learned that there are some $\mathcal{N} = 2$ theories which have BPS quivers in some region of the moduli but not at all points on the moduli.

From this example one may be tempted to conclude that all the $\mathcal{N} = 2$ theories have at least some points on their moduli where there is a BPS quiver description. However, this turns out not to be the case. In fact all the Gaiotto theories of rank 2 with $g > 2$ and with no punctures are believed to be of this type [24, 25]. These theories admit no mass deformation, and in some sense are the analog of the $\mathcal{N} = 2^*$ at $m = 0$ which are permanently stuck there. The case of $g = 2$ with no punctures is different. In one duality frame, that theory corresponds to an $SU(2)^3$ theory with two half-hypermultiplets in $(\mathbf{2}, \mathbf{2}, \mathbf{2})$. The two half-hypermultiplets, form one full hypermultiplet and that can receive a mass (though its IR Seiberg-Witten geometry, unlike the $m = 0$ point which is given by Gaiotto curve, is unknown). It is natural to conjecture that all the $\mathcal{N} = 2$ theories whose BPS phases do not form a dense subspace of the circle admit a BPS quiver description at such points in moduli (of course as discussed this is a necessary condition).

Given a BPS quiver, we can read off r, f as follows: Consider the skew-symmetric matrix which we can read off from the quiver links, that is $B_{ij} = \alpha_i \cdot \alpha_j$. The rank of B is $2r$ while f is the corank of B , i.e. $D - 2r$.

2.2. BPS spectra and representation theory. The BPS spectrum of a $\mathcal{N} = 2$ may also be understood in terms of the representation theory of the associated quiver Q [6, 7, 18–21]. A representation associates a vector space V_i to each node i of Q and a linear map $V_i \xrightarrow{\phi_a} V_j$ to each arrow $i \xrightarrow{a} j$. We write $d_i = \dim V_i$ ($i = 1 \dots, D$) for the dimension vector of the

representation; in terms of quiver quantum mechanics, d_i corresponds to the rank N_i of the gauge group at the i -th node.

As a first example, consider the BPS spectrum of the ADE Argyres–Douglas theories determined² in [2, 26]. The quiver $Q_{\mathfrak{g}}$ of these theories is simply the Dynkin diagram of the associated Lie algebra $\mathfrak{g} \in ADE$ with some orientation of the edges (all orientations being equivalent up to mutation [27]), so that the charge lattice gets identified with the root lattice of \mathfrak{g} , $\Gamma \simeq \sum_i \mathbb{Z} \alpha_i$. The ADE Argyres–Douglas theories have two³ special⁴ chambers, **(S)** and **(W)**, having a finite BPS spectrum consisting, respectively, in

- (S):** one BPS hypermultiplet for each simple root with charge vector α_i ;
- (W):** one BPS hypermultiplet for each positive root of \mathfrak{g} with charge vector the same positive root $\sum_i n_i \alpha_i$, ($n_i \geq 0$).

This result may be understood in terms of the Gabriel theorem [15–17] which puts the above Argyres–Douglas models in one-to-one correspondence with the quivers having finitely many non-isomorphic indecomposable representations. The Gabriel map sends the representation of a Dynkin quiver with dimension vector d_i into the element of the root lattice $\sum_i d_i \alpha_i \in \Gamma_{\mathfrak{g}}$. Under this map, the simple representations correspond to the simple roots α_i , and the indecomposable representations to the positive roots.

Gabriel theorem has been generalized to arbitrary quivers by Kac [13]. So the charge lattice may be always identified with the root lattice of some Lie algebra, and stable BPS states are mapped to positive roots under this identification. Real positive roots correspond to *rigid* indecomposable representation (no continuous moduli) so they are naturally related to BPS hypermultiplets; imaginary positive roots have moduli so, in general, they correspond to higher spin BPS multiplets. Which positive roots actually correspond to stable BPS particles depends on the particular chamber. Concretely, given a quiver Q we consider the central charge function $Z(\cdot)$ which associates to a representation R , having dimension vector $d_i(R)$, the complex number $Z(R) = \sum_i d_i(R) Z_i$, where $\arg Z_i \in [0, \pi[$. We say that a representation R is stable (with respect to the given $Z(\cdot)$) if [22]

$$(2.7) \quad \arg Z(S) < \arg Z(R)$$

for all proper subrepresentations S of R (this condition is called Π –stability in [18, 19]). Physically, this is the requirement that the BPS state of charge vector $\sum_i d_i(R) \alpha_i$ cannot decay into states having charge $\sum_i d_i(S) \alpha_i$ because there is no phase space.

² See [22] **Corollary 1.7** for an equivalent mathematical statement.

³ In fact many such chambers corresponding to different orientations of the Dynkin graph. These chambers have the same spectrum but differ for the BPS phase order [2]. See also appendix A.

⁴ For rank $\mathfrak{g} > 2$ there are other BPS chambers as well. The BPS spectrum is always finite.

Notice that simple representations, associated to the simple roots α_i , correspond to BPS hypermultiplets which are stable in all chambers. The existence of such a spanning set of universally stable hypermultiplets is a necessary condition for the $\mathcal{N} = 2$ theory to admit a quiver in the present sense.

As anticipated above, this representation-theoretical stability condition may be understood from the quiver quantum mechanics viewpoint as a consequence of the D -term equation in presence of FI terms which depend on the given central charges $Z_j = m_j e^{i\theta_j}$. Without changing the chamber, we may assume that the $\arg Z_i$'s are all very close together. Then, if $\arg Z(R) = \alpha$,

$$\begin{aligned} Z(S)/Z(R) &= \frac{\sum_j d_j(S) m_j e^{i(\theta_j - \alpha)}}{|Z(R)|} \\ (2.8) \quad &\approx \frac{1}{|Z(R)|} \left(\sum_j d_j(S) m_j + i \sum_j d_j(S) m_j (\theta_j - \alpha) \right) \\ &= r_1 + \frac{i}{|Z(R)|} \sum_j d_j(S) \vartheta_j \end{aligned}$$

where r_1 is real positive and $\vartheta_j = m_j(\theta_j - \alpha)$. Thus the stability condition (2.7) is equivalent to the condition that

$$(2.9) \quad \sum_i d_j(S) \vartheta_j < 0$$

for all proper subrepresentations S of R (this condition is called ϑ -stability [14]). A theorem by King (**Proposition 6.5** of [14]) states that an indecomposable representation R is ϑ -stable if and only if it satisfies the equation

$$(2.10) \quad \sum_{t(\alpha)=j} \Phi_\alpha^\dagger \Phi_\alpha - \sum_{h(\alpha)=j} \Phi_\alpha \Phi_\alpha^\dagger = \vartheta_j \mathbf{1},$$

which is the D -term equation in presence of the FI terms ϑ_j .

After the ADE Argyres–Douglas models, the next simplest instances are the $\mathcal{N} = 2$ theories having a quiver Q whose underlying graph is an affine \widehat{ADE} Dynkin diagram with arrows oriented in such a way that there are no oriented cycles. Up to equivalence, the affine quivers are

- (1) $\widehat{A}(p, q)$, with $p \geq q \geq 1$, corresponding to the \widehat{A}_{p+q-1} Dynkin diagram oriented in such a way that p arrows point in the positive direction and q in the negative one. We exclude $q = 0$ since $\widehat{A}(p, 0) \sim D_p$ and we get back a Argyres–Douglas model;
- (2) \widehat{D}_r , \widehat{E}_6 , \widehat{E}_7 , and \widehat{E}_8 . In these cases, since the Dynkin diagram is a tree, all orientations are mutation equivalent.

The charge lattice is identified with the root lattice $\Gamma_{\widehat{\mathfrak{g}}}$, and the only charge vectors which may possibly correspond to stable BPS states are:

- *real* positive roots \Rightarrow BPS hypermultiplets;

- the indivisible imaginary root $\delta \Rightarrow$ BPS vector-multiplet.

In particular, in any BPS chamber, we have *at most* one vector; indeed one of the result of the present paper is that affine $\mathcal{N} = 2$ theories correspond to a single $SU(2)$ SYM coupled to a vector-less $\mathcal{N} = 2$ system.

The simple roots are always stable. In fact, there exists a chamber, corresponding to the strong coupling regime, in which the *only* states are those associated to the simple roots⁵. Indeed, we may number the nodes of an affine quiver, without oriented cycles, from 1 to D in such a way that each vertex i is a source in the full subquiver of vertices $1, \dots, i$ [22, 23]. In this numeration, if we have

$$(2.11) \quad \arg Z_1 < \arg Z_2 < \dots < \arg Z_D$$

we see recursively that the indecomposable are just the simple roots.

In the weak coupling regime the state associated to δ , *i.e.* the W -boson, is stable together with a tower of hypermultiplets corresponding to a certain subset of Δ_+^{re} .

We close this section by checking these predictions for $SU(2)$ $\mathcal{N} = 2$ SQCD with $N_f = 0, 1, 2, 3$ fundamental hypermultiplets [24, 30, 31]. The case $N_f = 0$, corresponding to the quiver $\widehat{A}(1, 1)$, was already discussed around eqn.(2.2). It is easy to check that the stable representations in the weak coupling chamber, namely δ and the real positive roots, correspond to the BPS states present in the physical spectrum [21, 22].

$$\underline{N_f = 1}$$

Mutating⁶ the $N_f = 1$ quiver (2.6) at the hypermultiplet vertex (indicated by a curled arrow in the figure) we get the affine $\widehat{A}_2(2, 1)$ quiver

$$(2.12) \quad \begin{array}{ccc} \begin{array}{c} \bullet \xleftarrow{\quad} \bullet \\ \parallel \quad \nearrow \curvearrowright \\ \bullet \end{array} & \longrightarrow & \begin{array}{c} \alpha_1 \longrightarrow \alpha_2 \\ \downarrow \quad \nearrow \\ \alpha_0 \end{array} \end{array}$$

One has $2e \equiv \delta = \alpha_0 + \alpha_1 + \alpha_2$ while the flavor charge is proportional to $f = \alpha_2 - (\alpha_0 + \alpha_1)$, so in terms of the usual charges (e, m, f) the affine simple roots are

$$(2.13) \quad \alpha_0 = (0, 1, -1), \quad \alpha_1 = (1, -1, 0), \quad \alpha_2 = (1, 0, 1).$$

⁵ For an argument along the lines of [2], see appendix A.

⁶ Detailed definitions of the quiver mutations are given in section 5.2.

which is the correct strong coupling spectrum. The known weak coupling spectrum is also consistent with representation theory.

$N_f = 2$

Mutating the $N_f = 2$ quiver (2.6) at both hypermultiplet vertices we get the affine $\widehat{A}_3(2, 2)$ quiver



Again, the strong coupling BPS spectrum is given by four hypermultiplets of charges $\alpha_0, \alpha_1, \alpha_2, \alpha_3$. In the weak coupling we have a vector multiplet of charge $\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$ and a tower of BPS hypermultiplets whose charge vectors belong to $\Delta_+^{\text{re}}(\widehat{A}_3)$.

$N_f = 3$

Mutating the $N_f = 3$ quiver one gets the \widehat{D}_4 affine quiver



Again, the strong coupling spectrum consist of five hypermultiplets with charge vectors $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4$, while in the weak coupling we have one BPS vector multiplet with charge vector

$$(2.16) \quad \sum_{i \neq 1} \alpha_i + 2\alpha_1,$$

where α_1 is the simple root associated to the central node in (2.15), and the usual tower of dyons with charge vectors in $\Delta_+^{\text{re}}(\widehat{D}_4)$.

3. Definition of complete $\mathcal{N} = 2$ theories

In this section we motivate the definition of a special class of $\mathcal{N} = 2$ theories which we will call ‘complete $\mathcal{N} = 2$ gauge theories’. Consider an $\mathcal{N} = 2$ theory with $D = 2r + f$ BPS charges. This in particular means that we have D central charges $Z_i \in \mathbb{C}$, with $i = 1, \dots, D$ which appear in the BPS algebra. It is natural to ask if they can be arbitrarily varied. In other

words we are asking if the map from the moduli space \mathcal{M} to D -dimensional complex plane, given by the central charges,

$$Z : \mathcal{M} \rightarrow \mathbb{C}^D$$

is at least locally onto. For this to happen we need to have at least D complex parameters in the moduli space \mathcal{M} of the theory. Quite generally we can identify r complex parameters with labelling the Coulomb branch, and f parameters for varying the masses. In addition there could be additional coupling constants. In order to vary the central charges independently, we need at least r additional parameters. This suggests that if we can in addition vary the r coupling constants of the theory independently, then we have a complete $\mathcal{N} = 2$ theory. Note that this latter condition may not be possible in general. For example, for $SU(N)$ gauge theory we expect only one coupling constant but $r = N - 1$ dimensional Coulomb branch. We can in principle formally deform the coupling constants of the $U(1)$'s in the IR, but there is no guarantee that there is a UV complete theory which allows this (in fact we will argue in this paper that this is not possible). Moreover, there are some $\mathcal{N} = 2$ theories which do not even have a freedom to vary one coupling constant. For example the Minahan–Nemeschansky theories [32, 33] are of this type, where the coupling constant is completely fixed by the masses and the point on the Coulomb branch.

On the other hand it is clear that an $\mathcal{N} = 2$ theory consisting of asymptotically free matter spectrum with a gauge group $G = SU(2)^{\otimes r}$ is complete in the above sense, because we have r couplings, r Coulomb branch parameters, and one mass parameter for each matter representation. In particular all the rank 2 Gaiotto theories [34] are complete in this sense. One can also ask if the dimension of \mathcal{M} can be bigger than D . This is in principle possible, because the coupling constants of a $U(1)^r$ theory is a symmetric complex $r \times r$ matrix, which has $(r^2+r)/2$ entries. Nevertheless we will later argue that the dimension of \mathcal{M} is at most D , which gets saturated by complete theories.

The question we pose is the classification of all complete $\mathcal{N} = 2$ gauge theories which admit a BPS quiver. In order to accomplish this, we will use the 4d/2d correspondence of [2] that we will review in the next section.

4. 4d-2d correspondence reviewed

There has been a number of links between 4d $\mathcal{N} = 2$ theories and 2d QFT's. In particular two such correspondences were suggested in [2]. In this section we review one of those conjectured correspondences, which proves important for our applications.

This duality maps 4d theories with $\mathcal{N} = 2$ supersymmetry (with 8 supercharges) to 2d theories with $\mathcal{N} = 2$ (with 4 supercharges). The specific case where the map can be demonstrated explicitly is for $\mathcal{N} = 2$ theories in 4d which can be constructed in type II strings on local Calabi-Yau manifolds. The idea is that the worldsheet of the type II strings involves an $\mathcal{N} = 2$

superconformal theory, with $\hat{c} = 3$. Furthermore when the 4d $\mathcal{N} = 2$ theory can be decoupled from gravity, one is discussing the geometry near a local singularity of Calabi-Yau. In such a case, one can expect that the theory has a Liouville field, and that the $\mathcal{N} = 2$ worldsheet theory decomposes to a mixed product of the Liouville field and an $\mathcal{N} = 2$ 2d QFT. The accompanying $\mathcal{N} = 2$ QFT may be massive or conformal, which can be read off by freezing the value of the Liouville field. This worldsheet $\mathcal{N} = 2$ theory could be massive without contradiction as its coupling to Liouville can make it conformal. Moreover, since the central charge of the Liouville is $\hat{c} \geq 1$, this implies that the central charge of the accompanying 2d theory is $\hat{c} \leq 2$.

An example of this is the following: Consider again type IIA on the local Calabi-Yau threefold given by $\mathbb{C}^3/\mathbb{Z}_3$ or its blow ups, which is the total space of $O(-3)$ line bundle over \mathbb{P}^2 . Then the worldsheet theory has a mirror Landau-Ginzburg description given by [35, 36],

$$W = \exp(-Y_1) + \exp(-Y_2) + \exp(-Y_3) + \exp(+Y_1 + Y_2 - 3Y_3) \exp(-t)$$

where Y_i are chiral \mathbb{C}^* valued superfields, and t denotes the complexified Kahler class of \mathbb{P}^2 . We can treat an overall shift of Y as a Liouville field. Fixing that, will yield a theory with one less field given by

$$\begin{aligned} W &= \exp(-Y_3) [\exp(-Y'_1) + \exp(-Y'_2) + 1 + \exp(Y'_1 + Y'_2) \exp(-t)] \\ &= \exp(-Y_3) \cdot W'(Y'_1, Y'_2) \end{aligned}$$

where

$$Y'_1 = Y_1 - Y_3, Y'_2 = Y_2 - Y_3$$

One recognize $W'(Y'_1, Y'_2)$ as the superpotential for massive 2d theory which is the mirror of sigma model to \mathbb{P}^2 [1, 35].

Similarly, if we consider the type IIA on a Calabi-Yau corresponding to $\mathbb{C}^2/\mathbb{Z}_2 \times \mathbb{Z}_2$ or its blow up, the total space of the $O(-2, -2)$ bundle over $\mathbb{P}^1 \times \mathbb{P}^1$ similar manipulations (see [36]) will yield a factor W' given by

$$W' = \exp(-X_1) + \exp(X_1) \exp(-t_1) + \exp(-X_2) + \exp(X_2) \exp(-t_2) + 1$$

where the t_i are the two complexified Kahler classes of the \mathbb{P}^1 's. Again, one recognizes W' as the mirror to the 2d sigma model on $\mathbb{P}^1 \times \mathbb{P}^1$. By taking a special limit (corresponding to taking one of the \mathbb{P}^1 's much larger than the other) leads to geometric engineering of $\mathcal{N} = 2$ pure $SU(2)$ in 4 dimensions, leading to a 2d factor with superpotential (after an overall rescaling of W')

$$W' \rightarrow \exp(-X_1) + \exp(+X_1) + X_2'^2 + u$$

where one recognizes $W' = 0$ as the SW curve for the pure $SU(2)$ theory. This 2d factor is equivalent to the mirror of the sigma model on \mathbb{P}^1 (where the X'_2 part gives a trivial massive theory).

From these examples the general idea emerges that at least for all the $\mathcal{N} = 2$ theories which can be engineered in type II strings, we would obtain an accompanying 2d $\mathcal{N} = 2$ theory which is the factor of the worldsheet theory. However, there is more to this map. The BPS quivers

of the 4d theories naturally encode the soliton data of the corresponding 2d theory. The nodes of the 4d BPS quiver map to the 2d vacua, and the lines connecting them map to the soliton between them.⁷ In particular we recognize the 4d BPS quiver of the $\mathbb{C}^3/\mathbb{Z}_3$ model as encoding the three vacua of the \mathbb{P}^2 model and the corresponding bifundamentals as mapping to the kinks connecting them, and similarly that of the $\mathbb{C}^2/\mathbb{Z}_2 \times \mathbb{Z}_2$, which maps to the 2d data of the $\mathbb{P}^1 \times \mathbb{P}^1$ sigma model. Another example is the theory corresponding to $\mathcal{N} = 2$ theory for the pure $SU(2)$. As we just saw the corresponding 2d theory corresponds to the sigma model on \mathbb{P}^1 . This massive theory has two vacua and two solitons between the two. This is exactly the structure of the quiver for the $SU(2)$ theory as we already discussed.

The idea for this map is that there are canonical D-branes associated to LG vacua, as discussed in [36], corresponding to Lagrangian subspaces of LG. These we can identify with the worldsheet description of the BPS states. Moreover the intersection pairing between these Lagrangian cycles in 2d was mapped in [36] to the number of kinks connecting the vacua. On the other hand the intersection of D-branes give bifundamental fields, thus explaining this connection.

Based on many such examples it was suggested in [2] that for every $\mathcal{N} = 2$ theory in 4d, there is an associated 2d theory with $\mathcal{N} = 2$ supersymmetry. Moreover it was proposed that the quiver of the 4d theory get mapped to the vacua and kink structure of the 2d theory. On the other hand we know that not every 4d theory has a quiver description. This actually has a 2d counterpart: Not every 2d theory has isolated vacua and kinks between them. Thus the 4d/2d correspondence is more general than the map between their associated quivers. In this paper we assume the validity of this correspondence and use it to classify complete $\mathcal{N} = 2$ theories in 4d, which were defined in the previous section.

5. Complete $\mathcal{N} = 2$ theories and quivers of finite mutation type

In this section we argue that complete $\mathcal{N} = 2$ theories in 4d are mapped to 2d theories with $\hat{c} \leq 1$ in the UV. We will be interested in the case where both theories admit a quiver, though we believe the map is more general. Furthermore we review the mathematical classification of quivers of finite mutation type.

5.1. Completeness and finiteness of mutation type. The basic idea for showing the connection between completeness and finiteness of mutation type for the quiver is very simple: First we will assume that the 4d theory admits a BPS quiver. In such a case we are looking for theories whose dimension of moduli space is equal to the number of nodes. On the other hand, mapping this theory to 2d, and identifying the nodes, with vacua, it means that we are looking for 2d theories which have as many deformations

⁷ The extra data of orientation of the arrows is also encoded in the 2d theory in an implicit way, as we discuss later in the context of examples.

as the number of massive vacua. For 2d theories, with $\mathcal{N} = 2$ we know that, in the UV, the number of allowed deformations is given by the number of operators with dimension less than or equal to 1, *i.e.* relevant or marginal operators. On the other hand there are as many chiral fields as the vacua, with the highest chiral field having dimension \hat{c} . Since the dimension of deformations is equal to the number of vacua, this means all chiral fields can be used to deform it, including the one with maximal dimension. But given the bound on the allowable deformations, this implies that $\hat{c} \leq 1$.

On the other hand we can ask the question of what kinds of quivers are allowed for 2d theories with $\hat{c} \leq 1$. We argue that these must have a finite mutation type. In other words, there cannot be infinitely many mutation orbits of the quiver. Indeed, as noted before, the mutation of the quiver maps to wall crossing for the 2d BPS states. But since we have as many parameters to vary as the number of vacua, we can use this freedom to induce arbitrary wall crossings for the 2d theory. On the other hand each wall crossing leads to a mutation of the quiver. Thus arbitrary mutations of the quiver are physically realized. Moreover since we have enough parameters we can decouple as many vacua as we wish. In particular we can decouple all vacua except for any fixed pair. In this way we end up with a theory with only two vacua with some kinks between them. It is known [1] that the number of kinks between them is less than or equal to 2 for the theory to exist. This implies that no matter what quiver mutations we consider, the number of links between any pair cannot grow more than 2 for complete $\mathcal{N} = 2$ theories. This in particular implies that the quivers of complete $\mathcal{N} = 2$ theories should be finite in number (otherwise this number would grow at least for a pair of vacua).

It turns out that the quivers of finite mutation type have been classified by mathematicians [3–5]. Of course from what we have said above, we need to further restrict to quivers where there is no more than two links between any pairs of nodes. This turns out to be automatically true for all quivers of finite mutation type with more than two nodes and so we do not need to further impose this condition.

On the other hand for quivers with two nodes, we need to restrict to ones with less than three links.

Of course it is not clear that all the quivers of finite mutation type (apart from the restriction for the two node case) do arise for some complete $\mathcal{N} = 2$ gauge theory. We have only shown that complete gauge theories lead to finite mutation type quivers. Nevertheless we show this is also sufficient and identify each finite mutation type quivers with a unique $\mathcal{N} = 2$ theory in 4d. Before doing so, in the next subsection we review the mathematical result for classification of quivers of finite mutation type.

5.2. Quivers of finite mutation type. The class of quivers of interests in $\mathcal{N} = 2$ theories are the ‘2-acyclic’, namely the ones without loops

(arrows which start and end in the same node) and no arrows with opposite orientations between the same two nodes. Physically this is because a loop corresponds to an adjoint matter which can be given mass and thus disappear from consideration of BPS spectrum. For the same reason only the net number of bi-fundamentals between pairs of nodes enter the discussion because the others can be paired up by superpotential mass terms and disappear from the study of ground states of the SQM. In this paper when we discuss quivers we restrict to this class. Specifying such a quiver Q with D nodes is equivalent to giving an integral $D \times D$ skew-symmetric matrix B (called the *exchange matrix*) whose (i, j) entry is equal to the number of arrows from the i -th node to the j -th one (a negative number meaning arrows pointing in the opposite direction $j \rightarrow i$).

A mutation of such a quiver Q is given by a composition of elementary mutations. There is an elementary mutation for each vertex of Q . The elementary mutation at the k -th vertex, μ_k , has the following effect on the quiver [27, 37, 38] (for reviews see [39–41]):

- (1) It inverts the direction of all arrows going in/out the k -th vertex;
- (2) each triangle having k as a vertex gets mutated as in the following figure

Q	$\mu_k(Q)$	Q	$\mu_k(Q)$

where r, s, t are non-negative integers, and an arrow $i \xrightarrow{l} j$ with $l \geq 0$ means that l arrows go from i to j while an arrow $i \xrightarrow{l} j$ with $l \leq 0$ means $|l|$ arrows going in the opposite direction.

In terms of the exchange matrix B_{ij} the mutation μ_k reads [27, 37, 41]

$$(5.1) \quad \mu_k(B_{ij}) = \begin{cases} -B_{ij} & \text{if } i = k \text{ or } j = k; \\ B_{ij} + \text{sgn}(B_{ik}) \max\{B_{ik}B_{kj}, 0\} & \text{otherwise.} \end{cases}$$

The definition implies that μ_k is an involution:

$$(5.2) \quad (\mu_k)^2 = \text{identity.}$$

From the box we see that the mutation μ_k is particularly simple when the node k is either a sink (all arrows incoming) or a source (all arrows outgoing). In these cases, μ_k just inverts the orientation of the arrows through the k -th node.

Two quivers are said to be in the *same mutation-class* (or mutation-equivalent) if one can be transformed into the other by a finite sequence

of such elementary mutations. A quiver is said to be *mutation-finite* if its mutation-class contains only finitely many distinct quivers.

There is a Java applet due to B. Keller [42] which implements the quiver mutations and computes the mutation-class of a quiver up to sink/source equivalence (*i.e.* two quivers are identified if they differ by a mutation at a sink/source).

According to the Felikson–Shapiro–Tumarkin theorem [5] the complete list of mutation-finite quivers is the following:

- (1) quivers with at most two nodes;
- (2) quivers representing adjacency matrices of ideal triangulations of bordered surfaces with punctures and marked points on the boundaries [3] (to be discussed in the next subsection);
- (3) the quivers mutation equivalent to the nine E -type Dynkin diagrams⁸

$$\begin{array}{ll} \text{finite:} & E_6, E_7, E_8 \\ \text{affine:} & \hat{E}_6, \hat{E}_7, \hat{E}_8 \\ \text{elliptic:} & \hat{\hat{E}}_6, \hat{\hat{E}}_7, \hat{\hat{E}}_8, \end{array}$$

having rank D equal to the sum of the subscript plus the number of hats. The quivers associated to the unhatted and single hatted E -theories are the usual Dynkin diagrams of the E -type, and different orientation of the arrows give mutation equivalent quivers. For \hat{E}_r the arrows are cyclicly oriented in all triangles (all such orientations are mutation equivalent) see figure 1;

- (4) the two Derksen–Owen mutation classes X_7 and X_6 , (of rank 7 and 6, respectively) [4]. There are five distinct quivers in the class of X_6 , and just two in the one of X_7 . See figure 1.

In particular, all finite-mutation quivers with more than 10 nodes arise from ideal triangulations of surfaces in the sense of [3].

In [2] it was shown that the only two-node quivers which correspond to sensible $4d$ $\mathcal{N} = 2$ theories are (orientations) of the Dynkin graphs of $A_1 \times A_1$, A_2 and \hat{A}_1 . If Q is a finite-mutation quiver with $D \geq 3$, all its mutation-equivalent quivers have at most *double* arrows. The same is true for the three $D = 2$ Dynkin quivers $A_1 \times A_1$, A_2 and \hat{A}_1 . Then the property characterizing quivers corresponding to complete $\mathcal{N} = 2$ models is that in their mutation class there is no quiver with arrows of multiplicity > 2 . When in this paper we loosely refer to finite-mutation quivers, we mean those having this property. It is remarkable that all such quivers correspond to meaningful $4d$ $\mathcal{N} = 2$ theories, in fact to complete ones in the present sense.

⁸ In Saito's notation [43] the root system \hat{E}_r is written as $E_r^{(1,1)}$.

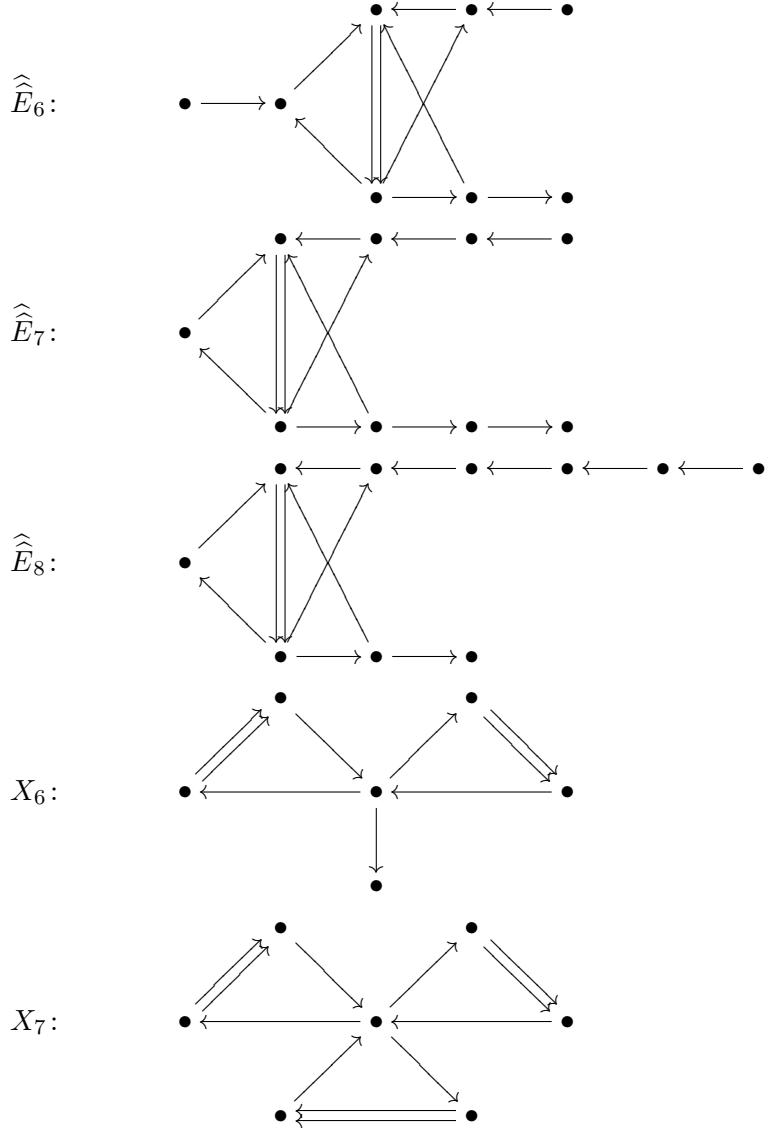


FIGURE 1. The three elliptic E -type Dynkin diagrams oriented as to give finite mutation quivers, and the two Derksen–Owen quivers.

5.2.1. Quivers from ideal triangulations of bordered surfaces. All but 11 mutation-finite classes arise from ideal triangulations of surfaces studied in ref. [3]. Here we summarize the results of [3] we need below. Let \mathcal{C} be an oriented surface of genus g with n punctures, b boundary components, and c_i marked points on the i -th boundary component ($i = 1, 2, \dots, b$). By a compatible collection of arcs we mean a set of curves, identified up to isotopy,

which end at the punctures or the marked points, do not intersect themselves or each other except at the end points, and cannot be contracted to a puncture or a boundary segment. Any maximal such compatible collection contains

$$(5.3) \quad D = 6g - 6 + 3n + \sum_i (c_i + 3)$$

arcs, and it is called an *ideal triangulation* of \mathcal{C} . This definition allows for *self-folded* triangles whose sides are not all distinct, see figure (5.4)



Given an ideal triangulation we number the arcs as $1, 2, \dots, D$, and define a skew-symmetric $D \times D$ integral matrix B as follows [3]: if i and j are not internal arcs of self-folded triangles (as is the arc int in figure (5.4)) we set B_{ij} to be the sum over all triangles Δ of which both arcs are sides of the weight w_{ij}^Δ . w_{ij}^Δ is equal $+1$ (resp. -1) if the side i of Δ follows (resp. precedes) the side j in the anticlockwise order. If i is an internal arc of a self-folded triangle we set $B_{ij} \equiv B_{ext(i)j}$, where $ext(i)$ is the external arc of the self-folded triangle containing i (see figure (5.4)). The matrix B is called the *adjacency matrix* of the ideal triangulation.

The adjacency matrix B defines a 2-acyclic quiver as before. From the definition, one has

$$(5.5) \quad B_{ij} = -2, -1, 0, 1, 2.$$

One shows [3] that two quivers, Q_1 and Q_2 , representing adjacency matrices of two different ideal triangulations of the same surface \mathcal{C} are mutation equivalent. Moreover, any quiver which is mutation equivalent to the adjacency quiver of a surface is the adjacency quiver for some ideal triangulation of that surface. This, together with eqn.(5.5), implies that all adjacency quivers are of finite-mutation type.

A mutation invariant of the quiver is automatically a topological invariant of \mathcal{C} . Since the rank of B is invariant under mutation [44], the corank of B is a topological invariant equal to the number of punctures plus the number of boundary components with c_i even [3]

$$(5.6) \quad f = D - \text{rank } B = n + \sum_{c_i \text{ even}} 1.$$

From the discussion in section 2.1 we see that this topological invariant is equal to the number of flavor charges in the $\mathcal{N} = 2$ theory.

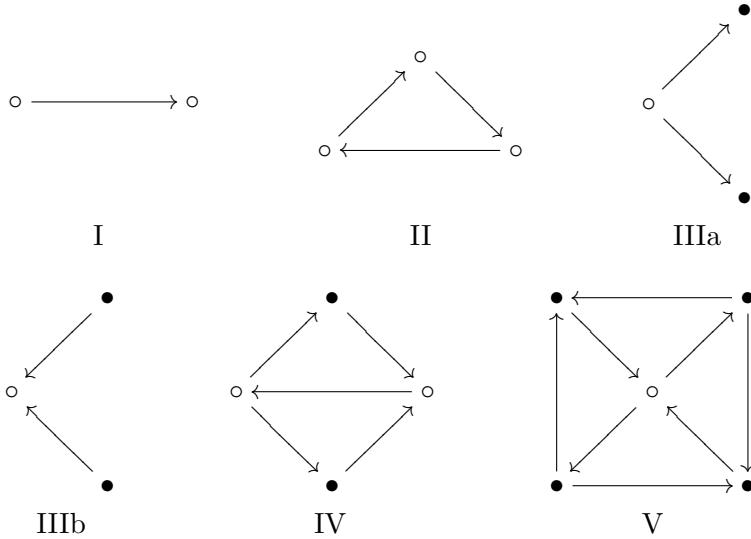


FIGURE 2. The quiver blocks of Type I–V [3].

A quiver is the adjacency quiver of a bordered surface if and only if it can be decomposed into quiver blocks [3]. There are five types of blocks (see figure 2). A quiver is an adjacency quiver of some bordered surface iff it can be obtained by gluing together a collection of blocks of types I, II, III, IV, and V by identifying together pairs of white nodes \circ . If the resulting quiver contains a pair of arrows connecting the same pair of vertices, but pointing in opposite directions, they must be removed.

White nodes represent arcs which are ordinary sides of triangles, and identifying pairs of them is equivalent to gluing the corresponding (generalized) triangles along that arc. More precisely, each block represents a piece of the triangulation [3]:

- a block of type I represents a triangle with one side along the boundary of the surface \mathcal{C} ;
- a block of type II represents a triangle with all three sides inner triangulation arcs;
- a block type III represents a punctured 2-gon⁹ with a side on the boundary;
- a block of type IV represents a 2-gon containing a folded triangle;
- a block of type V represents a 1-gon containing two folded triangles.

Finally, if a quiver may be decomposed into blocks in a unique way, there is (topologically) precisely one surface \mathcal{C} whose triangulations correspond to the quivers of its mutation class; it is possible (but very rare) that two

⁹ By an n -gon we mean a polygon with n sides, that is a disk with n marked points on the boundary.

topologically distinct surfaces have the same class of adjacency quivers. The physical meaning of this non uniqueness will be discussed in the next section.

The two black nodes of a type III block are terminal nodes and in particular sink/sources. To avoid special cases in some of the statements below, it is convenient to adopt the following convention: whenever we have a quiver Q with some type III blocks in its decomposition, we replace it by the physically equivalent quiver obtained by mutating Q at one terminal node for each type III block. We call this sink/source equivalent quiver the *normalized* quiver.

5.3. Some basic features of mutation-finite quivers. In this section we discuss some general features of mutation-finite quivers. One basic feature of mutation-finite quivers is that any full subquiver is also mutation finite. We interpret this in the 4d language as saying that there is a choice of moduli which reduces the light degrees of freedom of the theory to the corresponding subquiver. This is the correct interpretation also from the viewpoint of 4d/2d correspondence: From the 2d perspective the nodes correspond to 2d vacua and we can change the moduli of the 2d theory by taking all the nodes outside the subquiver to have infinitely large value for the superpotential. The inverse can also be done. Namely one can start with a mutation-finite quiver and add additional nodes and arrows subject to maintaining mutation-finiteness. This process should also be interpretable physically as coupling a giving physical theory to another one. It is also interesting to ask if this process would end, namely are there theories whose quivers are maximal and do not admit any additional nodes, subject to mutation-finiteness. The aim of this section is to analyze these questions.

As already noted, mutation-finite quivers have at most two arrows between any pairs of nodes. The double arrows of a finite-mutation quiver have a simple physical interpretation. In section 2 we considered the example of $SU(2)$ SYM coupled to N_f fundamental flavors. Its quiver, see figure (2.6), has a double arrow subquiver $\bullet \rightrightarrows \bullet$ (*a.k.a.* the Kronecker quiver), corresponding to the $SU(2)$ gauge sector, which is coupled by pairs of single arrows to each flavor node (which represents a fundamental hypermultiplet). The single arrows form together with the double one an oriented triangle, and stand for the gauge coupling of the SYM sector to the matter one. In section 2 we saw how this particular arrangement of arrows precisely corresponds to the physics of the gauge couplings.

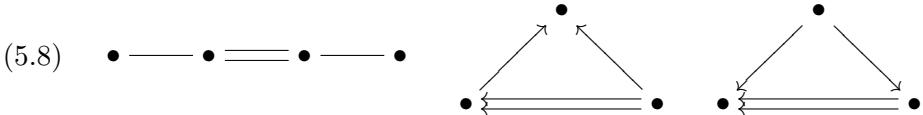
As already noted, a subquiver can be viewed as a subsector of the theory. In particular we can go to a point in moduli space where we have only the $SU(2)$ gauge theory degrees of freedom. On the other hand we could look at the couplings of the Kronecker subquiver $\bullet \rightrightarrows \bullet$ which represents a pair of dual electric/magnetic charges of an $SU(2)$ gauge sector, to the rest of the quiver and interpret this as the coupling of the $SU(2)$ gauge sector to the rest of the system. This can naturally be interpreted as saying that the

rest of the quiver has an $SU(2)$ gauge symmetry which is being gauged. We now discuss some general aspects of such couplings.

Let us ask then how the Kronecker quiver can be connected to the rest of the quiver. It turns out that generically quivers cannot have *overlapping* Kronecker subquivers; more precisely, if a mutation–finite quiver Q has a subquiver of the form¹⁰



then Q must be the Markov quiver (2.5), and we have the $\mathcal{N} = 2^*$ theory [4]. Thus other than this case, the Kronecker quivers are connected to the rest of the quiver only by single arrows. Consider then another node connected to the Kronecker quiver. It is either connected to both nodes of the Kronecker quiver or just to one. Note, however, that the following quivers



are not mutation–finite, and hence cannot appear as subquivers of finite–mutation quivers. Hence a Kronecker subquiver \mathbf{Kr} of a quiver Q which corresponds to a complete $\mathcal{N} = 2$ theory is attached to the rest of the quiver Q through oriented triangles, so that, locally around the double–arrow, the quiver looks like that of $SU(2)$ with N_f flavors (see figure (2.6)), where N_f is the number of oriented triangles in Q which have \mathbf{Kr} as a side.

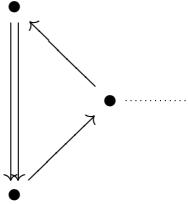
The quiver (2.6) is not of mutation–finite type for $N_f \geq 5$; this corresponds to the fact that the corresponding gauge theory is not UV complete having a Landau pole. For $N_f = 4$ the quiver (2.6) is of mutation–finite type, but no (connected) finite–mutation quiver may have it as a proper subquiver. Physically, this corresponds to the fact that $SU(2)$ with four flavor is conformal, and coupling extra matter makes the gauge beta function positive, losing UV completeness. Therefore

Kronecker Coupling: *Let Q be a quiver with a double–arrow describing a complete $\mathcal{N} = 2$ theory which is not pure $SU(2)$, $\mathcal{N} = 2^*$ $SU(2)$, or $SU(2)$ with $N_f = 4$. Then, locally near the double–arrow, Q has one of the following*

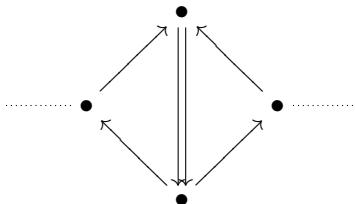
¹⁰ Here and below we use the following convention: Graphs with *unoriented* edges stand for the *full* family of quivers obtained by giving arbitrary orientations to the arrows.

three subquivers

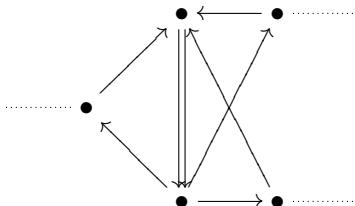
(5.9)



(5.10)



(5.11)



dashed lines standing for arrows connecting the subquiver to the rest of the quiver Q .

The above situation is naturally interpreted as the coupling of the $SU(2)$ SYM represented by the Kronecker subquiver **Kr** to, respectively, one, two, or three $\mathcal{N} = 2$ systems represented by the subquivers $\bullet \dots \dots$. The simplest instance is when these subquivers are just a node, \bullet , in which case we get $SU(2)$ SQCD with $N_f = 1, 2$ and 3 , respectively. We stress that, in general, the subquiver $\mathcal{N} = 2$ systems are coupled together also by other interactions, corresponding to the arrows connecting them in the full quiver Q . A specially simple case is when the elimination of the Kronecker subquiver **Kr** disconnects Q into a maximal number of ‘elementary’ components $\bullet \dots \dots$.

The allowed subquivers $\bullet \dots \dots$ are severely restricted by the mutation-finite condition for Q . As in the example of $SU(2)$ coupled to N_f flavors, this condition is physically interpreted as the UV completeness requirement that the beta-function of the $SU(2)$ is non-positive: hence the sum of the contribution to the beta function from the $\mathcal{N} = 2$ system(s) represented by the $\bullet \dots \dots$ subquivers should be less or equal to the contribution of 4 hypermultiplets in the fundamental representation. This observation will allow us to determine the contribution to the gauge beta function of

all the possible (complete) $\mathcal{N} = 2$ systems $\bullet \dots \bullet$ (which may have no Lagrangian description, in general).

Example. From figure 1 we see that the elliptic $\widehat{\tilde{E}}_r$ quivers correspond to a ‘weak coupling’ regime of the corresponding complete $\mathcal{N} = 2$ theories look as an $SU(2)$ SYM coupled to three decoupled $\mathcal{N} = 2$ systems. For $r = 7, 8$, one $\mathcal{N} = 2$ system (corresponding to the node in the left side of the figure) is an ordinary hypermultiplet. In section 7 we shall show that the $\widehat{\tilde{E}}_r$ theories have also strongly coupled regimes in which the spectrum consists only in a finite set of BPS hypermultiplets.

It is natural to ask how many Kronecker sub-quivers does a quiver have, and how this changes as the quiver undergoes mutation. In fact, in a typical mutation-class, most of the quivers have only single-arrows; very few quivers have the maximal number of double-arrows allowed for that class; for instance, for complete $\mathcal{N} = 2$ models which are quiver gauge theories and for which the matter fields can be massed up (*i.e.* all the mass terms are consistent with gauge symmetry), there is a *unique* BPS quiver with the maximal number of 2-arrows equal to the number of $SU(2)$ gauge groups. We stress that, in the general case, there is *no* one-to-one correspondence between $SU(2)$ gauge groups and Kronecker subquivers. Even if we take a quiver in the mutation-class with the maximal number of double-arrows, this may be still less than the actual number of $SU(2)$ gauge groups. This happens when we have several $SU(2)$ gauge sectors coupled together by half-hypermultiplets rather than full hypermultiplets, such that the half-hypermultiplets transform as different representation of the $SU(2)$ gauge groups and cannot receive mass. In such a case we cannot expect to isolate the pure $SU(2)$ theory, and so we do not expect to have a corresponding Kronecker subquiver.

As already noted, in principle we can add additional nodes and arrows to a given mutation-finite quiver and still keep it mutation-finite. This raises the question of whether there are maximal mutation-finite quivers for which we cannot add additional nodes maintaining this property, and their interpretation if they exist. We shall argue later in section 8) that:

The graphical conformality criterion: *A complete $\mathcal{N} = 2$ theory is UV conformal (as contrasted to asymptotically free) if and only if its normalized quiver is either a maximal mutation-finite one, or a vector-less quiver.*

By a *maximal mutation-finite* quiver we mean a quiver which is mutation-finite and not a proper subquiver of any connected mutation-finite quiver. Two basic examples of maximal mutation-finite quivers are the Markov quiver (2.5), corresponding to $SU(2)$ $\mathcal{N} = 2^*$, and the $SU(2)$ $N_f = 4$ quiver (2.6). By a *vector-less* quiver we mean a quiver such that

no quiver in its mutation class contains multiple arrows; such quivers correspond to $\mathcal{N} = 2$ theories having no phase which looks like SYM (with any gauge group G) coupled to some additional matter. In particular, vector-less quiver $\mathcal{N} = 2$ theories have no BPS chambers with *charged* BPS vector multiplets.

To complete the classification of conformally complete $\mathcal{N} = 2$ theories, we need to classify the *vector-less* quivers. Clearly a finite quiver such that all its mutations contain only simple arrows is, in particular, mutation-finite and must be in the Felikson–Shapiro–Tumarkin list. By inspection, the only classes with this property in the eleven exceptional cases are the three finite Dynkin diagrams E_6, E_7, E_8 . Likewise, going through the classification of the quivers associated to triangulated surfaces, we see that this property is true only if \mathcal{C} is the disk with zero or one puncture whose quivers are, respectively, the (finite) Dynkin diagrams of types A and D . Hence, the only $\mathcal{N} = 2$ theories with the properties that all quivers in their mutation classes have only single-lines are the *ADE* Argyres–Douglas ones (already studied in [2]). They are UV conformal.¹¹ Note that these are precisely the class that map to the 2d theories which are minimal, in the sense that they have a UV limit corresponding to minimal $\mathcal{N} = 2$ conformal theories in 2d (which in particular have $\hat{c} < 1$).

We end this subsection with a remark. The way mutation-finite quivers are classified in the math literature is by studying the maximal ones; once we have identified a maximal mutation-finite quiver, we may rule out all quivers containing it, and, by doing this systematically, we may eliminate all non-mutation-finite ones. Physically, this means that we keep adding ‘matter’ to the $SU(2)^k$ theory until the UV beta functions of all gauge couplings are negative. When we reach a conformal theory we stop, since adding further ‘matter’ will result in a UV incomplete theory. The corresponding quiver is automatically maximal, and we can forget about all quivers containing it. This gives us another way of understanding the correspondence

$$\text{mutation-finite quivers} \longleftrightarrow \text{complete } \mathcal{N} = 2 \text{ theories.}$$

6. Identification of a large class of quivers of finite mutation type as generalized Gaiotto theories

According to the discussion in §.5, to each mutation-finite class of 2-acyclic quivers there should correspond a complete $\mathcal{N} = 2$ theory in four dimension. To make this correspondence more explicit, in the following two sections we identify the supersymmetric theory associated to each mutation-finite class of quivers.

¹¹By Gabriel theorem [15–17, 41], these are in one-to-one correspondence with the finite-representation hereditary algebras. This is another confirmation of the deep connection between quiver representation theory and $\mathcal{N} = 2$ theories.

The quivers (with at least three nodes) which belong to all, but eleven, mutation-finite classes are adjacency matrices of ideal triangulations of some bordered surface. Therefore we divide the identification process into two steps: First we identify the theories corresponding to the infinite set of quiver classes arising from bordered surfaces \mathcal{C} , and then consider the residual eleven exceptional classes one by one.

The $\mathcal{N} = 2$ models corresponding to the non-exceptional quivers turn out to be generalizations of the $SU(2)$ theories recently studied by Gaiotto [34]. The existence of these more general theories already follows from the constructions in sections 3, 8 of [24].

More precisely, as we shall show momentarily, all the non-exceptional complete $\mathcal{N} = 2$ theories may be engineered by compactifying the A_1 six dimensional $(2, 0)$ theory on a curve \mathcal{C} of genus g and $n + b$ punctures supplemented with some particular boundary conditions at these punctures. The resulting four dimensional theory will preserve eight supercharges iff the internal $2d$ fields on \mathcal{C} , (A, ϕ) , satisfy the Hitchin equations [24, 34]

$$(6.1) \quad F + [\phi, \bar{\phi}] = 0$$

$$(6.2) \quad \bar{\partial}\phi = 0,$$

with prescribed singularities at the $n + b$ punctures. The conditions on (A, ϕ) are better stated in terms of the spectral cover $\Sigma \rightarrow \mathcal{C}$ of the Hitchin system (6.1)(6.2). Σ , which is the Seiberg–Witten IR curve of the resulting $4d \mathcal{N} = 2$ theory [24, 34], is the curve in the total space of the cotangent bundle $T^*\mathcal{C}$ defined by the spectral equation¹²

$$(6.3) \quad \det[y - \phi] \equiv y^2 - \phi_2 = 0.$$

The meromorphic quadratic differential ϕ_2 is required to have (for generic points in the Coulomb branch and values of the parameters) double poles at the ordinary punctures and poles of order $p_i = c_i + 2 \geq 3$ at the puncture representing the i -th boundary component having c_i marked points (section 8 of [24]). We may think of ordinary punctures as boundary components without marked points. This is because the quadratic differential $(dz/z)^2$ can be written as $(dw)^2$ where $w = \log z$, and w parameterizes a cylinder. When needed, we replace punctures with higher order poles of ϕ_2 with small circles with $p_i - 2$ marked points to reproduce their topological description.

The class of theories studied by Gaiotto in [34] corresponds to the special case of this construction in which all punctures are just ordinary double poles. This Gaiotto subset consists of models which are superconformal in the limit of zero masses (and Coulomb branch parameters). On the contrary, the general theory associated to a surface ‘with boundaries’ — that is, specified by a quadratic differential ϕ_2 with prescribed higher order poles — are *not* conformal in the UV but just asymptotically free (AF). The

¹² y is a coordinate along the fiber of $T^*\mathcal{C}$. The canonical differential $y dx$ is identified with the Seiberg–Witten one.

simplest examples [24] of such AF models are the well-known $SU(2)$ gauge theories with $N_f = 0, 1, 2, 3$ fundamental flavors; these theories may also be engineered in the present framework by considering a sphere with two or three punctures having pole orders

N_f	# punctures	order of poles	quiver class
0	2	3, 3	$\widehat{A}_1(1, 1)$
1	2	3, 4	$\widehat{A}_2(2, 1)$
2	2	4, 4	$\widehat{A}_3(2, 2)$
2	3	2, 2, 3	“ \widehat{D}_3 ” $\equiv \widehat{A}_3(1, 1)$
3	3	2, 2, 4	\widehat{D}_4

(the $N_f = 2$ model has two different, but physically equivalent, realizations in terms of a system of M -branes; in terms of the $6d$ $A_1(2, 0)$ theory [24] they correspond to the two surfaces listed in the table; at the quiver level the identity of the two theories expresses the well-known Lie algebra isomorphism $\widehat{SU(4)} \simeq \widehat{SO(6)}$).

The identification of the complete $\mathcal{N} = 2$ theories which are UV superconformal is presented in section 8, and agrees with the graphical rule of sect. 5.3.

It should be stressed, however, that the correspondences finite-mutation quiver \leftrightarrow triangulated surface \mathcal{C} \leftrightarrow Gaiotto $\mathcal{N} = 2$ theory require the surface \mathcal{C} to have at least one puncture to base the triangulation. In ref. [34] $\mathcal{N} = 2$ models are constructed also for genus $g > 1$ surfaces *without* punctures. With the exception of the $g = 2$ case (to be discussed in section 7 below), there are no additional mutation-finite quivers to be assigned to these puncture-less theories, given that the theories with at least one puncture already exhaust the full supply of finite-mutation quivers with more than 10 nodes. Moreover, the no-puncture $g \geq 3$ theories cannot be equivalent to some other model with punctures already in the classification, since *i*) they are conformal, *ii*) have no flavor charge, *iii*) have rank $\Gamma = 6g - 6 \geq 12$, and there are no mutation-finite quivers with these three properties. The solution of the puzzle is that these theories, like $\mathcal{N} = 4$, are not quiver theories, in the sense that there are no D -tuple of charge vectors $\gamma_a \in \Gamma$ such that all BPS charge vectors may be written as $\pm \sum_a n_a \gamma_a$ with positive n_a 's; for these theories the BPS phase are dense on the unit circle, and thus they do not admit a BPS quiver.

6.1. $4d/2d$ correspondence and ideal triangulations. The identification of the non-exceptional $\mathcal{N} = 2$ complete theories with the generalized Gaiotto theories, is confirmed by the $4d/2d$ correspondence of ref. [2], reviewed in section 4.

Roughly speaking, the $4d/2d$ correspondence says that the quiver of the $4d$ theory is to be identified with (minus) the BPS quiver of the

corresponding $2d$ $(2, 2)$ theory. At the technical level, things are a bit more involved because of some subtleties with the signs (*i.e.* arrow orientations) discussed in [1]. Moreover, as stressed in [2], the classification of $2d$ BPS quivers (modulo $2d$ wall-crossing [1]) is *coarser* than the classification of $4d$ quivers (modulo mutation-equivalence) because there are more $2d$ walls to cross than quiver mutations. A more precise dictionary is the following: let

$$(6.4) \quad S = \prod_{\text{half plane}}^{\curvearrowleft} \exp(-\mu_\theta)$$

be the product of all $SL(D, \mathbb{Z})$ monodromy group elements associated to BPS states with phase θ in the given half-plane¹³. S is related to the monodromy M by the formula $M = (S^{-1})^t S$ [1]. By the $2d$ wall-crossing formula, S is invariant under all wall-crossing except those which make a BPS state to exit from the given half-plane (while its PCT conjugate enters from the other side). Then S is defined up to the same mutations as B , except that S depends also on the sign conventions of the $2d$ vacua (changing the sign of the k -th vacuum makes $\mu_{kj} \rightarrow -\mu_{kj}$). Therefore, the refined statement is that we may choose the $2d$ conventions in such a way that the exchange matrix of the $4d$ quiver is

$$(6.5) \quad B = S - S^t.$$

In the case of complete theories, the corresponding $2d$ models are also complete in the same sense, and we may always reduce ourselves to a convex arrangement of vacua [1], in which case we have simply $B = -\mu$, and we may forget about subtleties (at the price of wall-crossing μ to a suitable $2d$ BPS chamber).

6.1.1. Lagrangian A -branes as ideal triangulations. We would like to identify the corresponding $2d$ theory associated with the $4d$ theory obtained by 2 5-branes wrapping a Riemann surface with punctures. We already know that if we have a type IIB geometry of the form

$$uv - W(y, z) = 0$$

The associated $2d$ theory is a LG theory with superpotential $W(y, z)$ as a function of chiral fields y, z . On the other hand it is also known that this type IIB geometry is dual to a 5-brane as a subspace of y, z parameterizing \mathbb{C}^2 given by wrapping the curve

$$W(y, z) = 0$$

and filling the spacetime [45]. Now let us consider the Gaiotto theories. In this case the 5-brane geometry is captured by the geometry

$$W(y, z) = y^2 - \phi_2(z)$$

¹³ $|(\mu_\theta)_{ij}|$ is equal to the number of BPS solitons connecting the i and j vacua and having BPS phase θ ; the sign of $(\mu_\theta)_{ij} = -(\mu_\theta)_{ji}$ follows, up to convention dependent choices, from the rules of ref. [1].

However, here y has a non-trivial geometry: y is a section of the canonical line bundle on the Riemann surface. To make y be ordinary coordinate we take a reference quadratic differential ω_0 , and define

$$\tilde{y} = y/\omega_0.$$

Under this transformation we get the equation

$$W(\tilde{y}, z) = \tilde{y}^2 - \frac{\phi_2(z)}{\omega_0}$$

Since the \tilde{y}^2 term does not affect the BPS structure and vacua of the theory, this is equivalent to a $2d$ theory with $(2, 2)$ supersymmetry and superpotential

$$(6.6) \quad W(z) = \frac{\phi_2(z)}{\omega_0(z)}.$$

The meromorphic one-form dW has a number of zeros (\equiv supersymmetric vacua)

$$(6.7) \quad \#\{\text{zeros of } dW\} = 2g - 2 + \text{polar degree of } dW = 6g - 6 + \sum_i (p_i + 1),$$

where p_i is the order of pole of ϕ_2 at the i -th puncture. Thus the number of supersymmetric vacua of the two dimensional theory is equal to D , the number of arcs in an ideal triangulation of the corresponding bordered surface. This is no coincidence: let us consider the Lagrangian A -branes defined, for this class of $(2, 2)$ theories, in [36]. They are the integral curves γ_i of the differential equation

$$(6.8) \quad \text{Im}(e^{i\theta} dW) = 0$$

(for some *fixed* but generic value of the angle θ) which start at $t = 0$ from the i -th zero of dW , X_i , and approach, as $t \rightarrow \pm\infty$, infinity in the W -plane — that is, a puncture in \mathcal{C} — along a direction such that

$$(6.9) \quad \text{Re}(e^{i\theta} W) \Big|_{t \rightarrow \pm\infty} \rightarrow +\infty.$$

We assume θ to have been chosen so that $\text{Im}(e^{i\theta} W(X_i)) \neq \text{Im}(e^{i\theta} W(X_j))$ for $i \neq j$. Then the branes γ_i are distinct.

If two arcs, γ_i, γ_j , cross at some finite value of t , they coincide everywhere $\gamma_i \equiv \gamma_j$. Hence the arcs γ_i do not cross themselves nor each other, except at the punctures. This is one of the properties defining the collection of arcs of an ideal triangulation [3]. To be a compatible collection of arcs on \mathcal{C} , the Lagrangian A -branes $\{\gamma_i\}$ should also be non-contractible to a puncture (or boundary arc) and pairwise isotopy inequivalent. If these properties hold, the Lagrangian branes $\{\gamma_i\}$ form automatically a maximal collection, and hence an ideal triangulation, since their number is the maximal one, eqn.(6.7). In ref. [36] it was shown that the Lagrangian A -branes $\{\gamma_i\}$ span the relative homology group $H_1(\mathcal{C}, \mathcal{B})$ (where $\mathcal{B} \subset \mathcal{C}$ is the region near the punctures

where $\text{Re}(e^{i\theta} W) \gg 1$, so all the axioms for an ideal triangulation are satisfied.

The above construction should be contrasted with the similar, but different, triangulation which arises in the study of 4d BPS states by considering straight lines on the SW curve, defined by the condition that the phase of the SW differential does not change along the path introduced in [46] and [26] and studied extensively in [24]. There, for the same class of models, one constructs an ideal triangulation using the integral curves of the (real part of) the Seiberg–Witten differential, namely the solutions to the equation

$$(6.10) \quad \text{Im}(e^{i\theta/2} y dz) = 0$$

instead of the one in eqn.(6.8). Again, one gets an ideal triangulation, but this time the ‘vacua’, that is the zeros of the Seiberg–Witten differential, are in one-to-one correspondence with the faces of the triangulation, rather than with the arcs. As a check, let us count the number of triangles

$$(6.11) \quad \begin{aligned} \# \text{ triangles} &= 2 - 2g + \# \text{ arcs} - \# \text{ punctures} \\ &= 4g - 4 + \sum_i p_i \equiv \# \text{ zeros of } \phi_2. \end{aligned}$$

The adjacency quivers obtained by these two procedures, corresponding to ideal triangulations of the same punctured surface, should be the same up to mutation equivalence. This is the underlying reason why the 2d BPS quiver is (up to natural equivalences) the same as the 4d Dirac quiver.

Before going to the adjacency quivers, let us illustrate in an example how the A -brane ideal triangulation works.

6.1.2. Example: torus with n ordinary punctures. We start with the torus with one puncture, which corresponds to $\mathcal{N} = 2^*$ and the Markov quiver (2.5). There is an essentially unique ideal triangulation



where the opposite sides of the rectangle are identified. The corresponding incidence matrix is

$$(6.13) \quad B_{1,2} = -2, \quad B_{1,3} = 2, \quad B_{2,3} = -2$$

giving the Markov quiver (2.5).

To recover this result from the 2d perspective, we go the universal cover of \mathcal{C} , namely \mathbb{C} , and consider the LG model with superpotential $W(X) = i\wp(X)$ taking, to have \mathbb{Z}_4 symmetry, a square torus of periods

$(1, i)$ so that

$$(6.14) \quad (\phi')^2 = 4\phi^3 - \frac{\Gamma(1/4)^8}{16\pi^2}\phi.$$

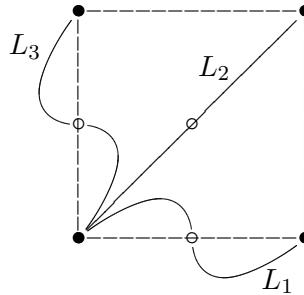
One has $\phi(iX) = -\phi(X)$ and $\phi(X) = X^{-2}f(X^4)$ with $f(\bar{z}) = \overline{f(z)}$. Viewing the torus as a double cover of the ploane given by $Z = 2\phi(X)$, and the 2-fold cover by $Y = \phi'$ we have

$$Y^2 = Z^3 - aZ$$

The three classical vacua correspond to the three solutions of $Y = \phi'(X_k) = 0$ at finite Z , and are at the half-lattice points

$$\begin{aligned} X_1 &= \frac{1}{2}, & X_2 &= \frac{1+i}{2}, & X_3 &= \frac{i}{2} \\ W(X_1) &= i \frac{\Gamma(1/4)^4}{8\pi}, & W(X_2) &= 0, & W(X_3) &= -i \frac{\Gamma(1/4)^4}{8\pi}. \end{aligned}$$

The Lagrangian branes map to straight lines on the W plane which in this case correspond to Z plane. There are a pair of kinks between any pair the three vacua corresponding to the two straight lines which connect them in the Z -plane. The Lagrangian brane L_2 passing through the \mathbb{Z}_4 invariant point X_2 and going to $\text{Re}(W) = +\infty$ is just the diagonal of the square along the bisectrix of the first/third quadrants. Then the two Lagrangian branes $L_{1,3}$ passing through $X_{1,3}$ should correspond to the two S -shaped curves in the figure (their curvature is exaggerated for drawing purposes)



comparing with eqn.(6.12) we see that the three Lagrangian branes L_i are (up to isotopy) the same as the ideal triangulation arcs.

The Landau–Ginzburg model with $W(X) = i\phi(X)$ was solved in ref. [47] (it corresponds to the three-point functions of the Ising model). It has two BPS states connecting each pair of vacua related by the symmetry $X(t) \leftrightarrow -X(t)$ (modulo periods) which fixes the three classical vacua. The S matrix is

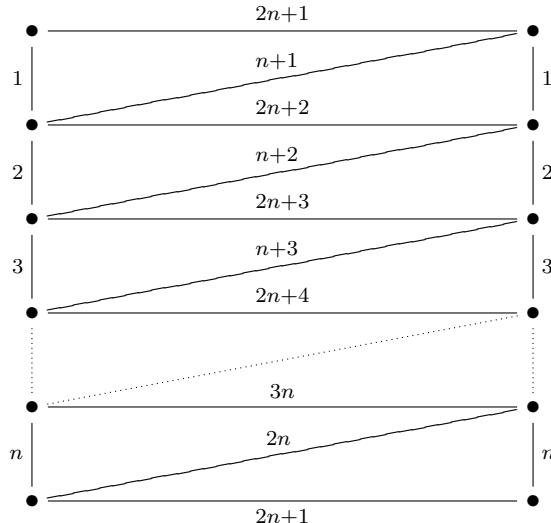
$$(6.15) \quad S = \begin{pmatrix} 1 & -2 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix}$$

(the eigenvalues of $M = (S^{-1})^t S$ are $-1, 1, -1$) and

$$(6.16) \quad B = S - S^t = \begin{pmatrix} 0 & -2 & 2 \\ 2 & 0 & -2 \\ -2 & 2 & 0 \end{pmatrix}$$

which is the exchange matrix of the Markov quiver (2.5).

A torus with $n > 1$ punctures has many different ideal triangulations. The one with the more transparent physical interpretation has the adjacency quiver with maximal number of double-arrows (Kronecker subquivers), namely n . This triangulation is the *zig-zag* one (*a.k.a.* the *snake* triangulation): See the figure



where corresponding segments of the sides should be identified (in the figure, identified segments carry the same label). The arc labelled k , with $1 \leq k \leq n$ shares *two* triangles with the arc labelled $n+k$. The two triangles have the same orientation, so the corresponding entries of the adjacency matrix are

$$(6.17) \quad B_{k,n+k} = -2, \quad k = 1, 2, \dots, n.$$

On the other hand, the k -th arc shares a single triangle with the arcs $2n+k$ and $2n+k+1$. One has

$$(6.18) \quad B_{k,2n+k} = +1$$

$$(6.19) \quad B_{k,2n+k+1} = +1.$$

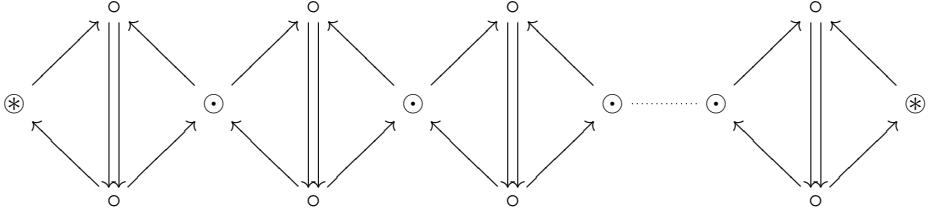
Finally, the arc $n+k$ shares a triangle with the arcs $2n+k$ and $2n+k+1$. Then

$$(6.20) \quad B_{n+k,2n+k} = -1$$

$$(6.21) \quad B_{n+k,2n+k+1} = -1.$$

All other entries of the adjacency matrix vanish. In particular, we have n double arrows, as anticipated. All triangles in the quiver are oriented. According to our discussion in section 5.3 these quivers correspond to a closed chain of n Kronecker subquivers (*i.e.* $SU(2)$ gauge groups) coupled to each other by bi-fundamental hypermultiplets (represented by the nodes \circ , \circledast on the figure)

(6.22)

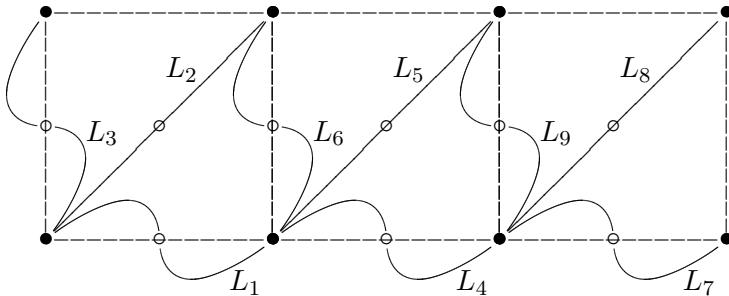


where the two bi-fundamentals denoted by the symbol \circledast should be identified. Thus, these $\mathcal{N} = 2$ models correspond to quiver $SU(2)$ gauge theories with underlying graph the affine Dynkin diagram \widehat{A}_{n-1} , as expected for the Gaiotto theory engineered by a torus with n -punctures. Notice that by the topological theorem (5.6) this $\mathcal{N} = 2$ model has precisely n flavor charges, corresponding to the n bi-fundamentals.

The above snake triangulation may be easily recovered from the two-dimensional point of view. One consider the same Landau–Ginzburg model with Weiertrass superpotential as before, except that we now identity the field X up to multiple periods

(6.23)
$$X \sim X + a n + b i, \text{ where } a, b \in \mathbb{Z},$$

so that now we have $3n$ distinct vacua and hence $3n$ distinct A –branes which are just the translation by $k = 0, 1, 2, \dots, n - 1$ of the basic ones for $n = 1$. The case $n = 3$ is represented in the figure



From the figure it is clear that the A –branes L_1, \dots, L_{3n} give precisely the snake triangulation.

Again, the adjacency quiver of the triangulation may be read from the $2d$ BPS spectrum. Between vacua $X = 1/2 + k$ and $X = 1/2(\tau + 1) + k$, $k = 0, 1, 2, \dots, n - 1$, we have still two solitons, going opposite way along the

B -cycle, but the vacuum at $\tau/2 + k$ is connected to the vacua $1/2 + (k-1)$, $1/2 + k$, $(\tau+1)/2 + (k-1)$ and $(\tau+1) + k$ by just one BPS soliton. *E.g.* for $n = 2$ the S matrix is

$$(6.24) \quad S = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

(eigenvalues¹⁴ of M : $-1, -1, 1, 1, -1, -1$) and $B = S - S^t$ is precisely the exchange matrix of the quiver (6.22) for $n = 2$.

6.1.3. *Adjacency matrix vs. 2d BPS spectrum.* In the above examples we saw that the adjacency quiver of the triangulation is given by the BPS quiver of the 2d $(2, 2)$ system whose A -branes triangulate the surface \mathcal{C} , in agreement with the basic idea of the $4d/2d$ correspondence. The examples discussed so far correspond to simple situations where certain sign subtleties play no role. The equality will be verified in many examples below, including some non-trivial cases, as the one discussed in detail in appendix B.3, where the subtleties of two-dimensional physics do play a significant role.

Let us consider the situation where $\mathcal{C} = \mathbb{C}$ (*i.e.* a sphere with a pole of order p at $z = \infty$), the exchange matrix B_{ij} of the 2d BPS quiver is given by the intersection number of the corresponding arcs (up to mutations) in the $\{\gamma_i\}$ ideal triangulation, see ref. [36]

$$(6.25) \quad B_{ij} = \pm \gamma_i \cdot \gamma_j.$$

Since the A -branes cross only at infinity, to get the correct counting of the intersection number one has to resolve the puncture by replacing it with a small circle with $p - 2$ marked points, as required to interpret the family $\{\gamma_i\}$ as an ideal triangulation. Then the intersection $\gamma_i \cdot \gamma_j$ is given by the signed sum of ± 1 over all triangles with sides γ_i, γ_j . In the case $\mathcal{C} = \mathbb{C}$, or, topologically, a disk with $p - 2$ marked points on the boundary, the quiver with exchange matrix $\gamma_i \cdot \gamma_j$ is, by the Milnor fiber theorem [48], given by the A_{p-3} Dynkin quiver (up to equivalence), which is the same as the adjacency quiver of the disk with p marked points [3].

In the general case, the intersection $\gamma_i \cdot \gamma_j$ again is concentrated at the poles, which, if irregular, must be resolved into boundary components. Locally, the situation is as in the previous case, and the counting still apply. It remains, however, the problems of specifying the signs (6.25) which are not determined at this level of analysis (except for the requirement that they must be compatible with the mutation-finiteness). There are two sources of signs: the classical sign of the A -brane curve, and the quantum sign

¹⁴ In general, the monodromy M for the n -punctured torus is equal, up to conjugacy, to the direct sum of n copies of the $n = 1$ monodromy.

given by the sign of the determinants in the quantization around that configuration. The methods of ref. [1], are very convenient to fix the signs (up to conventional choices) and in all examples we analyzed we get quivers consistent with the 4d/2d correspondence.

The identification of the 4d BPS quiver of a generalized Gaiotto theory with the topological adjacency quiver of an ideal triangulation of the corresponding bordered surface has a few immediate payoffs.

First of all, it follows from the above correspondence that any mutation invariant of the Dirac pairing matrix, B_{ij} is also a chamber-independent property of the four dimensional $\mathcal{N} = 2$ theory. The simplest such invariant is the *corank* of the matrix B_{ij} , that is the number of independent charge vectors $v \in \Gamma$ which have vanishing Dirac pairing with all the charges in the theory. Physically, such vectors should be seen as *flavor charges*, whereas the ones having non-trivial Dirac pairings have electric/magnetic nature. We shall, therefore, refer to the corank of B as the *number of flavor charges*. For quivers arising from triangulation of surfaces, the number of flavor charges is given by the number of punctures where ϕ_2 is allowed to have poles of *even* order [3] (in particular, all ordinary double poles will contribute).

This result may also be understood in terms of the geometry of the Seiberg–Witten curve Σ . Since Σ is a double cover of \mathcal{C} its genus is given by

$$(6.26) \quad g(\Sigma) = 2g - 1 + \frac{1}{2} n_B,$$

where n_B is the number of branch points. The branch points are given by *i)* the zeros of ϕ_2 (there are $4g - 4 + \sum_i p_i$ of them), *ii)* the poles of ϕ_2 of odd order. Then

$$(6.27) \quad g(\Sigma) = 4g - 3 + \frac{1}{2} \sum_{p_i \text{ even}} p_i + \frac{1}{2} \sum_{p_i \text{ odd}} (p_i + 1).$$

$g(\Sigma)$ is the number of linearly independent holomorphic one forms on Σ ; however, g of them are just pull-backs of holomorphic forms on the Gaiotto curve \mathcal{C} . These are even under the cover group \mathbb{Z}_2 , while the remaining $g(\Sigma) - g$ are odd. Dually, the number of odd 1-cycles is $2g(\Sigma) - 2g$. Given that the canonical one-form, $y dx$, is \mathbb{Z}_2 odd, we get that the total number of electric *and* magnetic charges is

$$(6.28) \quad 2g(\Sigma) - 2g = 6g - 6 + \sum_{p_i \text{ even}} p_i + \sum_{p_i \text{ odd}} (p_i + 1) = \text{rank } B,$$

as predicted by the Dirac quiver/triangulation quiver identification.

The second obvious pay-off is a very convenient way of constructing (and understanding) complicated theories in terms of simpler ones. Indeed, having related a large class of $\mathcal{N} = 2$ theories to surfaces with punctures and boundaries, one can easily take two such theories, view them as two decoupled sectors of a more complicated theory, and couple them by some suitable $\mathcal{N} = 2$ supersymmetric interactions. At the geometric level, this

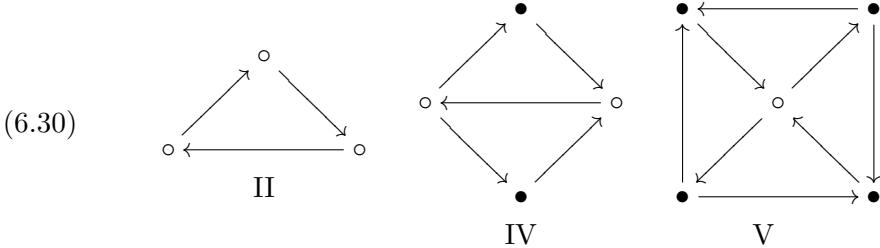
process of couplings various sub-sectors to construct a new model corresponds to surgery of triangulated surfaces. This viewpoint leads directly to simple rules for gluing together the sub-quivers associated to each sector into the quiver of the fully coupled theory. Thus one may get the quivers of complicated models without going through the triangulation process or the $4d/2d$ correspondence. There exist different kinds of surgery, corresponding to physically different ways of coupling together the various sub-sectors. The geometrical rules of triangulation guarantee that only couplings which are fully consistent at the quantum level may be realized by a sequence of these surgical operations on quivers. For complicated models, which have no regime in which all couplings are simultaneously weak, this would be hard to check directly. Surgery processes are described in detail in section 6.4 below.

6.2. Ideal triangulations vs. Gaiotto $SU(2)$ theories. We start by considering the original Gaiotto theories, namely closed surfaces with only ordinary punctures. Let \mathcal{C} be a surface of genus g with n ordinary punctures. The corresponding $\mathcal{N} = 2$ theory has a gauge group $SU(2)^{n+3g-3}$ [34], and hence a charge lattice Γ generated by $3g - 3 + n$ electric charges, $3g - 3 + n$ magnetic ones, and n flavor charges associated to the residues of $\sqrt{\phi_2}$ at the n punctures. Thus,

$$(6.29) \quad \text{rank } \Gamma = 6g + 3n - 6,$$

which is equal to the number of arcs in an ideal triangulation of the surface, and the number of nodes in its adjacency quiver.

From the description in section 5.2.1 it follows that we may simplify, for this class of surfaces, the rules to construct the adjacency quivers by gluing blocks. We may start with a collection of quiver blocks of just three kinds



and then glue them together by identifying *all* white nodes \circ in pairs, this last condition being equivalent to $\partial\mathcal{C} = \emptyset$ (*i.e.* only ordinary punctures).

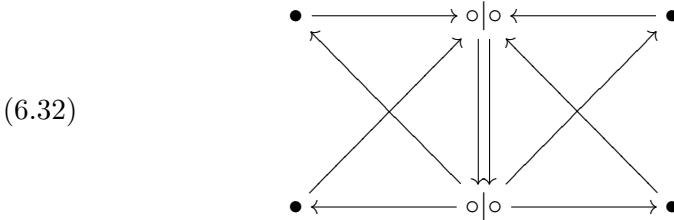
The topological invariants g and n may be read directly from the exchange matrix B of the quiver: n is just the corank f of B (= the number of flavor charges) and

$$(6.31) \quad g = \frac{D - 3f + 6}{6},$$

where D is the size of the matrix B , equal to the number of nodes in the quiver.

Now we discuss a few examples. The case of n -punctured torus was considered in section 6.1.2.

6.2.1. Example: the sphere with 4 punctures. The quiver for the sphere with four punctures, corresponding to $SU(2)$ gauge theory coupled to four flavors in the fundamental representation, is easy to construct. Just take two copies of the type IV block, and glue them together by identifying the white nodes \circ in such a way that the orientations of the arrows connecting them match. We get the quiver¹⁵



equal to (2.6) for $N_f = 4$. The underlying graph corresponds to Saito's $\widehat{\tilde{D}}_4$ elliptic root system [43].

The exchange matrix B has four zero eigenvalues: the corresponding eigenvectors are obtained by attaching a weight $1/2$ to the two white nodes, a weight 1 to any one of the black ones, and zero to the other three nodes. Then the corank of B is 4, and the quiver represents a triangulation of a surface with $(g, n) = (0, 4)$ (cfr. eqn.(6.31)).

The mutation-class of the quiver (6.32) contains four essentially distinct quivers, as it is easy to check with the help of Keller's quiver mutation Java applet [kellerappl]. The one shown in (6.32) is the one relevant in a weakly coupled chamber; it may be interpreted as the result of the coupling of four heavy electric hypermultiplets, represented by the black nodes, each carrying his own flavor charge, to the pure $SU(2)$ gauge theory, represented by the Kronecker subquiver, $\circ \longrightarrow \circ$. In this limit, the two white nodes correspond to the dyon of charge $(e_1, m_1) = (2, -1)$ and the monopole of charge $(e_2, m_2) = (0, 1)$ with Dirac pairing¹⁶

$$(6.33) \quad \langle (e_1, m_1), (e_2, m_2) \rangle \equiv e_1 m_2 - m_1 e_2 = 2.$$

¹⁵ Here and below, we use vertical bars $|$ to denote the decomposition of a quiver into its basic blocks.

¹⁶ Although the results of ref. [24] are not stated in the language of quivers, many of their findings may be rephrased in the present formalism, with full agreement, except that their discussion in section 10.7 corresponds to a quiver differing from (6.32) by the orientation of two arrows. That quiver is not of finite-mutation type, and hence does not correspond to a complete theory in our sense. It does appear, however, in another kind of finite-type classification for $\mathcal{N} = 2$ quivers, namely those which admit a chamber with a finite BPS spectrum consisting only of hypermultiplets. The quiver (6.32) does not have this last property.

According to the $4d/2d$ correspondence, the quiver (6.32) may be obtained as the BPS quiver of the $2d$ theory on the sphere with (say) the usual Fubini–Study Kähler potential, $K = -\log(1 + |Y|^2)$, and superpotential

$$(6.34) \quad W(Y) = \frac{1}{Y^2 + Y^{-2}},$$

which is symmetric under the interchange of the two poles $Y \leftrightarrow Y^{-1}$ of the sphere, as well as under $Y \leftrightarrow -Y$. One has

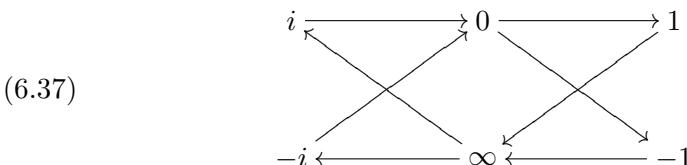
$$(6.35) \quad W'(Y) = -\frac{2Y - 2Y^{-3}}{(Y^2 + Y^{-2})^2} \equiv 2 \frac{Y^5 - Y}{(Y^4 + 1)^2}$$

From which we see that the classical vacua are the four roots of unity $Y = i^k$, the south pole $Y = 0$, and — by the $Y \leftrightarrow Y^{-1}$ symmetry — the north pole $Y = \infty$. In total, we have *six* vacua, as expected.

The critical values of the superpotential are $W = 0$ for the two polar vacua, and $W = y^2/2 \equiv \pm 1/2$ for the vacua at the roots of unity. In the W -plane all soliton are just segments along the real axis [49]. Thus the BPS equation, $W(Y) = t$, reduces to the quadratic equation in Y^2

$$(6.36) \quad (Y^2)^2 - \frac{1}{t} Y^2 + 1.$$

In the relevant interval of the real axis, $-1/2 < t < 1/2$, the discriminant is positive, and we have two real roots Y^2 . As $t \rightarrow -1/2$ both roots go to $Y^2 = -1$; analogously for $t \rightarrow 1/2$ both roots approach $Y^2 = 1$. As $t \rightarrow 0$ one solution goes to zero and one to ∞ . In conclusion, in each interval $-1/2 \leq W \leq 0$, and $0 \leq W \leq 1/2$, *both* roots of the quadratic equation in Y^2 do correspond to soliton: one going to the vacuum at the north pole, $Y^2 = \infty$, and the other to the vacuum at the south pole $Y^2 = 0$. Recalling that each solitonic solution in terms of Y^2 corresponds to *two* solutions in terms of Y related by the \mathbb{Z}_2 symmetry $Y \leftrightarrow -Y$: each starts at one of the two root-of-unity vacua sharing the given critical value (these two vacua are interchanged by \mathbb{Z}_2) and ends at one of the two polar vacua (which are \mathbb{Z}_2 invariant). Then the BPS quiver has the form (we label the vertices by the value of Y)

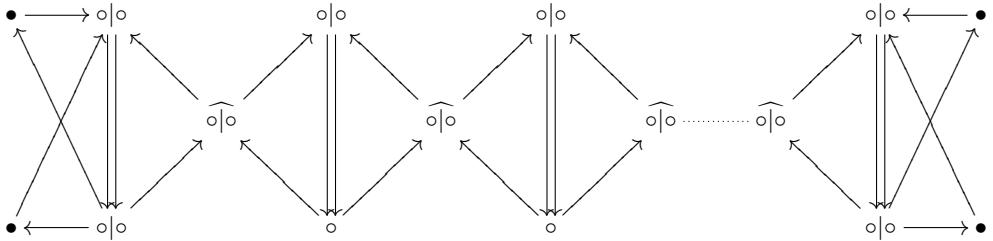


corresponding to the S matrix

$$(6.38) \quad S = \begin{pmatrix} 1 & 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 1 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(2d monodromy spectrum given by $-1, 1, 1, 1, 1, -1$, the four 1's being associated to the four flavor charges). The quiver (6.37) is mutation equivalent to (6.32).

6.2.2. Example: the sphere with $n \geq 5$ punctures. The extension to an arbitrary number $n \geq 4$ of ordinary punctures is straightforward. One takes two type IV blocks and $2(n-4)$ type II blocks and glue them together as in the figure



The incidence matrix of this quiver has n zero eigenvectors, corresponding to attaching a weight 1 to any one of the nodes \bullet or $\widehat{\circ|\circ}$, weight $1/2$ to the nodes $\circ|\circ$ connected to it by an arrow, and zero everywhere else. Since the total number of nodes is $D = 3n - 6$, from eqn.(6.31), we see that the above quiver corresponds to a surface with numerical invariants $(g, n) = (0, n)$.

The nodes \bullet and $\widehat{\circ|\circ}$ are in one-to-one correspondence with the flavor charges (*i.e.* zero eigenvectors of the incidence matrix B). Then they are interpreted as hypermultiplets carrying their own flavor charge and having electric charge -1 (that is, in the fundamental representation) with respect to each of the $SU(2)$ gauge groups (represented by the double-arrow Kronecker sub-quivers) connected to it by the arrows. Indeed, the node from which a double arrow starts/ends have charges $(e, m) = (2, -1)/(0, 1)$ with respect to the corresponding gauge group and the arrows in the figure are consistent with the Dirac pairings

$$(6.39) \quad \langle (2, -1), (-1, 0) \rangle = -1, \quad \langle (0, 1), (-1, 0) \rangle = 1.$$

The charge vectors in the kernel of B ,

$$(6.40) \quad \gamma_{\bullet_a} + \frac{1}{2} \sum_{\circ|\circ \rightleftarrows \bullet_a} \gamma_{\circ|\circ}, \quad \gamma_{\widehat{\circ|\circ}_b} + \frac{1}{2} \sum_{\circ|\circ \rightleftarrows \widehat{\circ|\circ}_b} \gamma_{\circ|\circ} \in \Gamma,$$

then correspond to purely flavor ones.

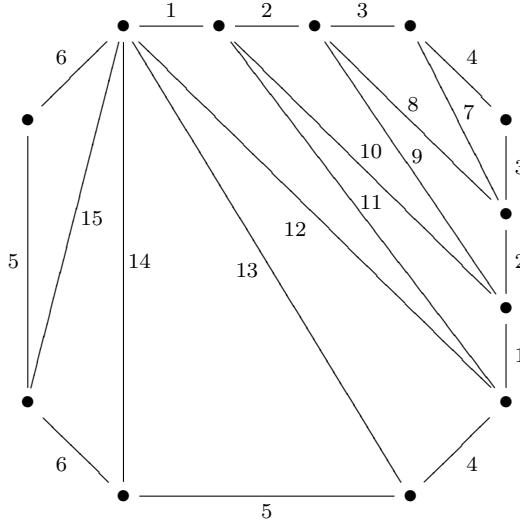


FIGURE 3. A $g = 2$ $n = 3$ ‘snake’ ideal triangulation.

Thus the nodes \bullet represent fundamental hypermultiplets, while the $\circ\bullet$ nodes stand for bi-fundamental ones. The above figure is the BPS quiver parallels the linear quiver representation of this gauge theory

$$(6.41) \quad [2] \longrightarrow (\overset{\circ}{2}) \longrightarrow (\overset{\circ}{2}) \longrightarrow (\overset{\circ}{2}) \dots \longrightarrow (\overset{\circ}{2}) \longrightarrow [2]$$

The number of distinct quivers in the mutation class of the one represented in the figure grows quite rapidly with n . The first few numbers are¹⁷

number of punctures	4	5	6	7
# of distinct mutation-equivalent quivers	4	26	191	1904

Since the theories are complete, each different quiver in the equivalence class corresponds to a physical regime of the $\mathcal{N} = 2$ theory and, in particular, to some BPS chamber. For genus zero surfaces with only ordinary double poles, one finds only one quiver in the mutation-class with the maximal number of double arrows (*i.e.* Kronecker subquivers), namely $n - 3$, which is the one we have drawn above, and which correspond to the standard regime admitting a Lagrangian description.

6.2.3. *Example: genus $g > 1$ with $n \geq 1$ punctures.* The analogue of the snake triangulation (see sect.6.1.2) for higher genus surface would be to cut open the surface to get a hyperbolic $4g$ -gon with sides pairwise identified, having care to choose one of the cuts in such a way that it goes through all the n punctures. See figure 3 for the example with $g = 2$, $n = 3$. Then one start

¹⁷ These numbers refer to the distinct quivers modulo sink/source equivalence as built in Keller’s mutation applet [42].

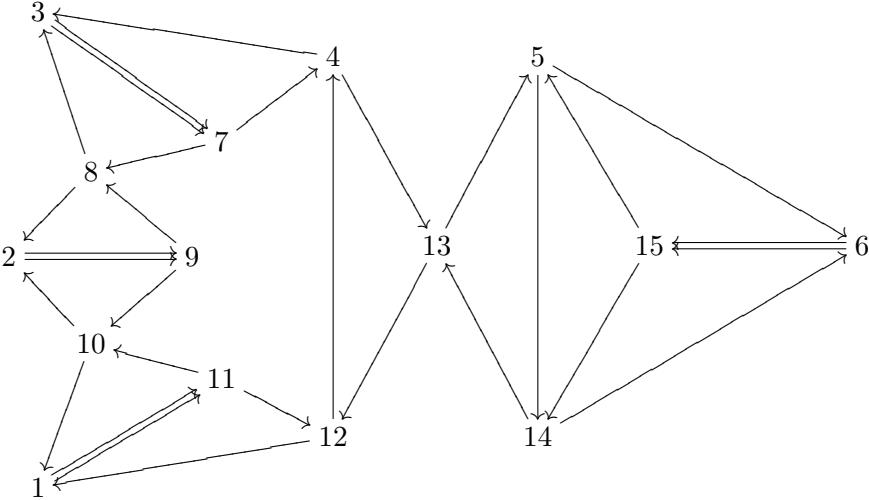


FIGURE 4. The adjacency quiver corresponding to the ideal triangulation 3 of a $g = 2$ surface with three punctures. The enumeration of the nodes corresponds to the enumeration of arcs in 3. In the left side of the quiver we see the ‘segment of a quiver $SU(2)$ theory’ associated to the three punctures.

doing the snake triangulation from the side on which the punctures lay (see the upper right corner of the figure). From that part of the triangulation we get n double arrows; for the example in the figure, they correspond to the following entries of the adjacency matrix

$$(6.42) \quad B_{3,7} = +2, \quad B_{2,9} = +2, \quad B_{1,11} = +2.$$

Then it remains to perform the triangulation of a $[4(g - 1) + 2]$ -gon with the first $4(g - 1)$ sides identified pairwise in the form $s_1, s_2, s_1, s_2, s_3, s_4, s_3, \dots$ while the last two sides are not identified. In the figure this corresponds to the part of the surface below arc 12. Let $c(g)$ be the maximal number of double arrows that we may get from such a triangulation. Then we have a triangulation with at most

$$(6.43) \quad n + c(g)$$

double arrows. It is easy to convince oneself that $c(g) = g - 1$. See figure 4 for the quiver corresponding to the ideal triangulation 3.

Therefore, for $g > 1$, the maximal number of double-arrows, $n + g - 1$, is less than the number of $SU(2)$ gauge groups, namely $n + 3g - 3$. As discussed before, this means that these theories have no chamber in which all the matter multiplets can be massed up. Indeed we will later show this is the case, by showing that there are no gauge invariant mass terms that can mass up all the matter fields. On the other hand, for $g > 1$ the quiver with

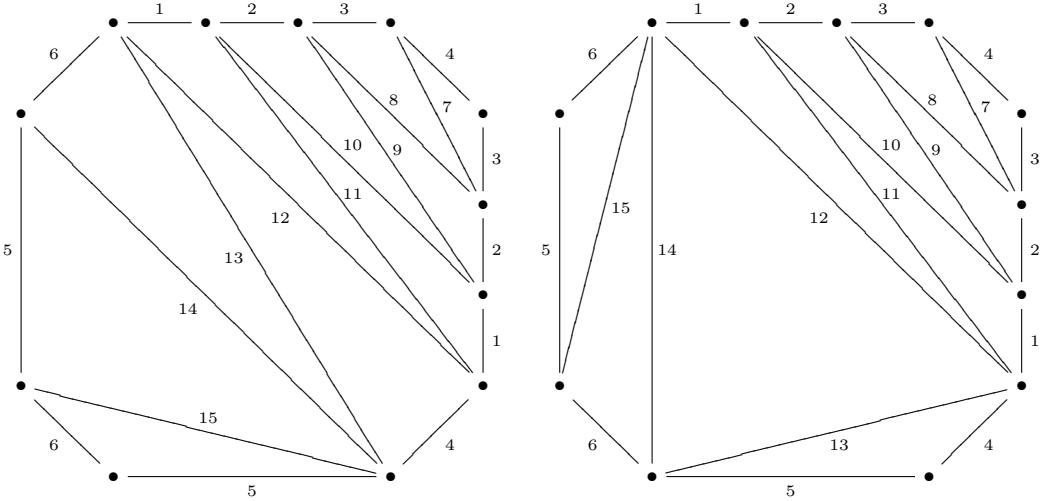


FIGURE 5. Inequivalent ‘snake’ triangulation of the same surface.

a maximal number of double arrows is not unique. For instance, for $g = 2$, $n = 3$ the triangulations in figure 5 also lead to four double arrows.

6.3. Generalized Gaiotto theories. The quivers of generalized Gaiotto theories are constructed by gluing together all five kinds of blocks, and there is no need to pair up every white node.

Generically, each quiver may be decomposed into blocks in a unique way. In this case, there is a unique bordered surface associated to the mutation-class of the quiver. There are a few exceptions to the uniqueness of the correspondence, and all these exceptions have a simple physical explanation: basically, these theories have more than one string/ M -theory engineering, and each of this realizations corresponds to a bordered surface. The quiver-mutation class, however, should be (and it is) independent of the geometrical realization. The typical example is $SU(2)$ with two flavors which has two such realizations [24, 34, 50].

6.3.1. Example: A_n and D_n Argyres–Douglas models. From the ideal triangulation point of view [fomin], The A_n , D_n models correspond, respectively, to the disk with $n + 3$ marked points on the boundary and to the punctured-disk with n marked points, that is to a sphere equipped with a quadratic differential ϕ_2 having one pole of degree $n + 5$, and, respectively, two poles of degrees 2 and $n + 2$.

These models are easily understood from the point of view of the $4d/2d$ correspondence. For the A_n series we choose a reference quadratic differential ω_0 having a pole of order 4 at infinity, while for the D_n series we pick up an ω_0 having a third order pole at infinity and a simple pole at the ordinary

puncture. We get the superpotentials:

$$(6.44) \quad A_n: \quad W(X) = X^{n+1} + \text{lower terms}$$

$$(6.45) \quad D_n: \quad W(X) = \frac{1}{X} + X^{n-1} + \text{lower terms}.$$

The first superpotential is just the usual one for the A_n minimal models [53, 54], and we know that, in some chamber, the BPS quiver is just the A_n Dynkin diagram with some orientation of the edges (which orientation being immaterial, since all orientations are mutation-equivalent for a tree quiver). This is, of course, the correct quiver for the A_n Argyres–Douglas model obtained by compactifying the Abelian six dimensional $(2, 0)$ theory on a complex curve of equation

$$(6.46) \quad y^2 = X^{n+1} + \text{lower terms},$$

unfolding the minimal A_n singularity.

On the other hand, eqn.(6.45) does not look like the usual superpotential for the D_n minimal models,

$$W(X, Z) = X^{n-1} + X Z^2 + \text{lower terms}.$$

However, to identify the BPS quiver we are free to deform the theory by adding ‘lower terms’ in $W(X, Z)$ in any convenient way. We take them to have the form,

$$(6.47) \quad W(X, Z) = X^{n-1} + X Z^2 - 2\lambda Z.$$

Now the chiral superfield Z is massive, and since it appears quadratically can be integrated out, giving

$$(6.48) \quad W(X) = X^{n-1} - \lambda^2/X,$$

in agreement with eqn.(6.45). Hence the BPS quiver is the same as the D_n minimal model one, that is (up to mutation equivalence) the D_n Dynkin diagram with some orientation of the arrows (again, all orientations are equivalent).

The four-dimensional $\mathcal{N} = 2$ models of these series are studied in detail in ref.[cnv].

6.3.2. Example: $SU(2)$ with $N_f = 0, 1, 2, 3$. • Pure $SU(2)$

The quadratic differential for the $N_f = 0$ theory has the general form

$$(6.49) \quad \phi_2 = \left(\frac{A}{z^3} + \frac{B}{z^2} + \frac{C}{z} \right) dz^2,$$

which has poles of order 3 at the north and south pole of \mathbb{P}^1 . Then its quiver should correspond to the triangulation of the annulus with a marked point on each boundary component, which is the Kronecker quiver, that is the affine $\widehat{A}_1(1, 1)$ quiver.

Let us check this result from the $4d/2d$ correspondence. We choose ω_0 equal to dz^2/z^2 , and write $z = e^X$ with X taking value in the cylinder, *i.e.* $X \sim X + 2\pi i$. The resulting Landau–Ginzburg model is

$$(6.50) \quad W(X) = A e^{-X} + B + C e^X,$$

which is equivalent to the \mathbb{CP}^1 sigma–model, whose BPS spectrum was computed in refs. [1, 51, 52]: the model has two vacua connected by two BPS particles, and hence its BPS quiver is $\widehat{A}_1(1, 1)$.

- $N_f = 1$

The $N_f = 1$ quadratic differential is

$$(6.51) \quad \phi_2 = \left(\frac{A}{z^4} + \frac{B}{z^3} + \frac{C}{z^2} + \frac{D}{z} \right) dz^2.$$

It has a pole of order 4 at the south pole $z = 0$ and one of order 3 at the north pole $z^{-1} = 0$; hence it corresponds to the triangulation of an annulus with one marked point on one boundary and two on the other, whose adjacency quiver is (up to equivalence) equal to the affine quiver $\widehat{A}_2(2, 1)$.

The same conclusion is obtained from the $4d/2d$ correspondence. Choosing ω_0 as in the $N_f = 0$ case, we get the Landau–Ginzburg model on the cylinder

$$(6.52) \quad W(X) = Ae^{-2X} + Be^{-X} + C + De^X,$$

which was solved in refs. [1, 51]. From the solution, one sees that BPS quiver of the model (6.52) is $\widehat{A}_2(2, 1)$.

- $N_f = 2$. *First realization*

$N_f = 2$ has two brane engineerings [24, 34, 50] which correspond to ideal triangulations of *different* bordered surfaces. The two triangulations, corresponding to the same physical theory, have the same adjacency quiver (up to mutation); indeed, this is one of the few cases in which the same mutation–class of quivers corresponds to a pair of topologically distinct surfaces, namely an annulus with two marked points on each boundary, and a disk with one ordinary puncture and three marked points on the boundary. The equality becomes less mysterious if we recall that the first surface has the $\widehat{A}_3(2, 2)$ affine Dynkin quiver, whereas the second should have the \widehat{D}_3 affine Dynkin quiver, and the two quivers are identified by the Lie algebra isomorphism $\mathfrak{su}(4) \simeq \mathfrak{so}(6)$.

The ϕ_2 for the first realization is

$$(6.53) \quad \phi_2 = \left(\frac{A}{z^4} + \frac{B}{z^3} + \frac{C}{z^2} + \frac{D}{z} + E \right) dz^2,$$

which indeed corresponds to an annulus with two marks on each boundary. The corresponding LG model, defined on the cylinder, has superpotential

$$(6.54) \quad W(X) = Ae^{-2X} + Be^{-X} + C + De^X + Ee^{2X},$$

Again, to compute the equivalence class of the BPS quiver we may adjust the constants to convenient values. Setting $B = D = 0$, we recover the $\sinh(2X)$ model solved in ref. [51]. From the explicit solution we see that the BPS quiver is $\widehat{A}_3(2, 2)$, as predicted by the $4d/2d$ correspondence.

- $N_f = 2$. *Second realization*

The ϕ_2 of the second realization is

$$(6.55) \quad \phi_2 = \left(\frac{A}{z^2} + \frac{B}{(z-1)^2} + \frac{C}{z(z-1)} + \frac{D}{z} \right) dz^2,$$

which manifestly corresponds to a disk with two punctures and one mark on the boundary. The corresponding LG model has superpotential

$$(6.56) \quad W(X) = A + \frac{Be^{2X}}{(e^X - 1)^2} + \frac{Ce^X}{(e^X - 1)} + De^X.$$

The check that the BPS quiver of the Landau–Ginzburg model (6.55) is mutation equivalent to $\widehat{A}_3(2, 2)$ is confined in appendix B.

- $N_f = 3$

This model has the quadratic differential

$$(6.57) \quad \phi_2 = \left(\frac{A}{z^2} + \frac{B}{(z-1)^2} + \frac{C}{z} + \frac{D}{z-1} + E \right) dz^2$$

corresponding to the twice-punctured disk with 2 marked points on the boundary, whose adjacency quiver is the affine \widehat{D}_4 .

The LG model is

$$(6.58) \quad W(X) = e^{2X} + \frac{1}{(1-e^{-X})^2} \equiv e^{2X} \frac{(e^X - 1)^2 + 1}{(e^X - 1)^2}.$$

In appendix B it is checked that the BPS quiver of the $2d$ theory is in the mutation class of \widehat{D}_4 .

6.3.3. Example: other affine \widehat{A}, \widehat{D} models. $SU(2)$ gauge theory with $N_f = 0, 1, 2, 3$ gives the first examples of four-dimensional $\mathcal{N} = 2$ models whose Dirac quiver is of the affine \widehat{A} or \widehat{D} type.

The general affine \widehat{A} model corresponds to a quadratic differential on the sphere having two poles of order $n+2$ and $m+2$, with $n, m \geq 1$, that is, to an annulus $\mathcal{A}_{n,m}$ with n (resp. m) marked points on the first (resp. second) boundary. The adjacency quiver of $\mathcal{A}_{n,m}$ is $\widehat{A}_{n+m-1}(n, m)$, i.e. the \widehat{A}_{n+m-1} Dynkin graph with n arrows pointing in the positive direction and m in the negative one.

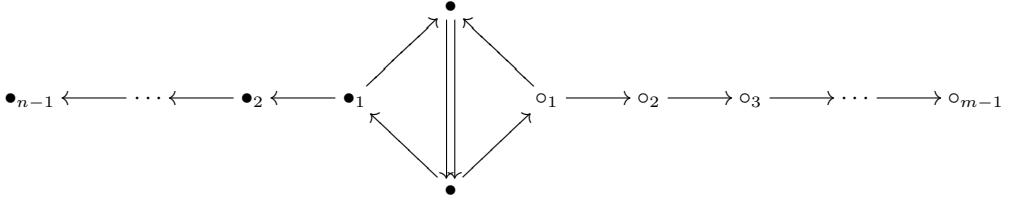


FIGURE 6. The quiver mutation-equivalent to the affine Dynkin quiver $\widehat{A}(n, m)$ (with $n, m \geq 1$) having a Kronecker subquiver.

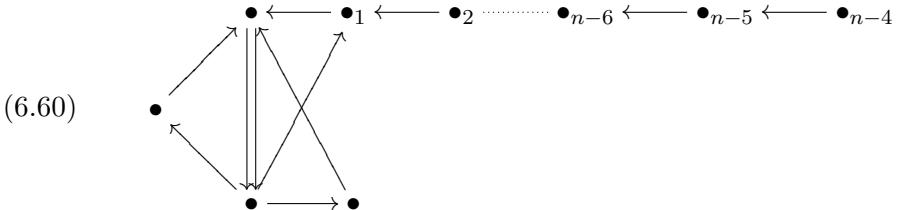
The quiver with the maximal number ($= 1$) of Kronecker subquivers in the mutation-class of the Dynkin quiver $\widehat{A}_{n+m-1}(n, m)$ is represented in figure 6; this quiver may be interpreted as an $SU(2)$ gauge sector coupled to two disconnected $\mathcal{N} = 2$ systems in the sense of section 5.3. Taking $n, m = 1, 2$ we recover $SU(2)$ with $N_f = 0, 1, 2$.

The corresponding 2d theory is

$$(6.59) \quad W(X) = e^{nX} + e^{-mX}.$$

Its BPS spectrum is given by the second case of example 4 in section 8.1 of ref. [1] (n of that reference corresponds to the present $n + m$, while k_0 is to be identified with m) corresponding to an affine \widehat{A}_{n+m-1} Dynkin graph. As a further check, we note that the conjugacy class of the 2d quantum monodromy computed in [1] precisely agrees with that of minus the Coxeter element of the $\widehat{A}(n, m)$ quiver computed in ref. [23].

The affine quivers \widehat{D}_{n-1} correspond to a triangulation of a disk with two punctures and $(n - 3)$ marked points on the boundary. The mutation-equivalent quiver with the maximal number (one) of Kronecker subquivers is obtained by gluing one block of type IV, one of type II, and $n - 5$ blocks of type I,

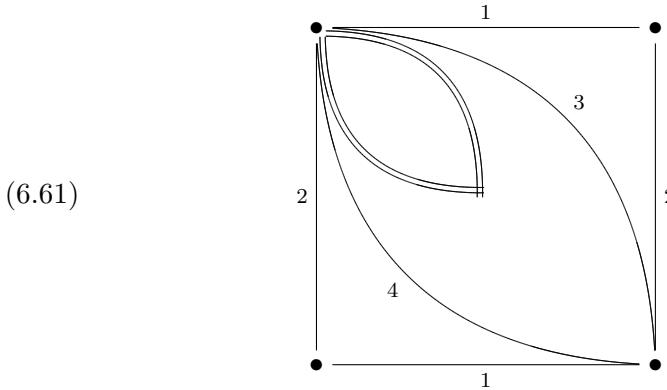


which corresponds to the coupling of $SU(2)$ to three $\mathcal{N} = 2$ systems, two of which being ordinary hypermultiplets. $n = 5$ reproduces $SU(2)$ with $N_f = 3$.

The 2d model is a generalization of the one for $SU(2)$ with three flavors.

There are some exceptional cases. From $SO(6) \simeq SU(4)$, we see that $\widehat{D}_3 \simeq \widehat{A}_3(2, 2)$, and the same quiver represent both the triangulation of a twice-punctured 1-gon and of an annulus with two marks on each boundary. As we have remarked these two surfaces correspond to two different M -theory realizations of $SU(2)$ coupled to two fundamental flavors.

6.3.4. Example: a remarkable unique-quiver AF model. $\mathcal{N} = 2$ and $\mathcal{N} = 4$ $SU(2)$ super-Yang-Mills share a rare property, namely their quivers — respectively the Kronecker and the Markov ones — are the *only* element of their mutation class. In this section, we illustrate a third $\mathcal{N} = 2$ theory with this uniqueness property: the generalized Gaiotto model on the torus with a pole of order three (*i.e.* a boundary with a marked point). Cutting open the torus, we have the ideal triangulation in the figure



where the double line stands for the boundary of the surface. With the numbering of arcs in figure, the adjacency matrix reads

$$(6.62) \quad B_{1,2} = +2 \quad B_{1,3} = -1 \quad B_{1,4} = -1$$

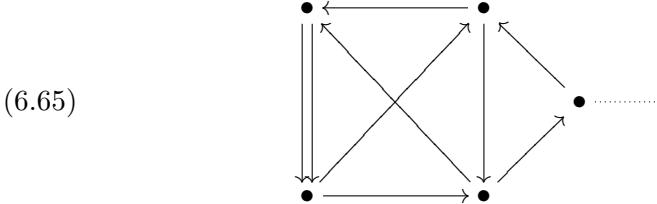
$$(6.63) \quad B_{2,3} = +1 \quad B_{2,4} = +1 \quad B_{3,4} = +1.$$

corresponding to the quiver



Using Keller's mutation applet [42], one checks that this quiver is the only one in its mutation class. This theory has no flavor charge, and it is not UV conformal according to our discussion in section 8, as well as the graphical rule of section 5.3; indeed, (6.64) is a proper subgroup of the finite-mutation

quiver obtained by gluing three type II blocks



In section 9.2 we give an alternative definition of this theory as $SU(2)$ SYM gauging the diagonal $SU(2)$ subgroup of the $SU(2) \times SU(2)$ global symmetry of a composite $\mathcal{N} = 2$ system.

From the $4d/2d$ correspondence perspective, the simplest Landau–Ginzburg superpotential corresponding to this geometry is

$$(6.66) \quad W(X) = \wp'(X).$$

One has

$$(6.67) \quad W'(X) = \wp''(X) = 6\wp(X)^2 - \frac{1}{2}g_2,$$

which gives four supersymmetric vacua at $\pm X_{\pm}$, where $\wp(X_{\pm}) = \pm\sqrt{g_2/12}$. This $2d$ model has all the subtleties we allude before; luckily, they were understood in [1]. The detailed analysis is presented in appendix B.3. The $2d$ computation confirms the quiver (6.64).

6.4. Surface/quiver surgeries. From the general discussion in §.5.3 as well as the examples in the previous two subsections, we see that the process of coupling several basic $\mathcal{N} = 2$ systems to construct more complicated ones is reflected at the quiver level in a kind of graphical gluing process. In the case of generalized Gaiotto theories, this gluing process should be related to a topological surgery of the corresponding bordered surface triangulated in a such a way that the triangulation of the resulting surface may be easily related to those of the several pieces we glue.

The surface surgery process is important from Gaiotto’s duality point of view [34], where $SU(2)$ gauge sectors are described, in their weak coupling limit, as long plumbing tubes connecting punctures in standard degeneration limit of Riemann surfaces. The plumbing parameter is given by $q = e^{2\pi i\tau}$, where τ is the complexified $SU(2)$ coupling. Thus the surgery processes allow us to fill a gap in the discussion of §.5.3 by showing that a Kronecker subquiver **Kr** may be identified with a plumbing tube, which may be taken to be tiny, thus setting the corresponding $SU(2)$ coupling to small values where a Lagrangian description is meaningful.

There are many possible surgery processes, corresponding to the variety of ‘fundamental’ $\mathcal{N} = 2$ systems and of possible supersymmetric couplings between them. Here we limit to the basic ones, without any claim to the completeness of the list. They are the ones with the more transparent physical interpretation.

6.4.1. Massive flavor surgery. Suppose we are in the following situation. In some regime, the Gaiotto theory associated to the closed surface $\mathcal{C}_{g,n}$ looks like two distinct sectors weakly coupled through some bi-fundamental hypermultiplet, carrying his own flavor charge, whose $SU(2) \times SU(2)$ symmetry is weakly gauged by vectors belonging to both of the above sectors. Giving mass to the coupling hypermultiplet, and taking the limit $m \rightarrow \infty$, the theory completely decouples into two distinct $\mathcal{N} = 2$ systems, each corresponding to a piece of the original surface $\mathcal{C}_{g,n}$ which gets broken in two parts in the infinite mass limit. We are interested in understanding the $\mathcal{N} = 2$ physical systems encoded in each surface piece, and their relation to the coupled $\mathcal{N} = 2$ model engineered by the original surface $\mathcal{C}_{g,n}$. Then we wish to learn how to revert the process and couple together the sub-systems by gluing various elementary ‘pieces’ to produce the higher genus surface $\mathcal{C}_{g,n}$.

The connected surface pieces arising from the $m \rightarrow \infty$ limit are necessarily surfaces with boundaries (*i.e.* whose Gaiotto construction has irregular poles). Indeed, the original theory was conformal, and hence the β -functions of all $SU(2)$ groups vanished, including the $SU(2)$ ’s gauging the symmetries of the hypermultiplet whose mass we take to infinity. When this last field is decoupled, the corresponding β -functions will not be zero any longer, but equal to minus the original contribution from the massive hypermultiplet. Therefore, neither of the two remaining decoupled sectors may be superconformal, and hence they cannot correspond to a closed surface. However, since the surgery is local, and only a puncture is involved, the two pieces will have just one boundary component each, and the original puncture associated to the massive flavor will remain as a marked point on each boundary.

From the point of view of the ideal triangulation, this is described as follows. The triangulation has an arc γ , starting and ending at the ‘massive’ (ordinary) puncture, which separates the surface into parts (see figure 7). We cut along the arc γ and separate the surface into two components \mathcal{C}_1 and \mathcal{C}_2 . The arc γ then becomes — on both pieces $\mathcal{C}_1, \mathcal{C}_2$ — a boundary with a marked point at the position of the original puncture. Notice that this process is essentially local, so our discussion applies also to the case in which cutting the separating arc γ will not disconnect the surface, but rather produce two boundaries each with a marked point.

The two pieces are of the form $\mathcal{C}_{g_1, n_1, 1, 1}$ and $\mathcal{C}_{g_2, n_2, 1, 1}$ with

$$(6.68) \quad g = g_1 + g_2$$

$$(6.69) \quad n = n_1 + n_2 + 1.$$

The original quiver associated to the closed surface $\mathcal{C}_{g,n}$ had rank $6g - 6 + 3n$, whereas the rank of each of the resulting subquivers is $6g_i + 3n_i - 2$ so

$$(6.70) \quad \text{rank}(\mathcal{C}_{g,n}) = \text{rank}(\mathcal{C}_{g_1, n_1, 1, 1}) + \text{rank}(\mathcal{C}_{g_2, n_2, 1, 1}) + 1$$

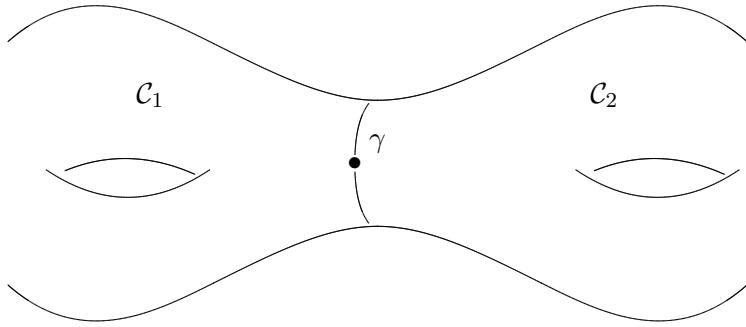


FIGURE 7. A separating arc γ passing through a ‘massive’ ordinary puncture •

which is the correct number since we loose one flavor charge in the infinite mass limit. Instead, if the surface remains connected after cutting γ , it has the form $\mathcal{C}_{g-1,n-1,2,2}$ whose rank $6g - 6 + 3n - 1$ is again one less than the original one.

From the quiver point of view, the process of breaking the surface into two parts is straightforward. One simply eliminates the separating node • and all the arrows connecting it to the rest of the quiver, thus obtaining two disconnected components corresponding to the ideal triangulations of the two pieces \mathcal{C}_1 , \mathcal{C}_2 of the surface $\mathcal{C}_{g,n}$ or a connected quiver corresponding to a surface $\mathcal{C}_{g-1,n-1,2,2}$ having two boundaries each with a marking.

The inverse process, the massive flavor surgery, is also easy to describe. Suppose we are given the quivers, Q_1 and Q_2 , associated to the two pieces each corresponding to a surface \mathcal{C}_i with a boundary γ_i having a single marked point (or the connected adjacency quiver of a surface with two boundary components with one marking each). In the triangulation of \mathcal{C}_1 , the boundary segment γ_1 is either a side of an ordinary triangle, or of a punctured 2-gon, or of a twice-punctured 1-gon (this last possibility occurring only if \mathcal{C}_1 itself is a twice-punctured 1-gon). In the block decomposition of Q_1 , the first two possibilities correspond to a ‘boundary block’ of type, respectively, I or III. In the third case $Q_1 \equiv \widehat{A}_3(2,2)$. The same applies to Q_2 .

The rule to glue together Q_1 , Q_2 ‘in the massive flavor way’ is just to replace, in the block decomposition of each Q_i , the block associated to the boundary γ_i with a block having one more white node ◦ according to figure 8.

Finally, we identify the white nodes ◦ added to the two quivers Q_i getting a connected quiver Q with rank $D(Q) = D(Q_1) + D(Q_2) + 1$. The extra node produced by the process is the massive flavor charge of the coupling hypermultiplet.

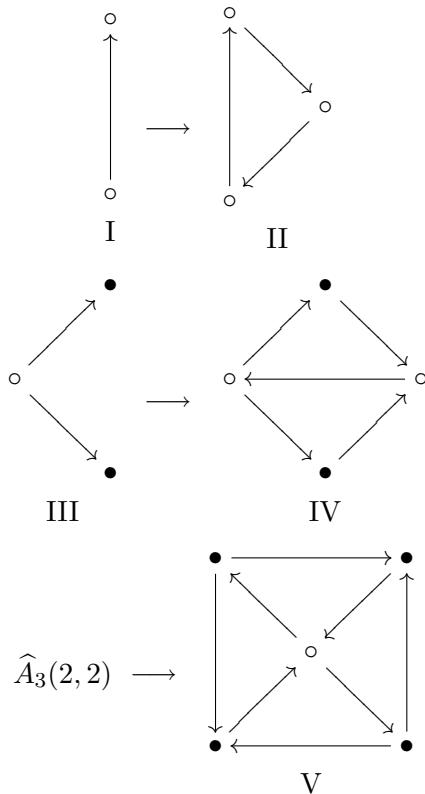
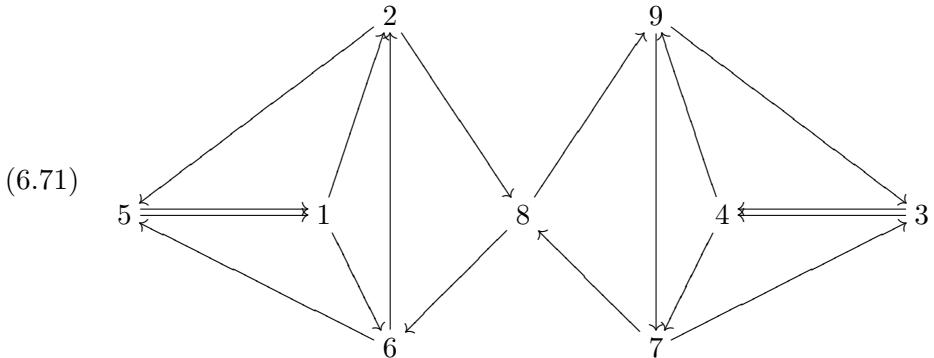


FIGURE 8. Quiver block replacements in massive flavor surgery.

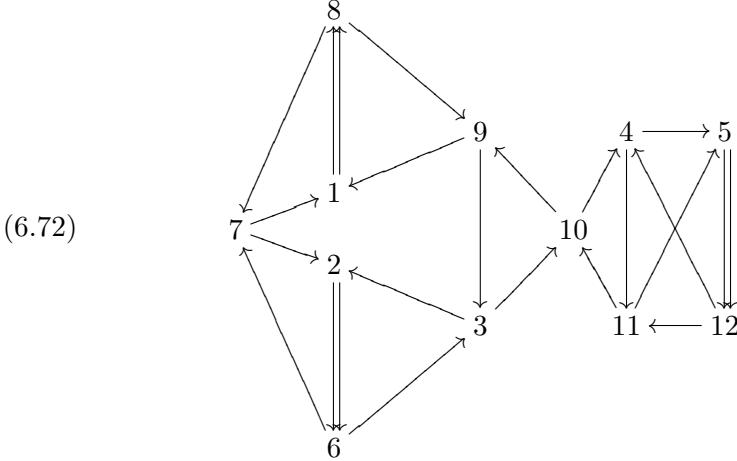
6.4.2. Examples. To simplify the figures, we represent double arrows as single arrows with a 2 in a box.

1. The $g = 2$ $n = 1$ quiver



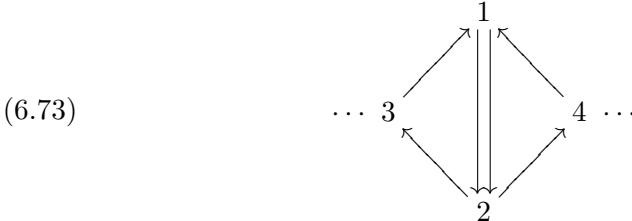
has a separating node, namely 8. Erasing this and the associated arrows, we get two disconnected copies of the quiver associated with a torus with a boundary having a marked point, eqn.(6.64).

2. The $g = 2, n = 2$ quiver



has a separating node, 10. Deleting it and its arrows we get on the right the quiver of a un-punctured torus having one boundary with a marked point, and on the left the quiver of a once-punctured torus with a boundary with a marked point.

6.4.3. Gauge surgery: the tube case. Assume we have a quiver with a Kronecker sub-quiver attached to two oriented triangles as in the figure



where the ellipsis \dots means that the nodes 3, 4 are attached to the rest of the quiver by any number of arrows consistent with the quiver being of the triangulation type. In practice, this means that the nodes 3, 4 should be identified with a white node of some block of the rest of the quiver.

Figure (6.73) corresponds to one of the three ways a Kronecker sub-quiver may appear in a finite-mutation quiver (see §. 5.3), and is physically interpreted as an $SU(2)$ SYM gauging the $SU(2)$ symmetries of the $\mathcal{N} = 2$ systems represented by the subquivers $\dots 3$ and $4 \dots$.

As we shall see momentarily, from the triangulation viewpoint the subquiver (6.73) represents a tube region of the surface $\mathcal{C}_{g,n,b,c}$. Of course, this is nothing else than Gaiotto's descriptions of $SU(2)$ gauge groups as plumbing tubes [34]. Then we can borrow his analysis of the relation between the (complexified) $SU(2)$ coupling τ and the plumbing parameter $q = e^{2\pi i\tau}$. The weak coupling limit then corresponds to a tube in the Riemann surface $\mathcal{C}_{g,n,b,c}$ which becomes infinitely long. In the limit $q = e^{2\pi i\tau} \rightarrow 0$, the tube

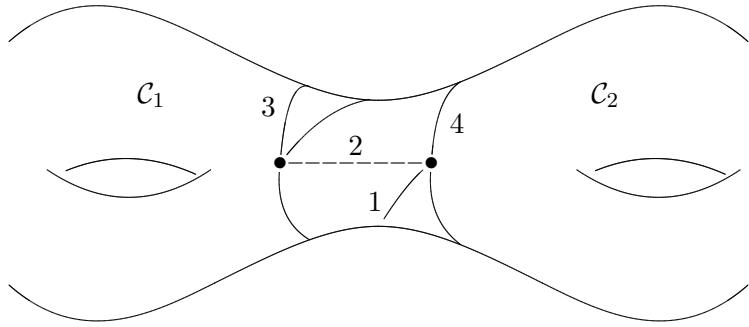


FIGURE 9. The punctures and arcs corresponding to the subquiver (6.73).

pinches, and we remain either with two disconnected surfaces, $\mathcal{C}_{g_1, n_1, b_1}$ and $\mathcal{C}_{g_2, n_2, b_2}$, where

$$(6.74) \quad g_1 + g_2 = g, \quad n_1 + n_2 = n + 2, \quad b_1 + b_2 = b,$$

or with a connected surface $\mathcal{C}_{g', n', b'}$ with

$$(6.75) \quad g' = g - 1, \quad n' = n + 2, \quad b' = b.$$

In either cases, the total number of nodes in the (possibly disconnected) quiver is conserved.

By the very concept of complete $\mathcal{N} = 2$, the decoupled $q \rightarrow 0$ theories should be also complete, and their quivers of finite-mutation type. Thus the coupling/decoupling process may be expressed in the quiver-theoretical language.

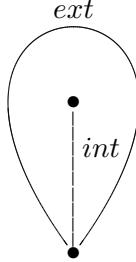
In the triangulation of $\mathcal{C}_{g,n,b,c}$, the sub-quiver (6.73) corresponds to a tube or, more precisely, to a cylinder $C_{1,1}$ with a marked point on each boundary which is glued through its two boundary arcs — corresponding to nodes 3 and 4 in (6.73) — to the rest of the surface $\mathcal{C}_{g,n,b,c}$ in such a way that the two markings on the boundaries of $C_{1,1}$ correspond to two (ordinary) punctures of the surface $\mathcal{C}_{g,n,b,c}$.

The cylinder with a marking on each boundary, $C_{1,1}$, is precisely the surface corresponding to pure $SU(2)$ $\mathcal{N} = 2$ super-Yang-Mills. We represent the cylinder $C_{1,1}$ as a rectangle with the two vertical sides identified. Then an ideal triangulation looks like



or, equivalently figure 9, where the arcs are numbered as the nodes in the subquiver (6.73).

To do the surgery, we cut away the cylinder $C_{1,1}$ along the two separating arcs 3 and 4. This operation produces two boundaries each with a marked point \bullet . Next we glue to each of these two boundaries a self-folded triangle along its external arc ext



which introduces the extra puncture replacing the pinched tube.

The net result of gluing the self-folded triangle, is replacing the block attaching the node 3 (resp. 4) to the rest of the quiver in the following way

type original block	type replacing block (*)
I	III
II	IV
III	$\widehat{A}_3(2,2)$
IV	V
V	excep. trian. 4-punct. sphere

TABLE. Gauge tube surgery

(*) Attaching blocks of type III and V are possible only for \mathcal{C}_1 equal to the twice-punctured 1-gon and, respectively, the 4-punctured sphere (with its exceptional triangulation).

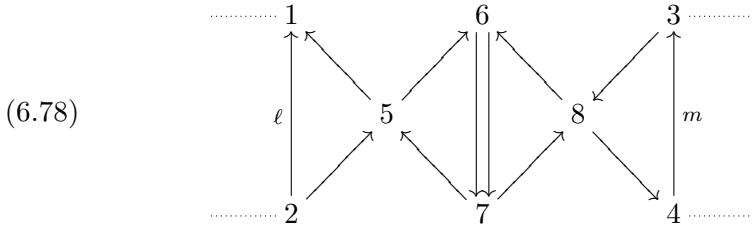
In the last step each of the two adjacency quivers of $\mathcal{C}_1, \mathcal{C}_2$ gets an extra node (associated to the internal arc int of the glued self-folded triangle); since in the process we have lost the two nodes associated to arcs 1 and 2 in figure (6.76), the total number of nodes is conserved, as expected.

The inverse process (*gluing*) is also easy. One takes two surfaces, $\mathcal{C}_{g_1, n_1, b_1, c_1}$ and $\mathcal{C}_{g_2, n_2, b_2, c_2}$, triangulated in such a way that the corresponding quivers have one of the blocks in the second column of table (6.77). These blocks correspond to a ‘puzzle piece’ of the triangulation containing a self-folded triangle. Then one cuts away the self-folded triangles from the corresponding ‘puzzle pieces’ of the two triangulated surfaces, producing a boundary with one marked point on each surface $\mathcal{C}_{g_1, n_1, b_1, c_1}, \mathcal{C}_{g_2, n_2, b_2, c_2}$. Finally one glues these boundaries to the boundaries of the cylinder (6.76) identifying the marked points.

The above ‘tube’ surgery is only a special instance of the coupling of two $\mathcal{N} = 2$ theories by replacing a pair of punctures by a thin tube. It works under the special assumption that both surfaces to be glued are triangulated

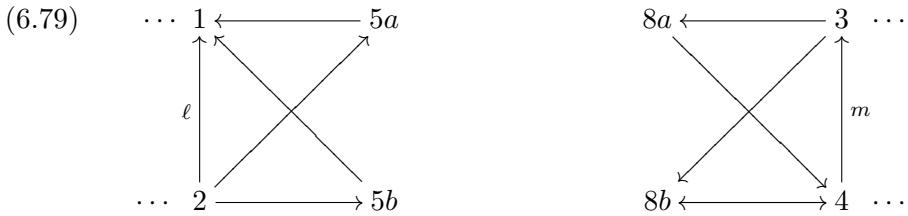
in such a way that a self-folded triangle exists (in particular, each surface must have at least either *two* punctures or a puncture and a boundary). There are more general ways of gluing quivers, which make sense under weaker assumptions on the two surfaces to be glued. We may glue, for instance, the quivers of a higher genus surface with one puncture to that of a surface with two punctures. However, it is not possible to relax this milder condition. The point is that, otherwise, we could get the quiver of a puncture-less surface by gluing two once-punctured lower genus ones. But this is clearly impossible.

6.4.4. Example: generalized hypermultiplet gaugings. Assume that the $SU(2)$ SYM associated to the tube to be pinched is coupled to the other sectors by two generalized ‘hypermultiplets’. At the quiver level, this means that we have a full subquiver of the form



where the \dots stands for any number of arrows connecting the four nodes 1, 2, 3, 4 of the subquiver to the nodes of the rest of the quiver, while the nodes 2 and 1 (resp. 4 and 3) are connected by ℓ arrows (resp. m arrows).

The triangles 1, 2, 5 and 3, 4, 8 correspond to blocks of type II. Decoupling the $SU(2)$, they get replaced by type IV blocks (cfr. table (6.77)). Then, as $\tau \rightarrow 0$ we get



(the full quiver may or may not be disconnected).

If $\ell = m = 2$, corresponding to an ordinary bi-fundamental hypermultiplet, we break the tube by replacing a gauge group and two bi-fundamentals by two pairs of fundamental hypermultiplets coupled to the two $SU(2)$'s associated to the pairs of nodes 1, 2 and 3, 4, respectively.

6.4.5. Examples: gauging $\mathcal{N} = 2$ subsystems. From the above we see that we can couple the $SU(2)$ gauge system any Gaiotto $\mathcal{N} = 2$ system whose surface \mathcal{C} has at least one ordinary puncture (subject to the condition that the glued surface has at least one puncture — if we wish a theory with a well-defined quiver). Such a system admits an $SU(2)$ global symmetry which can be gauged.

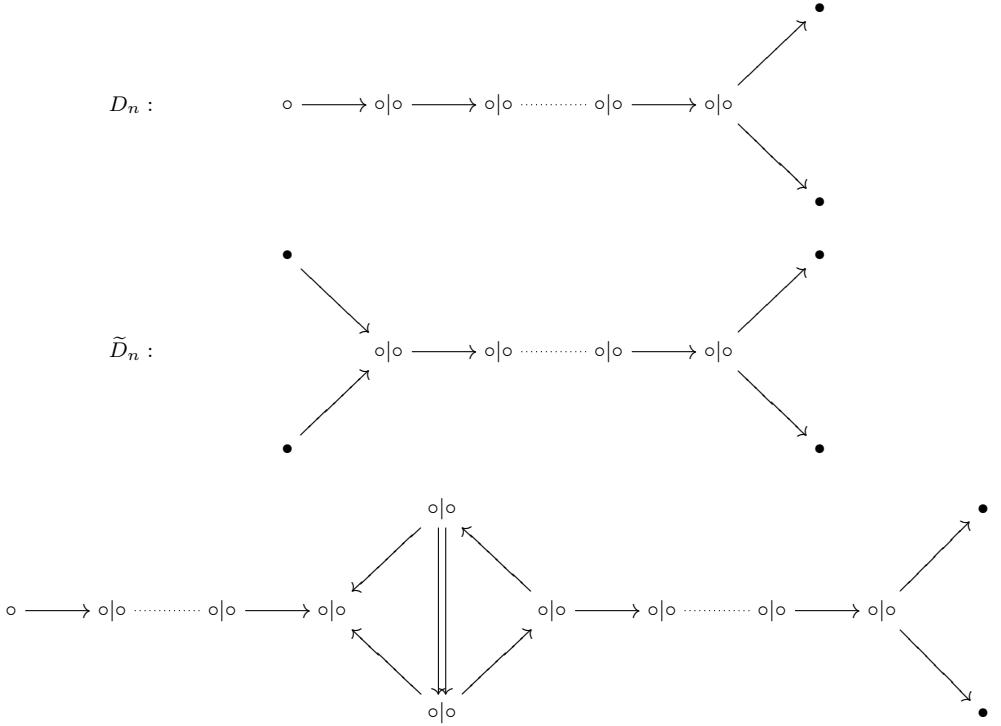


FIGURE 10. The D_n , \tilde{D}_n and $\Gamma_{n,m}$ quivers decomposed into blocks: the last block on the right is of type III. The blocks are divided by the vertical line $|$; the two \circ 's separated by a vertical line should be identified to get back the original quiver.

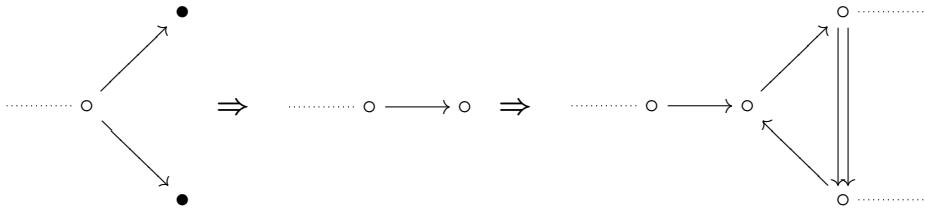
The more elementary such surfaces \mathcal{C} are:

- the punctured disk with m marked points on the boundary whose adjacency quiver is (up to mutation equivalence) the Dynkin quiver D_m ;
- the twice-punctured disk with m marking on the boundary corresponding to the affine \tilde{D}_{m+2} quivers, mutation equivalent to (6.60);
- the punctured annulus with (n, m) marking on the boundaries, last quiver in figure 10.

In their standard (Dynkin) form, the corresponding quivers contain one type III block (two for \tilde{D}_{m+2} , associated to the two ordinary punctures), as can be seen from the block decompositions in figure 10 .

By the rules of the gauge tube surgery, we may replace that type III block by a type I, and couple the ‘new’ white node to a Kronecker subquiver via an oriented triangle. We describe this process as a ‘gauging’ of the system described by the original surface \mathcal{C} .

Graphically, the gauging procedure looks as follows
(6.80)



There is a field theory explanation of the above surgery. The idea is that each block of type III in an adjacency quiver Q carries a global $SU(2)$ symmetry, and the surgery is just gauging it. Indeed, in presence of a type III block we have a special flavor charge¹⁸ J with weights +1 and -1 for the two black nodes of the type III block and zero elsewhere. The quiver (and hence the physics) is symmetric under the simultaneous interchange of the two black nodes and the corresponding mass parameters. This \mathbb{Z}_2 symmetry acts on the above charge as $J \rightarrow -J$, so the natural interpretation is that J is the Cartan generator of $\mathfrak{su}(2)$ and \mathbb{Z}_2 its Weyl group.

We can check this interpretation in a special case. From figure (6.80) we see that the gauging of an ordinary fundamental hypermultiplet corresponds to the gauging of the $D_2 \sim A_1 \times A_1$ Argyres–Douglas system: a fundamental hypermultiplet is *two* free hypermultiplets each with its own $SU(2)$ flavor charge. In other words we can consider the subquiver consisting of the two end nodes of the D_n series, which corresponds to two decoupled hypermultiplets, which can be gauged by the $SU(2)$. In this way the BPS quiver keeps only one of the two fundamentals (as discussed in the context of BPS quivers of $SU(2)$ coupled to one fundamental), as the other one can be obtained by the combination of elements of $SU(2)$'s Kronecker quiver, and one of the two fundamental states. This explains why effectively we get rid of one of the two end nodes of the D diagram and connect the remaining node to the Kronecker quiver in the standard way.

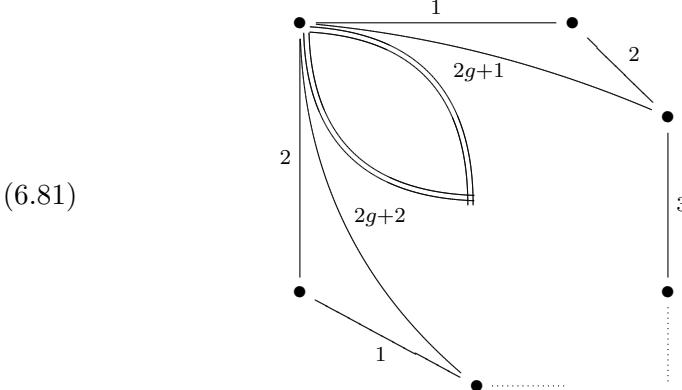
A preliminary discussion of the physical properties of these gauged $\mathcal{N} = 2$ systems are presented in section 9.

6.5. Vector-less quivers. In this section we show why the only ‘vector-less’ quivers are the ADE Dynkin ones. By this we mean that this is the only class which does not have any double arrows in any mutation of the corresponding quiver. For the eleven exceptional classes the fact that there are double lines in the quiver follows from direct inspection. It remains to consider the adjacency quiver of bordered surfaces.

The example in §.6.2.3 shows that all surfaces with $g \geq 1$ and at least one puncture have a triangulation with at least one double–arrow. On the other hand, suppose we have a surface with $g \geq 1$ and $b \geq 1$. We may cut

¹⁸ Recall that a flavor charge is a vector in Γ which is a zero eigenvector of the exchange matrix B .

open the surface to get a hyperbolic $4g$ -gon and start triangulating as in the figure



which gives $B_{12} = +2$. Hence all $g \geq 1$ triangulation quivers are mutation-equivalent to ones having at least one double-arrow.

For $g = 0$, all surfaces with $n \geq 4$ or $b \geq 2$ have quivers in the mutation-class with double arrows. Taking into account the restrictions on n , b , c for $g = 0$ [3], we remain with the possibility $b = 1$. If $b = 1$ and $n = 2$ we have affine- \widehat{D} quivers which are mutation-equivalent to those in figure (6.60) having a Kronecker subquiver.

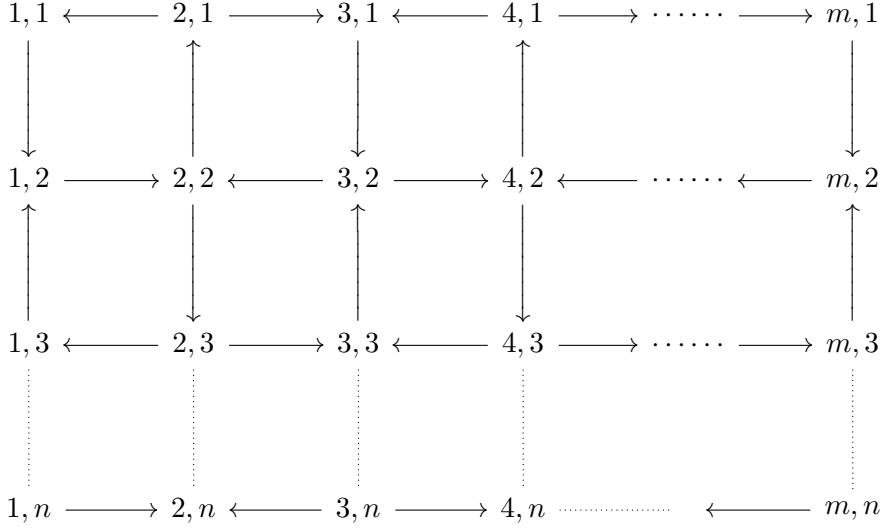
We remain with surfaces with $b = 1$, $n = 0$, corresponding to the mutation class of the A_r Dynkin quivers, and $b = 1$, $n = 1$, associated to the mutation class of the D_r Dynkin ones. These finite Dynkin quivers are known to be vector-free.

7. Identification of the exceptional theories

It remains to identify the complete $\mathcal{N} = 2$ theories associated to the eleven exceptional mutation classes which are mutation-finite but not associated to the ideal triangulation of any surface. They may be divided in four families (we write a standard representative for each mutation-class):

- (1) finite-type Dynkin quivers of type E_6, E_7, E_8 ;
- (2) affine-type Dynkin quivers of type $\widehat{E}_6, \widehat{E}_7, \widehat{E}_8$;
- (3) Saito's [43] elliptic-type Dynkin quiver (with oriented triangles) of type $\widehat{\widehat{E}}_7, \widehat{\widehat{E}}_7, \widehat{\widehat{E}}_8$;
- (4) the Derksen–Owen quivers X_6 and X_7 [4].

The models associated to the first family, E_6, E_7, E_8 , were already discussed in [2]. They are a generalization of the Argyres–Douglas model corresponding to the world-sheet theory of a M5-brane compactified to four dimension on a complex curve with equation the corresponding E -type

FIGURE 11. The $A_m \square A_n$ quiver.

minimal singularity

$$(7.1) \quad \begin{array}{c|c} E_6 & y^3 + x^4 = 0 \\ \hline E_7 & y^3 + yx^3 = 0 \\ \hline E_8 & y^3 + x^5 = 0 \end{array}$$

They are UV conformal, and vector-less.

7.1. Elliptic and affine E -models. The elliptic E -models turn out to be special instances of the class of models studied in [2] which are labelled by a pair (G, G') of simply-laced Dynkin graphs ($G, G' = ADE$). They correspond to the $4d$ $\mathcal{N} = 2$ theory obtained by compactifying Type IIB superstring on the local Calabi–Yau hypersurface $\mathcal{H} \subset \mathbb{C}^4$ of equation

$$(7.2) \quad \mathcal{H}: \quad W_G(x_1, x_2) + W_{G'}(x_3, x_4) = 0,$$

where $W_G(x_1, x_2) + x_0^2$ is the canonical surface singularity associated to the given Dynkin diagram G . The quiver of the (G, G') model is given by the *square tensor product* of the Dynkin graphs of G and G' , $G \square G'$ (for the product orientation rule see refs. [2, 55]). The quiver $A_m \square A_n$ is represented in figure 11.

Up to mutation-equivalence one has the following identifications [3]:

$$(7.3) \quad \widehat{\widehat{E}}_6 \sim A_2 \square D_4$$

$$(7.4) \quad \widehat{\widehat{E}}_7 \sim A_3 \square A_3$$

$$(7.5) \quad \widehat{\widehat{E}}_8 \sim A_2 \square A_5.$$

The first one may be further simplified using $D_4 \sim A_2 \square A_2$ [2]. Hence the corresponding 4d $\mathcal{N} = 2$ models may be engineered by Type IIB on the hypersurface \mathcal{H} :

quiver	CY hypersurface \mathcal{H}	(n_1, n_2, n_3)
$\widehat{\tilde{E}}_6$	$x_0^2 + x_1^3 + x_2^3 + x_3^3 + ax_1x_2x_3 = 0$	$(2, 2, 2)$
$\widehat{\tilde{E}}_7$	$x_0^2 + x_1^4 + x_2^4 + x_3^2 + ax_1x_2x_3 = 0$	$(3, 3, 1)$
$\widehat{\tilde{E}}_8$	$x_0^2 + x_1^3 + x_2^6 + x_3^2 + ax_1x_2x_3 = 0$	$(2, 4, 1)$

Notice that the section $x_0 = 0$ of each hypersurface is a quasi-homogeneous cone over an elliptic curve embedded in some weighted projective space. Indeed the Saito's elliptic roots systems are related to elliptic singularities. The only other elliptic Dynkin diagram which is a finite-mutation quiver is $\widehat{\tilde{D}}_4$ which corresponds to $SU(4)$ with $N_f = 4$ (*i.e.* the sphere with four punctures).

In a $\widehat{\tilde{E}}_r$ mutation class there are many quivers having a transparent physical interpretation. First of all, we have the tensor product quivers $G \square G'$, $G' \square G$, $G' \boxtimes G$, and $G \boxtimes G'$, which using the results of ref. [2] imply that the model is UV conformal with a quantum monodromy $\mathbb{M}(q)$ of order¹⁹

$$(7.7) \quad r = \frac{h(G) + h(G')}{\gcd\{h(G), h(G')\}} = \begin{cases} 2 & \text{for } \widehat{\tilde{E}}_7 \\ 3 & \text{for } \widehat{\tilde{E}}_6, \widehat{\tilde{E}}_8, \end{cases}$$

which means, in particular, that the the UV $U(1)_R$ charges r_i of the primary operators are of the form $\frac{1}{r} \mathbb{N}$. Moreover, the (G, G') $\mathcal{N} = 2$ model has two special chambers with a *finite* BPS spectrum consisting only of hypermultiplets. In the first such chamber they have charges [2]

$$(7.8) \quad \alpha_i \otimes \sum_a n_a^{(s)} \beta_a \in \Gamma_G \otimes \Gamma_{G'} \simeq \Gamma_{G \square G'},$$

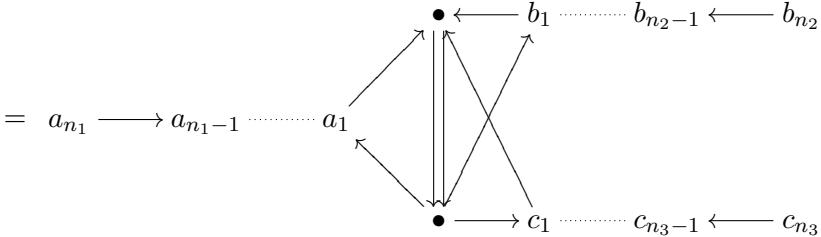
where $\alpha_i \in \Gamma_G$ are the *simple* roots of G and $\sum_a n_a^{(s)} \beta_a \in \Gamma_{G'}$ are all the positive roots. In the second chamber the two Dynkin diagrams interchange roles $G \leftrightarrow G'$. On the other hand, the $\widehat{\tilde{E}}_r$ quivers are not vector-less and hence have regimes described by mutation-equivalent quivers containing Kronecker subquivers; indeed the usual elliptic Dynkin forms have one Kronecker sub-quiver, see figure 1, and they correspond to pure $SU(2)$

¹⁹ Eqn.(7.7) holds for the groups in eqns.(7.3)–(7.5) but not in general. For the general case see [56].

coupled to three $\mathcal{N} = 2$ D -systems of the kind discussed in sections 5.3 and 6.4.5.

The family of coupled three $\mathcal{N} = 2$ D -systems has quivers of the suggestive form

$$(7.9) \quad Q(n_1, n_2, n_3) =$$



(notice that the quiver is symmetric under the interchanging of the nodes with a , b and c labels.) $Q(1, 1, 1) \simeq \widehat{D}_4$ is just the quiver of $SU(2)$ with three flavors.

The $\widehat{\tilde{E}}_r$ $\mathcal{N} = 2$ models are engineered by Type IIB on the CY hypersurface $x_0^2 + W_{n_1, n_2, n_3}(x_1, x_2, x_3) = 0$, where $W_{n_1, n_2, n_3}(x_1, x_2, x_3)$ is the equation of the elliptic curve in weighted projective space

$$(7.10) \quad W_{n_1, n_2, n_3}(x_1, x_2, x_3) \equiv x_1^{n_1+1} + x_2^{n_2+1} + x_3^{n_3+1} + \lambda x_1 x_2 x_3$$

and the integers (n_1, n_2, n_3) are as in the table (7.6). The corresponding quiver is simply $Q(n_1, n_2, n_3)$ for the same triplet of integers. Following our discussion in section 5.3, we expect that these models have BPS chambers, different from the two finite-spectrum ones analyzed in ref. [2], with BPS vector multiplets in the spectrum weakly coupled to the supersymmetric D -systems.

This completes the identification for the elliptic- E $\mathcal{N} = 2$ models as the models obtained by compactifying Type IIB on the corresponding CY hypersurface, see table (7.6).

More generally, we may ask for which triplet of integers (n_1, n_2, n_3) — besides the ones in table (7.6) — the quiver $Q(n_1, n_2, n_3)$ is of the finite-mutation type. Not surprisingly, the condition turns out to be

$$(7.11) \quad \frac{1}{n_1 + 1} + \frac{1}{n_2 + 1} + \frac{1}{n_3 + 1} \geq 1,$$

in one-to-one correspondence with Coxeter reflection groups for the sphere and the plane. The $\mathcal{N} = 2$ theories for which the inequality \geq in eqn.(7.11) is replaced by equality $=$ are actually UV superconformal (see next section).

The solutions to condition (7.11) are listed in table 1.

From the table we infer an interpretation of the affine- \widehat{E} quivers. They are precisely the asymptotically free, complete $\mathcal{N} = 2$ models associated to

n_1, n_2, n_3	equivalent Dynkin quiver	
(7.12)	\widehat{D}_{s+3}	disk with $n = 2, c = s + 1$
	\widehat{E}_6	asymptotically free
	\widehat{E}_7	asymptotically free
	\widehat{E}_8	asymptotically free
	$\widehat{\widehat{E}}_6$	superconformal
	$\widehat{\widehat{E}}_7$	superconformal
	$\widehat{\widehat{E}}_8$	superconformal

TABLE 1. The solutions (n_1, n_2, n_3) to condition (7.11) and the Dynkin quiver mutation-equivalent to the quiver $Q(n_1, n_2, n_3)$.

Type IIB on the (UV fixed point of the) hypersurface

$$(7.13) \quad x_0^2 + x_1^{n_1+1} + x_2^{n_2+1} + x_3^{n_3+1} + \lambda x_1 x_2 x_3 = 0$$

where n_1, n_2, n_3 are as specified in the table 1.

Table 1 gives us also an alternative construction of affine- \widehat{D} models in terms of Type IIB engineering.

As a further check of the identifications for the affine $\widehat{D}_r, \widehat{E}_r$ models in table 1, let us consider it from the point of view of the $4d/2d$ correspondence. The above identifications gives the $2d$ Landau–Ginzburg model with superpotential $W_{n_1, n_2, n_3}(x_1, x_2, x_3)$ in eqn. (7.10). The $\widehat{D}_r, \widehat{E}_r$ affine Dynkin diagrams correspond to the triplets of integers (n_1, n_2, n_3) with

$$(7.14) \quad \frac{1}{n_1 + 1} + \frac{1}{n_2 + 1} + \frac{1}{n_3 + 1} > 1.$$

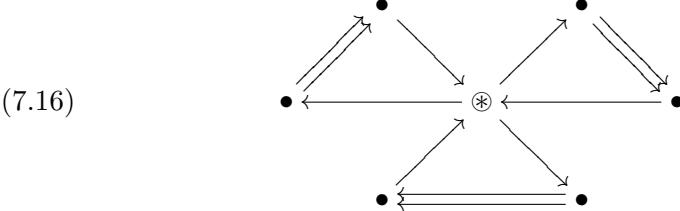
The identification requires the Witten index of the two dimensional model to be equal to the rank of the corresponding affine Lie algebra, *i.e.* $r + 1$. A direct computation shows that, under the condition (7.14), one has

$$(7.15) \quad 2d \text{ Witten index} = n_1 + n_2 + n_3 + 2 = \left\{ \begin{array}{ll} s + 4 & \text{for } \widehat{D}_{s+3} \\ 7 & \text{for } \widehat{E}_6 \\ 8 & \text{for } \widehat{E}_7 \\ 9 & \text{for } \widehat{E}_8 \end{array} \right\} \equiv r + 1.$$

This result supplements the classification of $2d \mathcal{N} = 2$ affine models [1].

7.2. The Derksen–Owen quivers X_7, X_6 . There remain only two mutation-finite classes: X_7 and X_6 .

7.2.1. X_7 . The mutation class of X_7 consists of just two distinct quivers [4]. The one with double-arrows is



The quiver (7.16) is maximal finite-mutation (**Theorem 13** of [4]), and hence it is expected to correspond to an UV conformal $\mathcal{N} = 2$ theories (this prediction will be confirmed momentarily).

X_7 has one flavor charge, associated to the node in (7.16) represented by the symbol \otimes . The corresponding vector in the charge lattice is

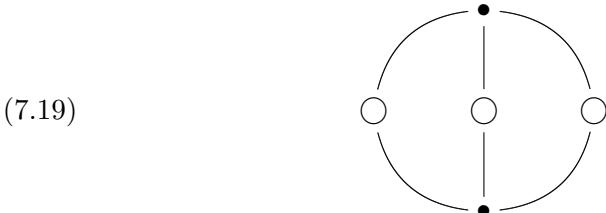
$$(7.17) \quad \text{flavor charge vector} = \gamma_{\otimes} + \frac{1}{2} \sum \gamma_{\bullet}.$$

The physical interpretation of this quiver is straightforward. Associated to the above flavor charge we have a mass parameter m . Taking $m \rightarrow \infty$, we approach a limit where a weakly coupled Lagrangian description is adequate: We have a *full* hypermultiplet in the quaternionic (pseudoreal) representation

$$(7.18) \quad (\mathbf{2}, \mathbf{2}, \mathbf{2})_{+1} \oplus (\mathbf{2}, \mathbf{2}, \mathbf{2})_{-1}$$

of its symmetry group $SU(2) \times SU(2) \times SU(2) \times SO(2)$ and the three $SU(2)$'s are weakly gauged by three copies of $SU(2)$ SYM represented by the three Kronecker subquivers, $\bullet \rightrightarrows \bullet$, in figure (7.16). Its unique flavor charge (7.17) corresponds to the $SO(2)$ symmetry of the hypermultiplet with mass parameter m .

Taking $m \rightarrow 0$, this model reduces to the conformal Gaiotto model with $g = 2$ and *no* puncture. Indeed, in some corner of its moduli space, the genus two curve with no punctures may be physically interpreted as in the figure



where the \circlearrowleft 's stand for $SU(2)$ gauge groups and the \bullet 's for tri-fundamental *half*-hypermultiplets. The two half-hypermultiplets have the same quantum numbers with respect to all gauge groups, and so we may combine them into a *complete* hypermultiplet in the $(\mathbf{2}, \mathbf{2}, \mathbf{2})$ of the $SU(2)^3$ gauge group. This process introduces — in the above Lagrangian corner of the moduli space

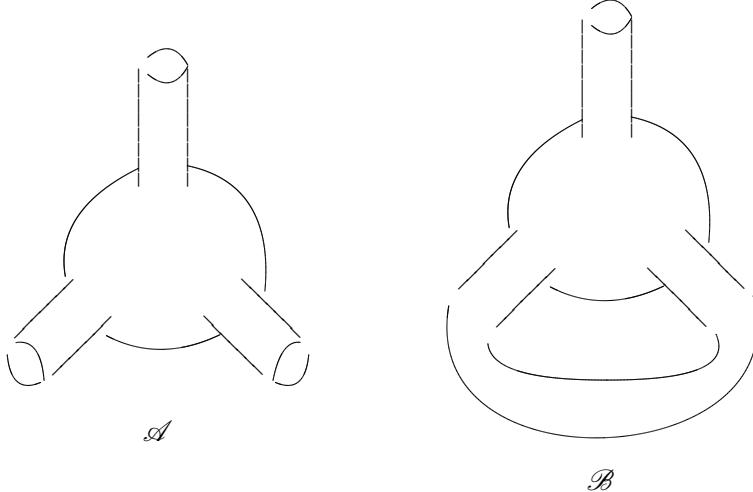


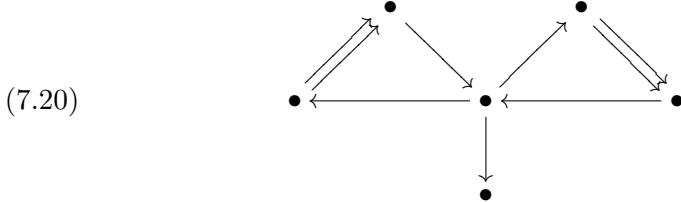
FIGURE 12. \mathcal{A} : A trifundamental half-multiplet corresponds to a thrice-punctured sphere attached to three long plumbing tubes. \mathcal{B} : Two of the three punctures may be connected by a long tube making a handle.

— an emergent $SO(2)$ symmetry — not present in the original Gaiotto construction — which is the one associated to the node \circledast of the X_7 quiver. In particular, the relation with Gaiotto’s $g = 2$ theory shows that the X_7 $\mathcal{N} = 2$ is UV conformal, as expected from the graphical rule.

This emergence of a flavor symmetry is special to $g = 2$, and does not generalizes to $g > 2$. This explains why X_7 is an isolated exception without higher rank analogues. Indeed, in the Gaiotto framework [34], the degeneration of a genus $g > 1$ surface without punctures into three-punctured spheres connected by long cylinders corresponds to a Lagrangian description in which each punctured sphere corresponds to a trifundamental *half*-hypermultiplet in the representation $(\mathbf{2}, \mathbf{2}, \mathbf{2})$ of $SU(3)^3$, which has *no* flavor symmetry (see figure 12. \mathcal{A}), while each long cylinder corresponds to a weakly coupled $SU(2)$ SYM. In order to have a flavor symmetry, we need at least two such half-hypermultiplets in the same representation of the gauge group. This may happen only if the three punctures of the sphere representing the second half-hypermultiplet are connected to the same three tubes as the sphere representing the first one. Then the two punctured spheres and the three tubes connecting them form a $g = 2$ surface disconnected from the rest. The only other possibility is that two punctures of the same sphere are connected together to form a handle (as in figure 12. \mathcal{B}). This also leads to $g = 2$, see next section.

From the figure (7.19) it is obvious that the model is UV conformal: Indeed, each $SU(2)$ ‘sees’ four fundamental hypermultiplets, and hence has a vanishing β -function.

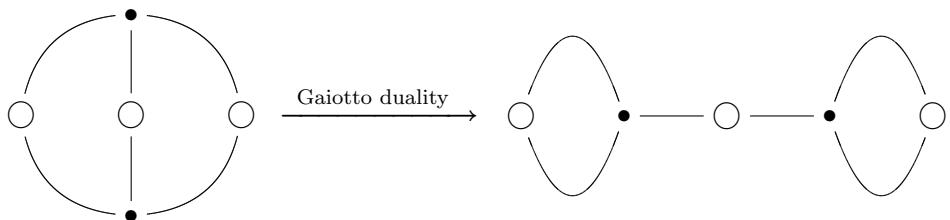
7.2.2. X_6 . The X_6 exceptional mutation class consists of five distinct quivers [4]. Two of them have double arrows (they are source/sink equivalent, and hence represent essentially the same physics),



The X_6 quiver has no flavor charge. X_6 is not maximal mutation-finite, but the only mutation-finite quiver containing it is X_7 itself (**Theorem 12** of [4]). Hence the corresponding $\mathcal{N} = 2$ theory must be UV asymptotically free, and must arise as a particular decoupling limit of the $X_7 \mathcal{N} = 2$ model. In fact, as already discussed any subsystem of a quiver can be viewed as arising in a particular limit of moduli space of that theory. Thus X_6 which is a subquiver of X_7 obtained by deleting one of the nodes of a double line can be obtained from a limit of X_7 theory.

That the X_7 theory has such a limit may be understood more explicitly. By the very concept of complete $\mathcal{N} = 2$ theory, the X_7 model has enough quantum consistent deformations that we may actually realize as sensible QFT all of its formal geometric limits. In particular, in the deformation space of X_7 there should be contained all relevant/marginal deformations of any theory related by Gaiotto dualities to the $g = 2$ conformal theory which is the $m \rightarrow 0$ limit of X_7 .

Between the Gaiotto dual theories, we have the one corresponding to the degeneration of the $g = 2$ surface shown in the right hand side of the figure
(7.21)



where, again, \circlearrowleft ’s stand for $SU(2)$ gauge groups and \bullet ’s for half-hypermultiplets $h_{\alpha\dot{\alpha}\dot{\alpha}}^{(a)}$ in the $(\mathbf{2}, \mathbf{2}, \mathbf{2})$ of $SU(2)^3$.

In the second limit, the same $SU(2)$ SYM gauges the first pair $\alpha, \dot{\alpha}$ of $SU(2)$ indices, so that the matter representation content in terms of the

three gauge groups is

$$(7.22) \quad \left((\mathbf{3}, \mathbf{1}, \mathbf{2}) \oplus (\mathbf{1}, \mathbf{1}, \mathbf{2}) \right) \oplus \left((\mathbf{1}, \mathbf{3}, \mathbf{2}) \oplus (\mathbf{1}, \mathbf{1}, \mathbf{2}) \right).$$

In this duality limit, we have two half-hypermultiplets with the same quantum numbers under all gauged symmetries, namely $(\mathbf{1}, \mathbf{1}, \mathbf{2})$, and hence an $SO(2)$ flavor symmetry rotating them. To this $SO(2)$ symmetry we may associate a mass deformation, μ . Since the X_7 theory is complete, this deformation should correspond to a region in its coupling space.

At this point we take the decoupling limit $\mu \rightarrow \infty$. We get a $\mathcal{N} = 2$ theory with a charge lattice of rank 6, no flavor charge, which is asymptotically free. Assuming there is a BPS quiver for this theory, it should be mutation-finite and contained in X_7 . There is only one such quiver, namely X_6 .

8. Conformal, complete theories

8.1. $U(1)_R$ symmetry. The $\mathcal{N}=2$ theories corresponding to mutation-finite quivers, being UV complete QFT, in the ultra-violet are either conformal or asymptotically free. In the first case there is a point in their parameter space (belonging to some specific chamber and hence corresponding to a particular quiver in the given mutation-class) in which the full superconformal invariance is restored.

In this section we address the question of classifying the subset of complete $\mathcal{N} = 2$ theories which have such a superconformal point. In $4d$, a necessary condition for $\mathcal{N} = 2$ superconformal invariance is the existence of a conserved $U(1)_R$ current. More precisely, the $U(1)$ associated to the overall phase of the Seiberg–Witten differential λ should become a symmetry at the conformal point.

For a generalized Gaiotto model, this $U(1)$ acts on the quadratic differential as

$$(8.1) \quad \phi_2 \rightarrow e^{2i\theta} \phi_2.$$

Hence, for this class of models, we have a conserved $U(1)_R$ symmetry iff there exists a complex automorphism of the surface \mathcal{C} , $f_\theta: \mathcal{C} \rightarrow \mathcal{C}$, such that

$$(8.2) \quad f_\theta^* \phi_2 \Big|_{\substack{\text{conformal} \\ \text{point}}} = e^{2i\theta} \phi_2 \Big|_{\substack{\text{conformal} \\ \text{point}}}.$$

For the kind of punctured bordered surfaces of interest here, we have a continuous group of automorphisms only if \mathcal{C} is a sphere with one or two punctures, where we may have either ordinary double poles or higher ones. Except for these special cases, (8.2) may be satisfied only by setting

$$(8.3) \quad \phi_2 \Big|_{\substack{\text{conformal} \\ \text{point}}} = 0.$$

Moreover, this should be achieved by finite deformation of the theory (otherwise, we would simply have an asymptotically free theory, which is

conformal at infinite distance in Coulomb branch). For poles higher than order 2, there will always be some Coulomb branch vevs which correspond to residues of the poles, and using the metric $\int |\delta\lambda_{SW}|^2$ we find this leads to infinite distance, where λ_{SW} denotes the Seiberg-Witten differential ydx . The regular poles can be set to zero by setting the corresponding mass to zero. Thus, the only superconformal $\mathcal{N} = 2$ theories associated to surfaces with $g > 0$ or $g = 0$ with at least three punctures (ordinary or otherwise) are the regular Gaiotto ones without higher order poles.

The sphere with a single puncture is a well-defined $\mathcal{N} = 2$ quiver theory only if we have a pole of order $p \geq 6$ — corresponding to a disk with $(p - 2)$ marked points *i.e.* a $(p - 2)$ -gon. This corresponds to the A_{p-5} Argyres–Douglas models which are known to have a superconformal point.

Likewise, the sphere with an ordinary puncture and one pole of order p is associated to D_p Argyres–Douglas theory which also has a superconformal point.

Instead, the sphere with two higher order poles is associated to an annulus with marked points on both boundaries. This theory is just asymptotically free: special instances are $SU(2)$ gauge theory coupled to $N_f = 0, 1, 2$ fundamental flavors. The fact that they are not superconformal is particularly evident from the $4d/2d$ perspective: they correspond to the $2d$ models

$$(8.4) \quad W(X) = e^{nX} + e^{-mX},$$

which has no continuous symmetry since the approximate $U(1)_R$ symmetries around the north and south poles do not agree in the intermediate region. In the language of ref. [1], this corresponds to a unipotent non-semisimple $2d$ monodromy.

It remains to discuss the 11 exceptional models. The models associated to the ordinary E_6, E_7, E_8 Dynkin quivers are a kind of exceptional Argyres–Douglas theories, already studied in [2]. They are known to have a conformal point.

The $\mathcal{N} = 2$ theories associated to affine and elliptic E –type Dynkin quivers are best studied by the Type IIB geometrical engineering described in section 7. Then the conformal $U(1)_R$ should arise from a $U(1)$ symmetry of the local Calabi–Yau hypersurface which acts on the holomorphic 3–form Ω as $\Omega \rightarrow e^{i\theta} \Omega$. In this way we see that the affine \widehat{E} –models have no conformal point, and thus are UV asymptotically free. This was to be expected, given that the affine \widehat{A} – and affine \widehat{D} –models are UV asymptotically free, and affine $\widehat{A}\widehat{D}\widehat{E}$ models form a family with uniform properties.

The elliptic $\widehat{\tilde{E}}$ –models, instead, have a conformal regime which was studied in detail in ref. [2] and reviewed in §. 7. Notice that the only other elliptic Dynkin diagram which gives a mutation–finite quiver, namely $\widehat{\tilde{D}}_4$, corresponds to $SU(2)$ with $N_f = 4$, and it is also UV superconformal.

Finally X_7 has a conformal limit, corresponding to $m \rightarrow 0$, as we may check from its Lagrangian formulation. In this limit the theory coincides

with the $g = 2$ Gaiotto model, so — as a conformal theory — it is already in the surface list, and we don't get a new model. X_6 is not UV conformal.

In conclusion, the full list of complete $\mathcal{N} = 2$ theories which have a UV superconformal limit are

- Gaiotto theories;
- ADE Argyres–Douglas theories;
- elliptic $\widehat{\tilde{E}}_6, \widehat{\tilde{E}}_7, \widehat{\tilde{E}}_8$ theories;
- X_7 .

8.2. Proof of the graphical rule. Finally, we wish to show that the above list just coincide with the set of all normalized mutation–finite quivers which are either vector–less or maximal.

The rule holds for the 11 exceptional classes by inspection: affine \widehat{E}_r and X_6 are neither maximal nor vector–free, and are non–conformal; the others are either vector–free, E_r , or maximal, $\widehat{\tilde{E}}_r, X_7$, and are conformal.

Then to prove the graphical rule it is enough to show that a *normalized* (non–exceptional) mutation–finite quiver which is maximal is the triangulation of a surface without boundaries (that is with only ordinary punctures).

Indeed, if a surface \mathcal{C} has a boundary component S^1 , we may glue to it another surface \mathcal{C}' with an S^1 boundary, and hence \mathcal{C} is not maximal. More precisely, at the level of block decomposition of the adjacency quiver, the S^1 boundary component corresponds to one of the following three possibilities²⁰: *i*) a free unpaired white node \circ ; *ii*) a block of type II; *iii*) a block of type III. To normalize the quiver, we replace the blocks of type III with a type II and a type I with arrows pointing in opposite directions, so case *iii*) is eliminated by the normalization assumption.

In case *i*) we may glue another block at the unpaired \circ node and the quiver is not maximal. In case *ii*) we replace the block II by a block III oriented in the same way, and the quiver is not maximal.

On the other hand, a surface without boundaries (corresponding to a Hitchin system with only regular singularities) has an adjacency quiver composed by blocks of type II, IV and V with all the white nodes \circ paired up. There is no possibility to attach extra nodes while getting a graph which is still an adjacency quiver. Finally, we have to check that no adjacency quiver of a surface with no–boundary is a subquiver of an exceptional one. This is true by inspection.

9. Physical properties of gauging $\mathcal{N} = 2$ D –sub-systems

In this paper we have found compelling evidence that many complete $\mathcal{N} = 2$ systems are best understood as a number of $SU(2)$ gauge sectors coupled to some $\mathcal{N} = 2$ systems with $SU(2)$ symmetry. In this section we

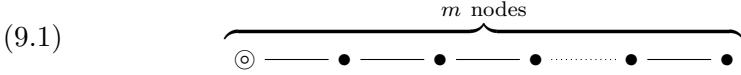
²⁰ In case of a boundary with many marked points, we have typically many of the following quiver blocks, and hence many possible extensions of the quiver which keep it mutation–finite.

discuss some physical properties of the *very simplest* examples of such $\mathcal{N} = 2$ systems, consisting of gauging $\mathcal{N} = 2$ D -subsystems.

9.1. β -functions of D -systems. We first focus our attention on the $\mathcal{N} = 2$ theories associated to the affine quivers $\widehat{A}(m, n)$ with²¹ $m, n \geq 1$, \widehat{D}_{n-1} and \widehat{E}_r . We have seen that they are mutation equivalent, respectively, to figure 6, eqn.(6.60), and eqn.(7.9) with (n_1, n_2, n_3) as in table 1. They are naturally interpreted as $SU(2)$ coupled to

- one D_{m+1} -system for $\widehat{A}(m+1, 1)$;
- one D_{m+1} -system and one $D_{m'+1}$ -system for $\widehat{A}(m+1, m'+1)$;
- two fundamental hypermultiplets and one D_{m+1} -system for \widehat{D}_{m+3} ;
- one fundamental hypermultiplet and two D_3 -systems for \widehat{E}_6 ;
- one fundamental hypermultiplet, a D_3 -system, and a D_4 -system for \widehat{E}_7 ;
- one fundamental hypermultiplet, a D_3 -system, and a D_5 -system for \widehat{E}_8 ;

Note that as discussed in §.6.4.5, a D_{m+1} system couples to an $SU(2)$ Kronecker quiver by the attachment of the subquiver



(the orientation being irrelevant) having a special node, $\textcircled{1}$, where we attach the oriented triangle coupling the subquiver to the Kronecker one. For $m = 1$, we get back the usual hypermultiplet; to get more elegant formulae, it is convenient to extend the definition to $m = 0$, corresponding to the empty $\mathcal{N} = 2$ system.

As we saw in the previous section, all affine complete $\mathcal{N} = 2$ theories are asymptotically free. Hence the β -function of the $SU(2)$ has to be negative. Comparing with the above list, we get that *the contribution to the $SU(2)$ β -function from the coupling to an D_{m+1} $\mathcal{N} = 2$ theory is less than twice the contribution from a fundamental hypermultiplet*.

To get a precise formula for the β -function contribution of a D_{m+1} system we have to look at the *elliptic* complete $\mathcal{N} = 2$ models: $\widehat{\widehat{D}}_4$, $\widehat{\widehat{E}}_6$, $\widehat{\widehat{E}}_7$, $\widehat{\widehat{E}}_8$, which may also be described as $SU(2)$ coupled to D_{m+1} -system (see figure 1 on page 38). These theories are UV superconformal, and hence have a vanishing β -function. These results are reproduced by taking the β -function of the D_{m+1} system to be

$$(9.2) \quad 2 \left(1 - \frac{1}{m+1} \right)$$

²¹ The affine quiver $\widehat{A}(m, 0)$ is mutation equivalent to the finite Dynkin quiver D_m .

times that of a fundamental hypermultiplet. Note that this formula gives the correct result for $m = 0$ and $m = 1$, and it is always less than 2, as required.

Eqn.(9.2) has a simple heuristic interpretation in terms of the string world-sheet theory. $SU(2)$ coupled to three D_{m+1} -system, is engineered by Type IIB on the hypersurface (7.13), and the world-sheet theory is the Landau–Ginzburg model with the RHS of (7.13) as superpotential *with Liouville superfield dependent couplings* (in order to get $2d$ superconformal invariance) [45]. The world-sheet Liouville couplings reflect the $4d$ β -function. These couplings, and hence the β -function, are proportional to $(\hat{c} - 1)$. In particular

$$(9.3) \quad \lambda X_1 X_2 X_3 \rightarrow \lambda_0 e^{(1-\hat{c})\phi} X_1 X_2 X_3$$

λ being the coupling which, in the conformal case, encodes the modulus of the torus τ . Let b the coefficient of the $SU(2)$ β -function (normalized so that the contribution of a fundamental hypermultiplet is +1); then

$$(9.4) \quad b = -4 + 2 \sum_{i=1}^3 \left(1 - \frac{1}{m_i + 1} \right) \equiv \sum_{i=1}^3 \left(1 - \frac{2}{m_i + 1} \right) - 1 = \hat{c} - 1,$$

and so eqn.(9.2) is suggestive of another manifestation of the general $4d/2d$ correspondence.

9.2. \widehat{D} -systems and new $\mathcal{N} = 2$ dualities. Similar arguments may be applied to other basic $\mathcal{N} = 2$ systems which are conveniently used as building blocks of more complex theories. *E.g.* the \widehat{D}_{m+1} theory has an $SU(2) \times SU(2)$ symmetry, that can be gauged, corresponding the two double ends. As discussed in §.6.4.5 this leads to attaching the subquiver

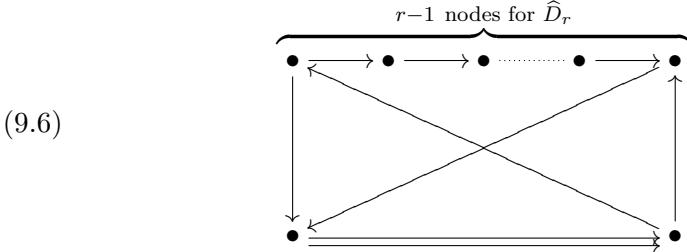
$$(9.5) \quad \text{m nodes} \overbrace{\circlearrowleft \cdots \bullet \cdots \bullet \cdots \bullet \cdots \bullet \cdots \bullet \cdots \circlearrowright}^{\text{m nodes}}$$

to two Kronecker systems one on each end. Since we may replace anyone of the type I blocks in figure (6.4.5) with a type II block, any quiver containing the subquiver (6.4.5) is not maximal, and hence corresponds to an UV asymptotically free theory. A naive analogy with the previous case would lead to the wrong conclusion that the contribution from an $m \geq 2$ such system to the $SU(2)$ β -functions of both SYM coupled at the nodes \circlearrowleft is less than the one from a bi-fundamental hypermultiplet. This is *not* correct: The contribution to the β -function of the gauging $SU(2)$'s is *equal* to that of a bi-fundamental hypermultiplet. Nevertheless the resulting model cannot be superconformal simply because the \widehat{D}_{m+1} sector is by itself asymptotically free, and the couplings which have negative β -functions are the ones inside the system described by the subquiver (9.5). Indeed, we have a dual picture of this $\mathcal{N} = 2$ theory: Up to mutation, the quiver \widehat{D}_{m+1} may be taken in the form (6.60) which is naturally interpreted as an $SU(2)$ SYM coupled to two

fundamental hypermultiplets and one D -system. The $SU(2) \times SU(2)$ flavor symmetry of the \widehat{D}_{m+1} system may be interpreted simply as the usual flavor symmetry of the two fundamental hypermultiplets. So, we may think of a model where the $SU(2) \times SU(2)$ symmetry of a \widehat{D}_{m+1} theory is gauged as a theory with one more gauge group, where the extra group gauges a pair of bi-fundamental half-hypermultiplets and a D -system.

A new kind of $\mathcal{N} = 2$ duality is obtained from the mutation-equivalence $\Gamma(n, m) \sim \Gamma(m, n)$ for the triangulation of a punctured annulus with (n, m) marking on the boundaries. In term of quivers, this may be seen as an $SU(2)$ which gauges the $SU(2)$ symmetry of a D -system and one of the two $SU(2)$ factor subgroups of the $SU(2) \times SU(2)$ symmetry of a \widehat{D} -system, while the other subgroup remains as a global symmetry (corresponding to the type III block in the $\Gamma(n, m)$ quiver). Again we have a duality interchanging the ranks of the two systems. This theory may be understood as an $SU(2)^2$ gauge theory where both $SU(2)$'s gauge the same half-bifundamental, then each of them gauge a D -system, and one of the two $SU(2)$'s also gauges a fundamental hypermultiplet.

The two \circlearrowleft nodes of the subquiver (9.5) may be gauged by two distinct $SU(2)$ SYM, or the same SYM may gauge the diagonal $SU(2)$ subgroup of the $SU(2) \times SU(2)$ of the \widehat{D} -systems. In the last case we get the quiver



$r = 2$ gives $\mathcal{N} = 2^*$, $r = 3$ the unique-quiver model of section 6.3.4, and more generally, the generalized Gaiotto theory associated to a torus with a boundary having $r - 2$ marks.

Note that, since $\widehat{D}_3 \sim \widehat{A}_3(2, 2)$ the ‘remarkable’ theory of §. 6.3.4 may be interpreted as $SU(2)$ SYM gauging a $SU(2)$ subgroup of the $SU(2) \times SU(2)$ flavor symmetry of $SU(2)$ SQCD with $N_f = 2$. This gives a Lagrangian formulation of the unique-quiver model of section 6.3.4, confirming that it is an asymptotic free theory without flavor charges.

9.3. BPS spectrum of $SU(2)$ SYM coupled to D -systems. In this section we determine the BPS spectra of $SU(2)$ SYM coupled to one, two, or three D_r -systems. With the exception of the elliptic models²² $\widehat{\tilde{E}}_r$, these theories are asymptotically free and have an affine quiver of the form

²² The strong coupling BPS spectrum of the elliptic models is described in §. It is likely that they have also ‘weak coupling’ chambers with BPS vector multiplets.

$\widehat{A}(m, n)$ ($m, n \geq 1$), \widehat{D}_r or \widehat{E}_r . The first $\mathcal{N} = 2$ models in these series are just $SU(2)$ SQCD with $N_f \leq 3$.

As in section 2.2 the BPS spectrum is determined by the Kac–Moody representation theory.

We have a strong coupling BPS chamber with only hypermultiplet dyons, one for each simple root of corresponding Kac–Moody algebra with the charge vector

$$(9.7) \quad \alpha_i = (0, \dots, 0, 1, 0, \dots, 0)$$

in the basis of the charge lattice Γ in which the quiver has the standard affine Dynkin graph form.

Then we have a weak coupling chamber with an infinite BPS dyon spectrum consisting of hypermultiplet of charge vector

$$(9.8) \quad \sum_i n_i \alpha_i \in \Delta_+^{\text{re}}$$

and a BPS vector multiplet of charge vector equal the indivisible imaginary root

$$(9.9) \quad \delta = \sum_i a_i \alpha_i,$$

where a_i are the Dynkin weights, equal, by the McKay correspondence, to the dimensions of the irreducible representations of the corresponding finite subgroup of $SU(2)$.

10. Conclusions

Appendix A. Strong coupling spectra of affine quiver models

In this appendix we show that the strong coupled spectrum of any $\mathcal{N} = 2$ theory having an affine quiver without oriented cycles is given by one hypermultiplet per simple root.

The basic point about affine quivers without oriented loops is the existence of frieze sequence [23]. In particular, we may number the vertices from 1 to D in such a way that each vertex i is a source in the full subquiver of vertices $1, \dots, i$. Let $\tilde{\mu}_k$ be the combination of the elementary quiver mutation, μ_k , with the corresponding change of basis in Γ as defined in equations (6.2)(6.3) of [2] (we adopt the same conventions). Then if the product

$$(A.1) \quad \tilde{\mu}_1 \circ \tilde{\mu}_2 \circ \dots \circ \tilde{\mu}_D,$$

acts on the quantum torus algebra \mathbb{T}_Γ as the inversion I , then the corresponding product of elementary quantum cluster mutations

$$(A.2) \quad \mathbb{K}(q) = \mathcal{Q}_1 \mathcal{Q}_2 \dots \mathcal{Q}_D,$$

is the quantum half-monodromy (the *omnipop* in the language of [24]) from which we may read the BPS spectrum²³ in the corresponding chamber (which is the strong coupled one) [cnv,ceclct].

The above identity follows from the simple observation that the vertex i is a source in the mutated quiver

$$(A.3) \quad Q_i = \mu_{i+1} \circ \mu_{i+2} \circ \cdots \circ \mu_D(Q),$$

so the i -th transformation $\tilde{\mu}_i$ in the sequence (A.1) just inverts $X_i \rightarrow X_i^{-1}$ while keeping invariant X_j for $j \neq i$. Thus the effect of the product (A.1) is just to invert all quantum cluster variables, that is the product in eqn.(A.1) is I .

The formula (A.2) also determines the BPS phase cyclic order in terms of the affine quiver orientation.

Appendix B. Details on some Landau–Ginzburg models

In this appendix we present some details on the two-dimensional computations for some of the Landau–Ginzburg models mentioned in the main body of the paper.

B.1. The second form of $N_f = 2$. This realization of $N_f = 2$ may be set in relation with the LG model

$$(B.1) \quad W(X) = e^X + \frac{1}{(1 - e^{-X})^2}.$$

This $2d$ theory has four classical vacua. One at $e^{-X} = \infty$, and the other three at the at e^{-X} equal to the three roots of

$$(B.2) \quad y^3 - y^2 + 3y - 1 = 0,$$

which has one *positive* real root $r = e^{-X_r}$, $X_r > 1$, and a pair of complex conjugate ones $\rho, \bar{\rho}$. The critical values are

$$(B.3) \quad W_\infty = 0, \quad W_r \text{ real positive } \approx 3.17748$$

$$(B.4) \quad W_\rho = (W_{\bar{\rho}})^* \text{ complex with negative real part.}$$

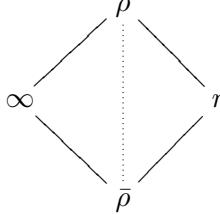
We know the following facts about the BPS quiver:

- should be connected and compatible with $\hat{c}_{uv} = 1$. Indeed, were it not connected, the connected components will have at most three nodes, and all such theories are already classified;
- the numbers of BPS states connecting ρ with ∞ (resp. r) is the same as the number of states connecting $\bar{\rho}$ with ∞ (resp. r) since they are related by complex conjugation;
- there are no solitons connecting r and ∞ .

²³ As well as the BPS phase order.

Then the graph underlying the quiver must have the form

(B.5)



where the dashed line means that there may be or not a soliton connecting the two complex vacua. We also know that the *orientation* of the arrows should be invariant under reflection with respect to the horizontal axis (*i.e.* under complex conjugation). Finally, we know that the direction of the arrows should be consistent with $\hat{c} = 1$, which requires that any proper sub-quiver should be a minimal model one. This leaves us with three possible BPS quivers which are all mutation-equivalent to $\widehat{A}_3(2, 2)$.

B.2. $N_f = 3$. One has

$$(B.6) \quad W' = 2e^{2X} \frac{(e^X - 1)^3 - 1}{(e^X - 1)^3}$$

so that we have two vacua at $X = -\infty$, and three vacua for $e^X = 1 + \varrho$, where ϱ is a primitive third root of 1. The critical values are 0 for the vacua at ∞ , and

$$(B.7) \quad \begin{aligned} W(e^X - 1 = \varrho) &= (1 + \varrho)^2 \frac{\varrho^2 + 1}{\varrho^2} = (1 + \varrho)^2(1 + \varrho^{-2}) = (1 + \varrho)^3 \\ &= \begin{cases} 8 & \varrho = 1 \\ (-\varrho^2)^3 \equiv -1 & \varrho \neq 1. \end{cases} \end{aligned}$$

Thus, all critical values are real (and hence aligned). There are no solitons between the two vacua at infinity, nor between the two vacua at $e^X = 1 + e^{\pm 2\pi i/3}$. Moreover, complex conjugation exchanges these last two vacua, and hence the number of soliton from each of these two vacua and the other vacua are equal.

Setting $y = e^{-X}$, the equation $W(X) = w$ becomes the quartic equation

$$(B.8) \quad y^4 - 2y^3 + (1 + 2/w)y^2 + 2y/w - 1/w = 0$$

whose discriminant is

$$(B.9) \quad -16(w - 8)(w + 1)^2/w^5.$$

Consider the solitons between infinity and the vacuum 0. In the W -plane they corresponds to the segment $0 \leq w \leq 8$ on the *real* axis. For $w \sim 0$ real positive, (B.8) gives $y \sim \zeta w^{-1/4}$, where ζ is a fourth-root of 1. Thus, for $w \sim 0^+$ we have one real positive, one real negative, and a pair of complex conjugate roots. Given that the constant term of (B.8) never vanishes, this configuration of roots (one positive, one negative, a pair of conjugate ones)

will persists as we move w along the real axis until we get at the first zero of the discriminant at $w = 8$. Here the two complex roots come together and become *real*. Indeed, at $w = 8$ the roots of (B.8) are

$$(B.10) \quad y = 1/2, 1/2, (1 - \sqrt{3})/2, (1 + \sqrt{3})/2,$$

and the two solutions which becomes purely imaginary as $w \rightarrow 0^+$, both have limit $y = 1/2$ as $w \rightarrow 8$. The other two roots at $w = 8$ corresponds to the two real roots at $w \sim 0$, respectively negative and positive.

$y = 1/2$, corresponds to $e^X = 2$, that is to the vacuum 0. Therefore, the two imaginary roots of (B.8) over the segment $0 \leq w \leq 8$ in the W -plane are precisely two BPS states connecting vacuum 0 to, respectively, ∞_1 and ∞_2 , where these two vacua correspond to $e^X = \mp i w^{1/4}$, as $w \rightarrow 0$.

In the W -plane, the solitons from infinity to $e^X = 1 + e^{2\pi i/3}$ correspond to the segment $-1 \leq w \leq 0$ on the real axis. For $w \sim 0$ real and *negative* we have from (B.8) $y \sim \zeta |w|^{-1/4}$ where ζ is a fourth-root of -1 . Thus for $w \sim 0^-$ we have two pairs of complex conjugate roots with phases $\pm i$ and, respectively, $e^{\pm i\pi/4}$.

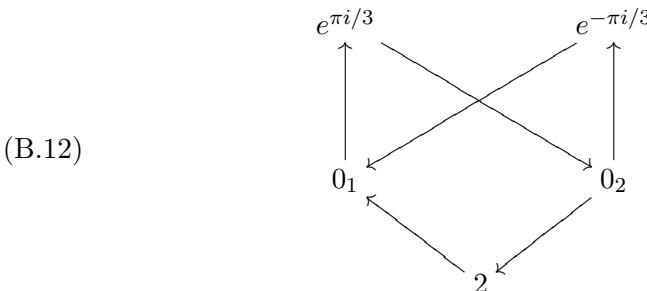
As we decrease w from 0 to -1 these pairs of complex roots will not mix, since the discriminant is not zero, until we reach $w = -1$ where the discriminant has a *double* zero. There the two complex pair — while remaining complex — gets together. Indeed, the roots of equation (B.8) with $t = -1$ are

$$(B.11) \quad y = e^{\pi i/3}, e^{\pi i/3}, e^{-\pi i/3}, e^{-\pi i/3}.$$

One has $e^{\mp\pi i/3} = 1 + e^{\mp 2\pi i/3}$. Hence two of the soliton starting from infinity will reach each complex classical vacua.

Finally, the solitons between $e^X = 2$ and $e^X = 1 + \varrho$ correspond to the segment $-1 \leq w \leq 0$ on the real axis. But these all passes through infinity. So no soliton here.

In conclusion, the above results suggest the following form for the quiver (where the nodes are labelled by the values of e^X)



which is mutation-equivalent to \widehat{D}_4 .

B.3. LG with $W(X) = \wp'(X)$. Let $\wp(z)$ be the Weierstrass function

$$(B.13) \quad (\wp')^2 = 4\wp^3 - g_2\wp - g_3.$$

where the cubic polynomial in the RHS has non-vanishing determinant $\Delta \neq 0$. We consider a LG model with the field X taking value on the corresponding torus and superpotential $W(X) = \wp'(X)$. The vacuum condition is

$$(B.14) \quad 0 = W'(X) = 6\wp(X)^2 - \frac{1}{2}g_2.$$

The function in the RHS has a pole of order 4 at the origin, and hence four zeros.

Lemma. *For $g_2 \neq 0$, all four classical vacua are massive (and hence distinct). Between any two vacua, the absolute number of BPS solitons is either 1 or 2.*

PROOF. Indeed, $W'' = 12\wp(X)\wp'(X)$. At a vacuum $\wp(X) = \pm\sqrt{g_2/12}$, and hence $\wp(X) \neq 0$ is non-zero. Then, in order to have $W'(X) = W''(X) = 0$, we must have $\wp'(X) = 0$ and hence

$$(B.15) \quad 0 = 4\wp^4 - g_2\wp - g_3 = \pm\sqrt{g_2/12}(4g_2/12 - g_2) - g_3$$

or

$$(B.16) \quad 0 = g_3^2 - \frac{1}{12} \cdot \frac{4}{9} \cdot g_2^3 = -\frac{\Delta}{432} \neq 0,$$

which is absurd. Then the four vacua are $\pm X_\pm$ where $\wp(X_\pm) = \pm\sqrt{g_2/12}$.

Let $W_\pm = \wp'(X_\pm)$. Consider the elliptic functions $F_{\epsilon,\epsilon'}(X) = \wp'(X) - \epsilon W_{\epsilon'}$, where $\epsilon, \epsilon' = \pm$. These meromorphic functions have a pole of order 3 at the origin, and hence should have three zeros on the torus whose sum must give zero. One the other hand,

$$(B.17) \quad F_{\epsilon,\epsilon'}(\epsilon X_{\epsilon'}) = 0, \quad F'_{\epsilon,\epsilon'}(\epsilon X_{\epsilon'}) \equiv \wp''(\epsilon X_{\epsilon'}) = 0,$$

and hence $F_{\epsilon,\epsilon'}(X)$ has a *double* zero at $\epsilon X_{\epsilon'}$.

Consider now the inverse image of the segment in W plane between the points $\epsilon W_{\epsilon'}$ and $\tilde{\epsilon} W_{\tilde{\epsilon}'}$; it may be written as

$$(B.18) \quad (1-t)F_{\epsilon,\epsilon'} + tF_{\tilde{\epsilon},\tilde{\epsilon}'} = 0 \quad 0 \leq t \leq 1.$$

For each t in the open interval $0 < t < 1$, we have three values of X (modulo periods) which satisfy this equation. Moreover, these values are all distinct, except at $t = 0, 1$, where two of the three values will go to the critical point $\epsilon X_{\epsilon'}$ and, respectively, to $\tilde{\epsilon} X_{\tilde{\epsilon}'}$ while the third root approaches at $-2\epsilon X_{\epsilon'}$ and $-2\tilde{\epsilon} X_{\tilde{\epsilon}'}$, respectively. Let $X_{(1)}(t), X_{(2)}(t)$ be the two solutions which for $t = 0$ go to the classical vacuum $\epsilon X_{\epsilon'}$. Two things may happen: either both $X_{(1)}(t), X_{(2)}(t)$ go to $\tilde{\epsilon} X_{\tilde{\epsilon}'}$ as $t \rightarrow 1$, or one of the two go to the third root $-2\tilde{\epsilon} X_{\tilde{\epsilon}'}$ while the other one will necessarily go to $\tilde{\epsilon} X_{\tilde{\epsilon}'}$. \square

To simplify the analysis, we consider a special case with enhanced symmetry, namely a lemniscatic (square) torus with periods $(1, i)$, corresponding to $g_3 = 0$, $g_2 = \Gamma(1/4)^8/16\pi^2$. Then $\wp'(iX) = i\wp'(X)$, and the model has

a \mathbb{Z}_4 symmetry, $X \rightarrow iX$, under which the four (distinct) vacua form an orbit. The four vacua are at

$$(B.19) \quad X_k = i^{k-2} \left(\frac{1}{2} + i\alpha \right), \quad k = 1, 2, 3, 4, \quad \alpha \approx 0.1988783 \in \mathbb{R}.$$

The critical values form a square in W -plane with vertices at

$$(B.20) \quad W(X_k) = i^{k-1} a, \quad k = 1, 2, 3, 4, \quad a \approx 22.3682 \in \mathbb{R}.$$

By the \mathbb{Z}_4 symmetry, it is enough to determine the number of BPS states along a side and a diagonal of this square. Consider the diagonal corresponding to the segment along the imaginary axis from $-ia$ to $+ia$; a diagonal soliton is a curve on the torus connecting $1/2 - i\alpha$ to $1/2 + i\alpha$ which maps to this segment in the W -plane. Let the X -plane be the universal cover of the torus. Along the straight-line $1/2 + i\mathbb{R}$ the function $\wp'(X)$ is purely imaginary, so the segment in the X -plane connecting $1/2 - i\alpha$ to $1/2 + i\alpha$ is mapped into the diagonal of the square, and hence it is a soliton. Likewise, the segment in the X plane from $1/2 - i\alpha$ to $1/2 - i(1 - \alpha)$ is also a segment between the same two vacua on the torus. So there are at least two solitons along each diagonal; since there cannot be more than two by the lemma, we conclude that along the diagonal we have precisely two solitons.

It remains to determine the number μ of solitons along the sides of the square. We have $|\mu| = 1, 2$ by the lemma. In order to get μ , we may use the general classification of \mathbb{Z}_4 symmetric models in [1]. Eqn.(8.5) of ref. [1] implies that

$$(B.21) \quad Q(z) \equiv z^4 + \mu z^3 \pm 2z^2 + (-1)^{q+1} \mu z + (-1)^{q+1}$$

should be a product of cyclotomic polynomials for some choice of signs \pm and $(-1)^q$. The solutions to this condition with $\mu = \pm 1, \pm 2$ are

$$(B.22) \quad \Phi_3(z) \Phi_4(z) = z^4 + z^3 + 2z^2 + z + 1$$

$$(B.23) \quad \Phi_6(z) \Phi_4(z) = z^4 - z^3 + 2z^2 - z + 1$$

$$(B.24) \quad \Phi_4(z) \Phi_1(z)^2 = z^4 - 2z^3 + 2z^2 - 2z + 1$$

$$(B.25) \quad \Phi_4(z) \Phi_2(z)^2 = z^4 + 2z^3 + 2z^2 + 2z + 1$$

which also implies $(-1)^q = -1$. Then eqn.(8.4) of ref. [1] gives for the characteristic polynomial of the $2d$ monodromy M

$$(B.26) \quad \det[z - M] = \begin{cases} \Phi_3(-z) \Phi_1(-z)^2 & |\mu| = 1 \\ \Phi_1(-z)^4 & |\mu| = 2. \end{cases}$$

The second case corresponds to the four point correlation of the Ising model. The spectrum of M is not compatible with a unitary theory with $\hat{c}_{uv} \leq 1$.

The first case of eqn.(B.26) is perfectly compatible with a AF model with $\hat{c}_{uv} = 1$, and having four chiral primary operators of dimension in the UV 0, $1/3$, $2/3$ and 1. Since the two allowed deformations of $W(X)$, namely $\wp(X)$ and $\zeta(X)$, are expected to have UV dimensions $2/3$ and $1/3$, respectively, this solution must correspond to the model $W(X) = \wp'(X)$.

Then we learn that along the sides of the critical square in the W -plane we have just one soliton, $|\mu| = 1$.

The quiver

We write the elements S_θ of the Stokes group corresponding to the four BPS rays $e^{i\theta}$ in the lower half plane, as borrowed from section 8 of ref.[classification] for the relevant \mathbb{Z}_4 -symmetric model²⁴

$$(B.27) \quad S_0 = 1 - 2 E_{3,1} \quad S_{-\pi/4} = 1 - E_{2,1} + E_{3,4}$$

$$(B.28) \quad S_{-\pi/2} = 1 + 2 E_{2,4} \quad S_{-3\pi/4} = 1 - E_{1,4} - E_{2,3}$$

where $(E_{ij})_{kl}$ is the matrix which is 1 for $k = i, l = j$ and zero otherwise. One has (the conventions of [classification] correspond to taking the product of the S_θ in the *clockwise* order)

$$(B.29) \quad S \equiv S_{-3\pi/4} S_{-\pi/2} S_{-\pi/4} S_0 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 1 & 1 & -1 & 1 \\ -2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

and then

$$(B.30) \quad B \equiv S - S^t = \begin{pmatrix} 0 & -1 & 2 & -1 \\ 1 & 0 & -1 & 1 \\ -2 & 1 & 0 & 1 \\ 1 & -1 & -1 & 0 \end{pmatrix},$$

which corresponds to the quiver

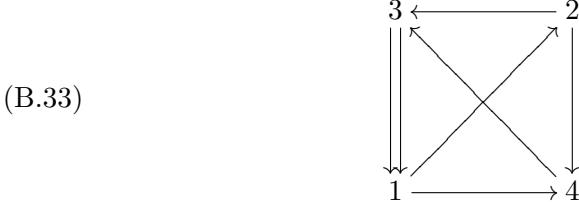


which is the one associated to the unique ideal triangulation of a torus with a boundary having a marked point. If we change the half-plane used to define B , nothing is going to change: in fact by \mathbb{Z}_4 symmetry, we have only to check the rotation of the half-plane by $-\pi/4$; this amounts to replacing

$$(B.32) \quad S \rightarrow S' = I_3 (S_0^{-1})^t S S_0^{-1} I_3$$

²⁴ With respect to that reference, we change the sign to vacua 3 and 4, which is natural since the topological metric η changes sign as $X \leftrightarrow -X$.

(where $I_3 = \text{diag}(1, 1, -1, 1)$ just a vacuum sign redefinition to reestablish the correct conventions). Then $B' = S' - (S')^t$ gives the quiver



which is the same as before, up to a relabeling of the nodes.

References

- [1] S. Cecotti and C. Vafa, “On classification of $\mathcal{N} = 2$ supersymmetric theories,” *Commun. Math. Phys.* **158** (1993) 569–644, <http://www.arXiv.org/abs/hep-th/9211097>.
- [2] S. Cecotti, A. Neitzke and C. Vafa, “ R -Twisting and $4d/2d$ correspondences,” <http://www.arXiv.org/abs/1006.3435>.
- [3] S. Fomin, M. Shapiro and D. Thurston, “Cluster algebras and triangulated surfaces. Part I: cluster complexes”, <http://www.arXiv.org/abs/math/0608367>.
- [4] H. Derksen and R. Owen, “New graphs of finite mutation type”, <http://www.arXiv.org/abs/0804.0787>.
- [5] A. Felikson and M. Shapiro and P. Tumarkin, “Skew-symmetric cluster algebras of finite mutation type”, <http://www.arXiv.org/abs/0811.1703>.
- [6] M.R. Douglas and G. Moore, “ D -branes, quivers, and ALE instantons”, <http://www.arXiv.org/abs/hep-th/9603167>.
- [7] F. Denef, “Quantum quivers and Hall/Holes Halos,” *JHEP* 0210 (2002) 023, <http://www.arXiv.org/abs/hep-th/0206072>.
- [8] M. Kontsevich and Y. Soibelman, “Stability structures, motivic Donaldson-Thomas invariants and cluster transformations,” <http://www.arXiv.org/abs/0811.2435>.
- [9] D. Gaiotto, G. W. Moore, and A. Neitzke, “Four-dimensional wall-crossing via three-dimensional field theory,” <http://www.arXiv.org/abs/0807.4723>.
- [10] T. Dimofte and S. Gukov, “Refined, Motivic, and Quantum,” *Lett. Math. Phys.* **91** (2010) 1, <http://www.arXiv.org/abs/0904.1420>.
- [11] T. Dimofte, S. Gukov, and Y. Soibelman, “Quantum Wall Crossing in $N=2$ Gauge Theories,” <http://www.arXiv.org/abs/0912.1346>.
- [12] S. Cecotti and C. Vafa, “BPS Wall Crossing and Topological Strings,” <http://www.arXiv.org/abs/0910.2615>.
- [13] V.G. Kac “Infinite roots systems, representations of graphs and invariant theory,” *Inventiones mathematicae* **56** 57–92 (1980).
- [14] A.D. King, “Moduli of representations of finite-dimensional algebras,” *Quart. J. Mat. Oxford Ser (2)* **45** (1994) 180, 515.
- [15] P. Gabriel and A.V. Roiter, *Representations of finite-dimensional algebras*, Encyclopaedia of Mathematical Sciences, ALGEBRA VIII, vol. 73, A.I. Kostrikin and I.R. Shafarevich Eds., Springer-Verlag (1991).
- [16] I. Assem, D. Simson, and A. Skowroński, *Elements of the representation theory of associative algebras. Vol. 1*, vol. 65 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 2006. Techniques of representation theory.
- [17] M. Auslander, S. O., and I. Reiten, *Representation theory of Artin algebras*, vol. 36 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1995.

- [18] M.R. Douglas, B. Fiol and C. Römelsberger, “Stability and BPS branes,” [hep-th/0002037](#).
- [19] M.R. Douglas, B. Fiol and C. Römelsberger, “The spectrum of BPS branes on a noncompact Calabi–Yau,” [hep-th/0003263](#).
- [20] B. Fiol and M. Marino, “BPS states and algebras from quivers,” [hep-th/0006189](#).
- [21] B. Fiol, “The BPS spectrum of $N=2$ $SU(N)$ SYM and parton branes,” [hep-th/0012079](#).
- [22] B. Keller, “On the cluster theory and quantum dilogarithm identities,” <http://www.arXiv.org/abs/1102.4148>.
- [23] B. Keller and S. Scherotzke, “Linear recurrence relations for cluster variables of affine quivers,” [1000.0613](#).
- [24] D. Gaiotto, G. W. Moore, and A. Neitzke, “Wall-crossing, Hitchin Systems, and the WKB Approximation,” <http://www.arXiv.org/abs/0907.3987>.
- [25] D. Gaiotto, private communication.
- [26] A. D. Shapere and C. Vafa, “BPS structure of Argyres-Douglas superconformal theories,” <http://www.arXiv.org/abs/hep-th/9910182>.
- [27] S. Fomin and A. Zelevinsky, “Cluster algebras IV: Coefficients,” *Compos. Math.* **143** (2007) 112–164 [math.RA/0602259](#).
- [28] V.G. Kac, *Infinite dimensional Lie algebras*, Third edition, Cambridge University press, 1990.
- [29] M. Wakimoto, *Infinite-dimensional Lie algebras*, Translations of Mathematical Monographs vol 195, AMS, 1999.
- [30] N. Seiberg and E. Witten, “Electric-magnetic duality, monopole condensation, and confinement in $\mathcal{N} = 2$ supersymmetric Yang-Mills theory,” *Nucl. Phys.* **B426** (1994) 19–52, <http://www.arXiv.org/abs/hep-th/9407087>.
- [31] N. Seiberg and E. Witten, “Monopoles, duality and chiral symmetric breaking in $N = 2$ supersymmetric QCD,” *Nucl. Phys.* **B431** (1994) 485–550, <http://www.arXiv.org/abs/hep-th/9408099>.
- [32] J.A. Minahan and D. Nemeschany, “An $N = 2$ superconformal fixed point with E_6 global symmetry,” *Nucl. Phys.* **B482** (1996) 142–152 [hep-th/9608047](#).
- [33] J.A. Minahan and D. Nemeschany, “Superconformal fixed points with E_n global symmetry,” *Nucl. Phys.* **B489** (1997) 24–26 [hep-th/9610076](#).
- [34] D. Gaiotto, “ $N=2$ dualities,” <http://www.arXiv.org/abs/0904.27150904.2715>.
- [35] K. Hori and C. Vafa, “Mirror symmetry,” <http://www.arXiv.org/abs/hep-th/0002222>.
- [36] K. Hori, A. Iqbal, and C. Vafa, “D-branes and mirror symmetry,” <http://www.arXiv.org/abs/hep-th/0005247>.
- [37] S. Fomin and A. Zelevinsky, “Cluster algebras. I. Foundations,” *J. Amer. Math. Soc.* **15** (2002), no. 2, 497–529 (electronic).
- [38] S. Fomin and A. Zelevinsky, “Cluster algebras. II. Finite type classification,” *Invent. Math.* **154** (2003), no. 1, 63–121.
- [39] S. Fomin and N. Reading, “Root systems and generalized associahedra,” in *Geometric combinatorics*, vol. 13 of *IAS/Park City Math. Ser.*, pp. 63–131. Amer. Math. Soc., Providence, RI, 2007. <http://www.arXiv.org/abs/math/050551>.
- [40] S. Fomin and A. Zelevinsky, “Cluster algebras: notes for the CDM-03 conference,” in *Current developments in mathematics, 2003*, pp. 1–34. Int. Press, Somerville, MA, 2003.
- [41] B. Keller, “Cluster algebras, quiver representations and triangulated categories,” <http://www.arXiv.org/abs/0807.19600807.1960>.
- [42] B. Keller, “Quiver mutation in Java”, available from the author’s homepage, <http://www.institut.math.jussieu.fr/~keller/quivermutation>
- [43] K. Saito, “Extended affine root systems.I. Coxeter transformations,” *Pucl. Res. Inst. Math. Sci.* **21** (1985) 75–179.

- [44] A. Berenstein, S. Fomin and A. Zelevinsky, “Cluster algebras III: Upper bounds and double Bruhat cells,” Duke Math. J. **126** (2005) 1–52, [math.RT/0305434](https://arxiv.org/abs/math.RT/0305434).
- [45] H. Ooguri and C. Vafa, “Two-dimensional black hole and singularities of CY manifolds,” Nucl. Phys. **B463** (1996) 55–72, <http://www.arXiv.org/abs/hep-th/9511164>.
- [46] A. Kleemann, W. Lerche, P. Mayr, C. Vafa, and N. P. Warner, “Self-dual strings and $n=2$ supersymmetric field theory,” Nucl. Phys. **B477** (1996) 746–766, <http://www.arXiv.org/abs/hep-th/9604034>.
- [47] S. Cecotti and C. Vafa, “Ising model and $N=2$ supersymmetric theories,” Commun. Math. Phys. **157** (1993) 139–170 [hep-th/9209085](https://arxiv.org/abs/hep-th/9209085).
- [48] J. Milnor, “Singular points of complex hypersurfaces,” Annals of Math. Studies **61**, Princeton, 1968.
- [49] P. Fendley, S.D. Mathur, C. Vafa and N.P. Warner, Phys. Lett. **B243** (1990) 257.
- [50] D. Gaiotto, “Surface Operators in $N=2$ 4d Gauge Theories,” <http://www.arXiv.org/abs/0911.1316>.
- [51] S. Cecotti and C. Vafa, “Topological-antitopological fusion,” Nucl. Phys. **B367** (1991) 359–461.
- [52] S. Cecotti and C. Vafa, “Exact results for supersymmetric sigma models,” Phys. Rev. Lett. **68** (1992) 903–906 [hep-th/9111016](https://arxiv.org/abs/hep-th/9111016).
- [53] E. Martinec, Phys. Lett. **217B** (1989) 431.
- [54] C. Vafa and N.P. Warner, Phys. Lett. **43** (1989) 730.
- [55] B. Keller, “The periodicity conjecture for pairs of Dynkin diagrams,” <http://www.arXiv.org/abs/1001.1531>.
- [56] S. Cecotti, “Trieste lectures on wall-crossing invariants,” available from the author’s homepage, http://people.sissa.it/~cecotti/ictp_text.pdf.

SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI, VIA BONOMEA 265,
I-34100 TRIESTE, ITALY

E-mail address: `cecotti@sissa.it`

JEFFERSON PHYSICAL LABORATORY, HARVARD UNIVERSITY, CAMBRIDGE, MA
02138, USA

E-mail address: `vafa@physics.harvard.edu`

This page intentionally left blank

Existence, uniqueness and removable singularities for nonlinear partial differential equations in geometry

F. Reese Harvey and H. Blaine Lawson, Jr.

ABSTRACT. This paper surveys some recent results on existence, uniqueness and removable singularities for fully nonlinear differential equations on manifolds. The discussion also treats restriction theorems and the strong Bellman principle.

CONTENTS

1. Introduction	103
2. Subequations—A Geometric Approach	111
3. Jet Equivalence of Subequations	118
4. Monotonicity.	123
5. Comparison and Strict Approximation	128
6. Removable Singularities	131
7. Boundary Convexity	133
8. The Dirichlet Problem	136
9. Restriction Theorems	143
10. Convex Subequations and the Strong Bellman Principle	145
11. Applications to Almost Complex Manifolds	147
Appendix A. A Pocket Dictionary	148
Appendix B. Examples of Basic Monotonicity Cones	150
References	152

1. Introduction

Calibrated geometries are considered generalizations of Kähler geometry. They resemble Kähler geometry in having large families of distinguished

Partially supported by the N.S.F.

subvarieties determined by a fixed differential form. On the other hand, they seemed at first to be unlike Kähler geometry in having no suitable analogue of holomorphic functions. However, it was realized several years ago that the analogues of plurisubharmonic functions do exist (in abundance) on any calibrated manifold, and a potential theory was developed in this context [HL_{2,3}]. This led us naturally to the study of “maximal” or “extremal” functions, the analogues of solutions to the homogeneous complex Monge-Ampère equation, first considered by Bremermann [B] and Walsh [W] and later developed, in the inhomogeneous case, by Bedford-Taylor [BT_{*}] and others. The techniques and results developed in our study turned out to have substantial applications outside of calibrated geometry – in particular to many of the highly degenerate elliptic equations which appear naturally in geometry.

This paper is a survey of those techniques and results. We will address questions of existence and uniqueness for the Dirichlet Problem, the question of removable singularities for solutions and subsolutions, and the problem of restriction. The techniques apply broadly to fully nonlinear (second-order) equations in geometry, and in particular, to those which arise “universally” on riemannian, hermitian, or calibrated manifolds. A number of examples and applications will be discussed, including a proof of the Pali Conjecture on almost complex manifolds. Many more examples appear in the references.

It is conventional in discussing nonlinear differential equations to introduce the notions of a subsolution and supersolution, and define a solution to be a function which is both. In this paper we adopt an intrinsic approach by specifying a subset F of constraints on the value of a function and its derivatives. The classical subsolutions are defined to be the C^2 -functions u whose 2-jet (u, Du, D^2u) lies in F at each point. The set F will be called a **subequation**, and the functions u with $(u, Du, D^2u) \in F$ are called F -**subharmonic**.

The notion of supersolution is captured by the **dual** subequation

$$\tilde{F} \equiv -\{\sim \text{Int}F\} = \sim \{-\text{Int}F\},$$

and classical solutions u are just those where u is F -subharmonic and $-u$ is \tilde{F} -subharmonic. They have the property that $(u, Du, D^2u) \in \partial F$ at each point, since $\partial F = F \cap (\sim \tilde{F})$, and they will be called F -**harmonic** functions.

The simplest example is the Laplace equation, where $F = \{\text{tr}(D^2u) \geq 0\} = \tilde{F}$.

The most basic example is the Monge-Ampère subequation $\mathcal{P} = \{D^2u \geq 0\}$ with $\partial\mathcal{P} \subset \{\det D^2u = 0\}$. The dual $\tilde{\mathcal{P}}$ -subharmonics are the *subaffine functions* (see 2.1.8).

Adopting this point of view brings out an internal duality:

$$\tilde{\tilde{F}} = F,$$

and enables the roles of F and \tilde{F} to be interchanged in the analysis. This symmetry is often enlightening. It is particularly so when discussing the boundary geometry necessary for solving the Dirichlet problem.

A dictionary relating this approach to the more classical one is given in Appendix A.

The first step in our analysis is to extend the notion of F -subharmonicity to general upper semi-continuous $[-\infty, \infty)$ -valued functions. This is done in §2 where it is noted that these generalized F -subharmonic functions enjoy essentially all the useful properties of classical subharmonic functions. However, for this to be meaningful, F must satisfy a certain *positivity condition*, corresponding to weak ellipticity. We also require a *negativity condition*, corresponding to weak “properness”.

For the sake of clarity our exposition will often jump between the two extreme cases:

- (1) Constant coefficient (parallel) subequations in \mathbf{R}^n , and
- (2) General subequations on manifolds.

In fact, for many equations of interest in geometry and, in particular, those which are the principal focus of this survey, these two cases are directly related by the notion of **jet-equivalence**, introduced in §3. This basic concept plays a fundamental role in our work. Jet-equivalence is a certain transformation of all the variables. It can often be quite radical – turning mild equations into nasty ones, homogeneous equations into inhomogeneous ones, etc.

As stated, many important nonlinear equations on manifolds are locally jet-equivalent, in local coordinates, to constant coefficient equations. In this case the results of Slodkowski [S₁] and Jensen [J₁], and methods of viscosity theory [CIL], [C] can be applied to prove *local weak comparison*, and therefore *global weak comparison* — the first main step in the analysis of the Dirichlet Problem.

This leads to another concept of basic importance here: that of a **monotonicity cone**, introduced in §4. It gives the approximation tools needed to promote weak comparison to *full comparison* (see Definition 5.1) which, together with appropriate boundary geometry, yields both uniqueness and existence for the Dirichlet Problem. A subequation M is called a *monotonicity cone* for a subequation F if

$$F + M \subset F \tag{1.1.1}$$

and each fibre M_x , for $x \in X$, is a convex cone with vertex at the origin. One has that

$$F + M \subset F \iff \tilde{F} + M \subset \tilde{F},$$

so a monotonicity cone for F is also one for \tilde{F} .

Monotonicity cones play a role in the theory of removable singularities. For M as above, we define a closed subset $E \subset X$ to be *M -polar* if

$E = \{x : \psi(x) = -\infty\}$ for some M -subharmonic function which is smooth on $X - E$.

If M is a monotonicity cone for a subequation F , then M -polar sets are removable for F -subharmonic and F -harmonic functions on X .

(See Theorems 6.2.1 and 6.2.2.) This applies, for example, to all branches of the complex Monge-Ampère equation (see 2.1.10). Moreover, if a constant pure second-order subequation F in \mathbf{R}^n is M -monotone, where $M \equiv \mathcal{P}(p) \subset \text{Sym}^2(\mathbf{R}^n)$ is defined in terms of the ordered eigenvalues by $\lambda_1(A) + \dots + \lambda_{[p]}(A) + (p - [p])\lambda_{p+1}(A) \geq 0$, then

*any closed subset of locally finite Hausdorff $p - 2$ measure
is removable for F and \tilde{F} .*

This applies to the calibration case. It generalizes certain results in [CLN], [AGV] and [La*].

Monotonicity cones also play a key role in comparison. The monotonicity condition (1.1.1) is equivalent to

$$F + \tilde{F} \subset \widetilde{M}.$$

For many basic monotonicity cones, the \widetilde{M} -subharmonic functions satisfy the Zero Maximum Principle (see Appendix B). In such cases, comparison (see 5.1) comes down to an *addition theorem*: if u is F -subharmonic and v is \tilde{F} -subharmonic, then $u + v$ is \widetilde{M} subharmonic.

There is a last ingredient needed for the Dirichlet Problem – the necessary boundary geometry. Associated to each subequation F , there is a notion of strict F -convexity for oriented hypersurfaces. There are certain equations, like the k -Laplacian for $1 < k \leq \infty$ (see 7.4(a)), for which all hypersurfaces are strictly F -convex. This convexity is defined in terms of the asymptotic geometry of F at infinity (see §7). It is quite often easy to compute, and it can be expressed directly in terms of the second fundamental form.

This notion of boundary convexity implies existence, via the Perron process, once comparison has been established.

If comparison holds for a subequation F on a manifold X , then the Dirichlet Problem is uniquely solvable for F -harmonic functions on every domain $\Omega \subset X$ with smooth boundary which is strictly F and \tilde{F} convex.

Unique solvability for the Dirichlet Problem means that for every $\varphi \in C(\partial\Omega)$, there exists a unique $u \in C(\overline{\Omega})$ such that

$$u|_{\Omega} \in F(\Omega) \quad \text{and} \quad u|_{\partial\Omega} = \varphi$$

This theorem combines with results discussed above to prove the following general result.

THEOREM 8.1.2. *Let F be a subequation with monotonicity cone M . Suppose that:*

- (i) F is locally affinely jet-equivalent to a constant coefficient subequation, and
- (ii) X carries a smooth strictly M -subharmonic function.

Then existence and uniqueness hold for the Dirichlet problem for F -harmonic functions on any domain $\Omega \subset\subset X$ whose boundary is both strictly F - and \tilde{F} -convex.

The global condition (ii) is essential for a result of this generality. For example, suppose X is a riemannian manifold and $F \equiv \{\text{Hess } u \geq 0\}$, where $\text{Hess } u$ is the riemannian hessian. Given a domain $\Omega \subset\subset X$ with strictly convex boundary, one can completely change the geometry and topology in the interior of Ω without affecting the boundary. The subequation F continues to satisfy (i), but solutions to the Dirichlet Problem won't exist unless (ii) is satisfied. Another good example is the complex analogue $F = \mathcal{P}^C$ on an almost complex hermitian manifold (the homogeneous complex Monge-Ampère equation). Here condition (ii) amounts to the hypothesis that X carries at least one strictly plurisubharmonic function.

In homogeneous spaces one can apply a trick of Walsh [W] to establish existence without uniqueness.

THEOREM 8.1.3. *Let $X = G/H$ be a riemannian homogeneous space and suppose that $F \subset J^2(X)$ is a subequation which is invariant under the natural action of G on $J^2(X)$. Let $\Omega \subset\subset X$ be a connected domain whose boundary is both F and \tilde{F} strictly convex. Then existence holds for the Dirichlet problem for F -harmonic functions on Ω .*

These results apply to a wide spectrum of equations. Many examples have been discussed in [HL4,6,7] and are summarized in §2 below.

- (**Constant Coefficients**). Theorem 8.1.3 establishes existence for any constant coefficient subequation F in \mathbf{R}^n , and uniqueness also follows, by 8.1.2, whenever F has monotonicity cone M and there exists a strictly M -subharmonic function on $\bar{\Omega}$. If F is pure second-order, for example, the function $|x|^2$ works for any M , and so uniqueness always holds.

For invariant equations on a sphere, existence always holds by Theorem 8.1.3. However, for domains which do not lie in a hemisphere, where there exists a convex function, comparison and its consequences can fail, even for pure second-order equations (see Appendix D in [HL6]).

- (**Branches**). The homogeneous Monge-Ampère equations over \mathbf{R}, \mathbf{C} or \mathbf{H} each have branches defined by $\lambda_k(D^2u) = 0$ where $\lambda_1 \leq \dots \leq \lambda_n$ are the ordered eigenvalues. (See 2.1.3 and 2.1.10.) In fact the equation given by the ℓ^{th} elementary symmetric function $\sigma_\ell(D^2u) = 0$ also has ℓ distinct branches. This is a general phenomenon which applies to any homogeneous polynomial on $\text{Sym}^2(\mathbf{R}^n)$.

which is Gårding hyperbolic with respect to the identity. (See [HL_{7,8}] and 4.3.4 below.)

- (**The Special Lagrangian Potential Equation**). This equation $F(c)$, given in 2.2.1(d), can be treated for all values of c and has the nice feature that $\tilde{F}(c) = F(-c)$.
- (**Geometrically Determined Subequations – Calibrations**). These are subequations determined by a compact subset \mathbf{G} of the Grassmann bundle of tangent p -planes by requiring that $\text{tr}_W(\text{Hess } u) \geq 0$ for all $W \in \mathbf{G}$. These include many interesting examples, including the subequations in calibrated geometry discussed at the outset. It also includes a new polynomial differential equation in Lagrangian geometry (see 2.1.11(d)). Incidentally, this equation has branches whose study is a non-trivial application of the Gårding theory above.
- (**Equations Involving the Principal Curvatures of the Graph and the k -Laplacian**). For all such invariant equations on G/H , Theorem 8.1.3 gives existence (but not uniqueness). Strict boundary convexity is easily computable (see [HL₆, §17] for example). Existence holds on *all* domains for the k -Laplacian $|\nabla u|^2 \Delta u + (k-2)(\nabla u)^t (\text{Hess } u)(\nabla u) = 0$, when $1 < k \leq \infty$ and when $k = 1$ on mean-convex domains, where uniqueness fails catastrophically.

A fundamental point is that all such equations can be carried over to any riemannian manifold with an appropriate (not necessarily integrable!) reduction of structure group. This is done by using the **riemannian hessian** given in §8.2. Theorem 8.1.2 can then be applied, and we obtain the following corollary. Let \mathbf{F} and \mathbf{M} be constant coefficient subequations in \mathbf{R}^n with invariance group G .

THEOREM 8.2.2. *Let F be a subequation with monotonicity cone M canonically determined by \mathbf{F} and \mathbf{M} on a riemannian manifold X with a topological G -structure. Let $\Omega \subset\subset X$ be a domain with smooth boundary which is both F and \tilde{F} strictly convex. Assume there exists a strictly M -subharmonic function on $\bar{\Omega}$. Then the Dirichlet Problem for F -harmonic functions is uniquely solvable for all $\varphi \in C(\partial\Omega)$.*

- (**Universal Riemannian Subequations**). Any constant coefficient subequation \mathbf{F} which is invariant under the natural action of $O(n)$ carries over directly to any riemannian manifold, and Theorem 8.2.2 applies. This includes most of the examples above.
- (**Universal Hermitian Subequations**). A constant coefficient subequation \mathbf{F} invariant under $U(n)$ carries over to any almost complex hermitian manifold. There is a quaternionic analogue. More generally, we have:
- (**Equations on Manifolds with G -Structure**). A constant coefficient subequation \mathbf{F} invariant under a subgroup $G \subset O(n)$ carries

over to any manifold equipped with a topological G -structure (see 8.2.1). This includes manifolds with topological (or quasi) calibrations based on any fixed form in $\Lambda^p \mathbf{R}^n$. Even the extreme case $G = \{e\}$ is interesting here. An $\{e\}$ -structure is a topological trivialization of TX . It transplants every constant coefficient equation to X , and Theorem 8.2.2 applies. This holds, for example, for every orientable 3-manifold and every Lie group.

Theorem 8.1.2 actually treats much more general equations on manifolds. Affine jet-equivalence gives great flexibility to the result.

Many variable-coefficient, inhomogeneous subequations on manifolds can be transformed by local affine jet-equivalence to universally defined subequations, such as those in Theorem 8.2.2, while preserving the domains of strict boundary convexity.

- (**Calabi-Yau-Type Equations**). This is a good example of the power of affine jet equivalence. It applies to treat equations of type $(i\partial\bar{\partial}u + \omega)^n = F(x, u)\omega^n$ on almost complex hermitian manifolds, where $F > 0$ is non-decreasing in u . See 3.2.8.
- (**Inhomogeneous Equations**). Many homogeneous equations can be transformed into inhomogeneous equations by affine jet equivalence. For example, from the k^{th} branch of the Monge-Ampère equation one can obtain: $\lambda_k(\text{Hess } u) = f(x)$ for any continuous function f . See 3.2.7.
- (**Obstacle Problems**). The methods here apply also to the Dirichlet Problem with an Obstacle. In this case not all boundary data are allowed. They are constrained by the obstacle function. This is another example of an inhomogeneous equation. See §8.6.
- (**Parabolic Equations**). Each of these subequations has a parabolic cousin, where existence and uniqueness results are generally stronger. See 8.5.

For any subequation F on a manifold X , one has the very natural

Restriction Question: When is the restriction of an F -subharmonic function on X to a submanifold $j : Y \subset X$, a $j^*(F)$ -subharmonic function on Y ?

For C^2 -functions, this always holds, and if fact defines the induced subequation j^*F . However, it is important and non-trivial for general upper semi-continuous subharmonics. There are several restriction results established in [HL9]. They are relevant to calibrated and riemannian geometry. Sometimes they lead to characterizing F -subharmonics in terms of their restrictions to special submanifolds.

An important case of this latter phenomenon occurs in almost complex manifolds. The “standard” way of defining plurisubharmonic functions is to require that the restrictions to (pseudo) holomorphic curves are subharmonic. There also exists an intrinsic subequation, whose subharmonics agree with the standard plurisubharmonic functions in the integrable case. Via the

restriction theorem, these two definitions have been shown to agree on any almost complex manifold [HL₁₀].

There is also the notion of a plurisubharmonic distribution on a general almost complex manifold. Nefton Pali [P] has shown that those which are representable by continuous $[-\infty, \infty)$ -valued functions are of the type above, and he conjectured that this should be true generally. This leads to another topic.

For convex subequations which are “second-order complete”, a Strong Bellman Principle can be applied. It enables one to prove that distributionally F -subharmonic functions correspond in a very precise sense to the upper semi-continuous F -subharmonic functions considered here. This is done in [HL₁₃]. Such arguments apply to prove the Pali Conjecture [HL₁₀].

Some Historical Notes. There is of course a vast literature on the principal branches of \mathcal{P} and \mathcal{P}^C of the real and complex Monge-Ampère equations. Just to mention a few of the historically significant contributions beginning with Alexandrov: [Al], [Po_{*}], [RT], [B], [W], [TU], [CNS_{*}], [CKNS], [BT_{*}], [HM], [S₁], [CY_{*}], and [Yau]. Quaternionic subharmonicity and the principal branch \mathcal{P}^H of the quaternionic Monge-Ampère equation have been studied in [A_{*}] and [AV]. On compact complex manifolds without boundary, viscosity solutions to equations of the form $(i\partial\bar{\partial}u + \omega)^n = e^\varphi v$, where $v > 0$ is a given smooth volume form, were studied in [EGZ]. By establishing a comparison principle they obtain existence and uniqueness of solutions in important borderline cases ($\omega \geq 0$, $v \geq 0$ with $\int v > 0$), and also show that these are the unique solutions in the pluripotential sense.

The parabolic form of the 1-Laplacian gives rise to mean curvature flow by the level set method. Some of the interesting results on this topic (see [ES_{*}], [CGG_{*}], [E], [Gi]) can be carried over from euclidean space to the riemannian setting by the methods of [HL₆].

The first basic work on the Dirichlet Problem for the convex branches of the Special Lagrangian potential equation appeared in [CNS₂], and there are further results by Yuan [Y], [WY].

In [AFS] and [PZ] standard viscosity theory has been adapted to riemannian manifolds by using the distance function, parallel translation, Jacobi fields, etc. For the problems considered here this machinery is not necessary.

In [S_{2,3,4}], Z. Slodkowski developed an axiomatic perspective on generalized subharmonic functions, and addressed the Dirichlet Problem in this context. He studied certain invariant “pseudoconvex classes” of functions on euclidean space and complex homogeneous spaces. There is a version of duality which plays an important role in his theory. It is formulated differently from the one here. However, in the cases of overlap the two notions of duality are equivalent. Interestingly, his results are used to prove a duality theorem for complex interpolation of normed spaces [S₅]

Concerning Regularity. In this paper there is no serious discussion of regularity for solutions of the Dirichlet Problem. Indeed, with the level of degeneracy allowed here, no regularity above continuity can be claimed generally. Consider $u_{xx} = 0$ in \mathbf{R}^2 for example. (See also [Po₁] and [NTV] and references therein.) A good account of regularity results can be found in [E]. A general exposition of viscosity methods and results appears in [CIL] and [C].

Concerning $-\infty$. Our approach here is to steadfastly treat subsolutions from the point of view of classical potential theory. We allow subsolutions (F -subharmonic functions) to assume the value $-\infty$, in contrast to standard viscosity theory where subsolutions are finite-valued. This has the advantage of including basic functions, like the fundamental solution of the Laplacian, Riesz potentials, and $\log|f|$ with f holomorphic, into the class of subsolutions. It also allows the constant function $u \equiv -\infty$, which is crucial for the restriction theorems discussed in Chapter 9. This issue is not important for the Dirichlet Problem.

2. Subequations—A Geometric Approach

The aim of this chapter is to present a geometric approach to subequations, pioneered by Krylov [K]. This point of view clarifies and conceptually simplifies many aspects of the theory. For transparency we begin with the basic case.

2.1. Constant Coefficient Subequations in \mathbf{R}^n . The 2-jets of functions on \mathbf{R}^n (i.e., Taylor polynomials of degree two) take values in the vector space

$$\mathbf{J}^2 \equiv \mathbf{R} \times \mathbf{R}^n \times \text{Sym}^2(\mathbf{R}^n) \quad \text{with traditional coordinates } (r, p, A). \quad (2.1.1)$$

DEFINITION 2.1.1. A *second-order constant coefficient subequation* on \mathbf{R}^n is a proper closed subset $\mathbf{F} \subset \mathbf{J}^2$ satisfying the **Positivity Condition**

$$\mathbf{F} + \mathcal{P} \subset \mathbf{F} \quad (P)$$

and the **Negativity Condition**

$$\mathbf{F} + \mathcal{N} \subset \mathbf{F} \quad (N)$$

where

$$\mathcal{P} \equiv \{(0, 0, A) \in \mathbf{J}^2 : A \geq 0\} \quad \text{and} \quad \mathcal{N} \equiv \{(r, 0, 0) \in \mathbf{J}^2 : r \leq 0\},$$

and the **Topological Condition**

$$\mathbf{F} = \overline{\text{Int}\mathbf{F}}. \quad (T)$$

We say \mathbf{F} is *pure second-order* if $\mathbf{F} = \mathbf{R} \times \mathbf{R}^n \times \mathbf{F}_0$ for a closed subset $\mathbf{F}_0 \subset \text{Sym}^2(\mathbf{R}^n)$. In this case only (P) is required, since (N) is automatic and one can show that (P) \Rightarrow (T). Such subequations are often simply denoted by the subset \mathbf{F}_0 of $\text{Sym}^2(\mathbf{R}^n)$.

EXAMPLE 2.1.2. Some basic pure second-order examples are:

(a) **The Laplace Subequation:**

$$\mathbf{F}_0 = \{A \in \text{Sym}^2(\mathbf{R}^n) : \text{tr}A \geq 0\}.$$

(b) **The Homogeneous Monge-Ampère Subequation:**

$$\mathbf{F}_0 = \{A \in \text{Sym}^2(\mathbf{R}^n) : A \geq 0\} \cong \mathcal{P}.$$

(c) **The k^{th} Elementary Symmetric Function Subequation:**

$$\mathbf{F}_0 = \{A \in \text{Sym}^2(\mathbf{R}^n) : \sigma_\ell(A) \geq 0, 1 \leq \ell \leq k\}.$$

(d) **The Special Lagrangian Potential Subequation:**

$$\mathbf{F}_0 = \{A \in \text{Sym}^2(\mathbf{R}^n) : \text{tr}(\arctan A) \geq c\}.$$

(e) **The Calabi-Yau Subequation:** (This is not pure second-order, but it is gradient-independent.)

$$\mathbf{F} = \{(r, p, A) \in \text{Sym}^2(\mathbf{R}^n) : \text{tr}(A + I) \geq e^r \text{ and } A + I \geq 0\}.$$

REMARK 2.1.3. In $\mathbf{C}^n = (\mathbf{R}^{2n}, J)$ each of the examples above has a complex analogue given by replacing A with its hermitian symmetric part $A_{\mathbf{C}} \equiv \frac{1}{2}(A - JAJ)$. The same applies in quaternionic n -space $\mathbf{H}^n = (\mathbf{R}^{4n}, I, J, K)$ with A replaced by $A_{\mathbf{H}} \equiv \frac{1}{4}(A - IAI - JAJ - KAK)$.

DEFINITION 2.1.4. Given a constant coefficient subequation \mathbf{F} on \mathbf{R}^n , the **dual** subequation $\tilde{\mathbf{F}}$ is defined by

$$\tilde{\mathbf{F}} \equiv \sim(-\text{Int}\mathbf{F}) = -(\sim \text{Int}\mathbf{F}).$$

LEMMA 2.1.5. \mathbf{F} is a subequation $\iff \tilde{\mathbf{F}}$ is a subequation, and in this case

$$\tilde{\tilde{\mathbf{F}}} = \mathbf{F} \quad \text{and} \quad \widetilde{\mathbf{F} + J} = \tilde{\mathbf{F}} - J$$

for all $J \in \mathbf{J}^2$.

The proof can be found in [HL₄, §4]. In the examples above the dual subequations are easily computed in terms of the eigenvalues of A (or $A_{\mathbf{C}}$, etc.). One finds that the Laplace subequation is self-dual ($\tilde{\mathbf{F}} = \mathbf{F}$) but the others are generally not. Of particular interest is example (b) where the dual of $\mathcal{P} \equiv \{A \geq 0\}$ is

$$\tilde{\mathcal{P}} \cong \{A \in \text{Sym}^2(\mathbf{R}^n) : \text{at least one eigenvalue of } A \text{ is } \geq 0\} \quad (2.1.2)$$

We now present a concept of central importance which comes from viscosity theory [CIL]. For any manifold X , let $\text{USC}(X)$ denote the set of upper semi-continuous functions $u : X \rightarrow [-\infty, \infty)$. Given $u \in \text{USC}(X)$ and a point $x \in X$, a **test function for u at x** is a C^2 -function φ defined near x so that

$$u \leq \varphi \quad \text{and} \quad u(x) = \varphi(x).$$

DEFINITION 2.1.6. Let \mathbf{F} be a constant coefficient subequation on \mathbf{R}^n and fix an open set $X \subset \mathbf{R}^n$. A function $u \in \text{USC}(X)$ is said to be **\mathbf{F} -subharmonic** on X if for each $x \in X$ and each test function φ for u at x , the 2-jet (or total second derivative) of φ satisfies

$$J_x^2\varphi \equiv (\varphi(x), (D\varphi)_x, (D^2\varphi)_x) \in \mathbf{F}. \quad (2.1.3)$$

It is important that this condition (2.1.3) is only required at points where test functions actually exist. The set of such functions is denoted by $F(X)$.

It is striking that the space $F(X)$ of F -subharmonics shares many of the important properties enjoyed by classical subharmonic functions (see 2.3.1 below). The C^2 -functions $u \in F(X)$ are exactly those with $J_x^2u \in \mathbf{F}$ for all $x \in X$. This basic fact requires the Positivity Condition (P) on \mathbf{F} . Interestingly, the other properties in 2.3.1 do not require (P).

For the subequation \mathcal{P} in example (b) we have the following.

PROPOSITION 2.1.7. (see [HL4, Rmk. 4.9] and [HL9, Prop. 2.7])

- (i) $\mathcal{P}(X)$ is the set of convex functions on X .
- (ii) $\tilde{\mathcal{P}}(X)$ is the set of subaffine functions on X .

DEFINITION 2.1.8. A function $u \in \text{USC}(X)$ is called **subaffine** if for each compact subset $K \subset X$ and each affine function a ,

$$u \leq a \text{ on } \partial K \quad \Rightarrow \quad u \leq a \text{ on } K.$$

Note that subaffine functions satisfy the maximum principle. In fact, for a pure second-order subequations, the subequation $\tilde{\mathcal{P}}$ is universal for this property. That is, if the functions in $\mathbf{F}(X)$ satisfy the maximum principle, then $\mathbf{F} \subset \tilde{\mathcal{P}}$. We note also that functions which are locally subaffine are globally subaffine, while the corresponding statement for functions satisfying the maximum principle is false.

DEFINITION 2.1.9. Let \mathbf{F} and X be as in Definition 2.1.6. A function $u \in \text{USC}(X)$ is said to be **\mathbf{F} -harmonic** on X if

$$u \in F(X) \quad \text{and} \quad -u \in \tilde{F}(X) \quad (2.1.4)$$

Condition (2.1.4) implies that u is continuous. If u is twice differentiable at a point x , then (2.1.4) implies that

$$J_x^2u \in \mathbf{F} \cap (-\tilde{\mathbf{F}}) = \mathbf{F} \cap (\sim \text{Int}\mathbf{F}) = \partial\mathbf{F}.$$

Thus if \mathbf{F} is defined classically as the closure of a set $\{f(r, p, A) > 0\}$ for a continuous function $f : \mathbf{J}^2 \rightarrow \mathbf{R}$, then any $u \in C^2(X)$ which is \mathbf{F} -harmonic satisfies the differential equation

$$f(u, Du, D^2u) = 0 \quad \text{on } X,$$

however, the converse is not always true.

NOTE 2.1.10. (Branches) It is instructive to consider the most basic of subequations, \mathcal{P} . A C^2 -function u which is \mathcal{P} -harmonic satisfies the homogeneous Monge-Ampère equation

$$\det(D^2u) = 0. \quad (2.1.5)$$

However, u is required to have the additional property of being convex (cf. Alexandroff [Al]). (In the complex analogue u is plurisubharmonic.)

The equation (2.1.5) has other solutions corresponding to other “branches” of the locus $\{\det A = 0\}$, which can also be handled by this theory. Given a symmetric matrix A , let $\lambda_1(A) \leq \lambda_2(A) \leq \dots \leq \lambda_n(A)$ be the ordered eigenvalues of A . Since $\det A = \lambda_1(A) \cdots \lambda_n(A)$, equation (2.1.5) can be split into branches

$$\lambda_k(D^2u) = 0. \quad (2.1.5)_k$$

for $k = 1, \dots, n$. By monotonicity of eigenvalues, each $\Lambda_k \equiv \{\lambda_k \geq 0\}$ is a subequation. Interestingly, the dual of a branch is another branch:

$$\widetilde{\Lambda}_k = \Lambda_{n-k+1}$$

This phenomenon of branches occurs in many equations of geometric significance.

EXAMPLE 2.1.11. (Geometrically Defined Subequations) There is a large class of subequations which arise naturally in our set-theoretic setting. Let $G(p, \mathbf{R}^n)$ denote the Grassmannian of p -planes in \mathbf{R}^n . For each compact subset $\mathbf{G} \subset G(p, \mathbf{R}^n)$ we define the pure second-order subequation

$$\mathbf{F}(\mathbf{G}) \equiv \{A \in \text{Sym}^2(\mathbf{R}^n) : \text{tr}_W A \geq 0 \text{ for all } W \in \mathbf{G}\} \quad (2.1.6)$$

with dual

$$\widetilde{\mathbf{F}(\mathbf{G})} = \{A \in \text{Sym}^2(\mathbf{R}^n) : \text{tr}_W A \geq 0 \text{ for some } W \in \mathbf{G}\}$$

The $\mathbf{F}(\mathbf{G})$ -subharmonic functions are called \mathbf{G} -plurisubharmonic. This terminology is justified by the following. Let $X \subset \mathbf{R}^n$ be an open set.

THEOREM 2.1.12. *A function $u \in \text{USC}(X)$ is \mathbf{G} -plurisubharmonic if and only if for every affine \mathbf{G} -plane L the restriction $u|_{X \cap L}$ is subharmonic for the standard Laplacian on L . The same statement holds with the affine \mathbf{G} -planes expanded to include all minimal \mathbf{G} -submanifolds of X . (A \mathbf{G} -submanifold is one whose tangent planes are elements of \mathbf{G}).*

This follows from a Restriction Theorem in [HL9], which is discussed in Chapter 9.

- (a) $\mathbf{G} = G(1, \mathbf{R}^n)$: In this case $\mathbf{F}(\mathbf{G}) = \mathcal{P}$ and the \mathbf{G} -plurisubharmonic functions are the classical convex functions, i.e., those which are convex on affine lines.
- (b) $\mathbf{G} = G_C(1, \mathbf{C}^n) \subset G(2, \mathbf{R}^{2n})$ the set of complex lines in \mathbf{C}^n : In this case $\mathbf{F}(\mathbf{G}) = \mathcal{P}^C$ (see 4.3.1), and the \mathbf{G} -plurisubharmonic functions are the standard plurisubharmonic functions, i.e., those which are subharmonic on complex lines.

- (c) $\mathbf{G} = G(p, \mathbf{R}^n)$: Here the \mathbf{G} -plurisubharmonic functions are the standard p -plurisubharmonic functions, i.e., those which are subharmonic on affine p -planes. This subequation has the feature that each p -plurisubharmonic function is also \mathbf{G} -plurisubharmonic for every closed $\mathbf{G} \subset G(p, \mathbf{R}^n)$. The analogue $\mathbf{G} = G(p, \mathbf{C}^n)$ in the complex case plays a role in analysis in several complex variables.

The \mathbf{G} -harmonic functions in these cases are viscosity solutions to differential equations which are $O(n)$ (or $U(n)$) invariant polynomials in the variables D^2u . Each of these equations has branches which will be discussed further in 4.3.1 and 4.3.2 below.

- (d) $\mathbf{G} = \text{LAG} \subset G(n, \mathbf{R}^{2n})$ the set of Lagrangian planes in $\mathbf{C}^n = \mathbf{R}^{2n}$: In this case the LAG-plurisubharmonic functions are relatively new and interesting. The corresponding harmonics are viscosity solutions to a differential equation which is a $U(n)$ -invariant polynomial in the variables D^2u (see [HL14]). This equation also has branches.

Many important examples come directly from the theory of calibrations. A *parallel calibration* in \mathbf{R}^n is a constant coefficient p -form whose restriction satisfies $\pm\varphi|_W \leq \text{vol}_W$ for all oriented p -planes W . For such a φ , we define $\mathbf{G} \equiv G(\varphi)$ to be the set of $W \in G(p, \mathbf{R}^n)$ such that $|\varphi|_W = \text{vol}_W$. In this case $G(\varphi)$ -submanifolds (or simply φ -submanifolds) are automatically minimal. When $\varphi = \omega$ is the Kähler form in \mathbf{C}^n , we recover case (b) above, where the ω -submanifolds are the holomorphic curves. (This carries over to any symplectic manifold (X, ω) with a compatible almost complex structure in the sense of Gromov [Gr].) The $G(\varphi)$ -plurisubharmonic (or simply φ -plurisubharmonic) functions are essentially **dual** to the φ -submanifolds (see [HL_{2,3}]), and they provide calibrated geometry with new tools from conventional analysis.

- (e) $\mathbf{G} = G(\varphi) = \text{SLAG} \subset G(n, \mathbf{R}^{2n})$ where $\varphi = \text{Re}(dz_1 \wedge \cdots \wedge dz_n)$ is the Special Lagrangian Calibration (cf. [HL₁]). The notions of Special Lagrangian submanifolds and of SLAG-plurisubharmonic and SLAG-harmonic functions carry over to any Ricci-flat Kähler manifold (cf. [HL₁]). The SLAG-subvarieties play a central role in the conjectured differential-geometric interpretation of mirror symmetry presented in [SYZ_{1,2}].

- (f) $\mathbf{G} = G(\varphi) \subset G(3, \mathbf{R}^7)$ where $\mathbf{R}^7 = \text{Im}\mathbf{O}$ is the imaginary octonions and $\varphi(x, y, z) \equiv \langle x \cdot y, z \rangle$ is the **associative** calibration. There is a rich geometry of associative submanifolds, and an abundance of φ -plurisubharmonic and φ -harmonic functions. The same applies to the **coassociative** calibration $\psi = * \varphi$. Both calibrations make sense on any 7-manifold with G_2 -holonomy.

- (g) $\mathbf{G} = G(\Phi) \subset G(4, \mathbf{R}^8)$ where $\mathbf{R}^8 = \mathbf{O}$, the octonions, and $\Phi(x, y, z, w) \equiv \langle x \times y \times z, w \rangle$ is the **Cayley** calibration. There is a rich geometry of Cayley submanifolds, and an abundance of Φ -plurisubharmonic and Φ -harmonic functions. All this carries over to any 8-manifold with Spin_7 -holonomy.

Note. While the φ -harmonic functions in examples (e), (f) and (g) are of basic interest in calibrated geometry, they appear **not** to satisfy any polynomial equation in u, Du and D^2u . This is one justification for the approach to subequations adopted here.

2.2. Subequations on General Manifolds. Suppose now that X is a smooth manifold of dimension n . The natural setting for second-order differential equations on X is the bundle of **2-jets** of functions on X . This is the bundle $J^2(X) \rightarrow X$ whose fibre at $x \in X$ is the quotient $J_x^2(X) = C_x^\infty / C_{x,3}^\infty$ of germs of smooth functions at x modulo those which vanish to order 3 at x .

Restriction from 2-jets to 1-jets gives a basic short exact sequence

$$0 \longrightarrow \text{Sym}^2(T^*X) \longrightarrow J^2(X) \longrightarrow J^1(X) \longrightarrow 0 \quad (2.2.1)$$

where $\text{Sym}^2(T_x^*X)$ embeds into $J_x^2(X)$ as the 2-jets of functions having a critical value zero at x . The dual exact sequence is

$$0 \longrightarrow J_1(X) \longrightarrow J_2(X) \xrightarrow{\sigma} \text{Sym}^2(TX) \longrightarrow 0. \quad (2.2.2)$$

Sections of $J_k(X)$ are linear differential operators of degree $\leq k$ on X , and σ is the *principal symbol map* on operators of degree 2.

There are two important, intrinsically defined subbundles of $J^2(X)$ which correspond to the subspaces \mathcal{P} and \mathcal{N} in Definition 2.1.1, namely:

$$\begin{aligned} \mathcal{P} &\equiv \{A \in \text{Sym}^2(T^*X) : A \geq 0\} & \text{and} \\ \mathcal{N} &\equiv \{\text{2-jets of constant functions} \leq 0\}. \end{aligned}$$

DEFINITION 2.2.1. A *subequation* of order ≤ 2 on X is a closed subset $F \subset J^2(X)$ satisfying (under fibre-wise sum) the *Positivity Condition*:

$$F + \mathcal{P} \subset F, \quad (P)$$

the *Negativity Condition*:

$$F + \mathcal{N} \subset F, \quad (N)$$

and the *Topological Condition*:

$$(i) \ F = \overline{\text{Int}F}, \quad (ii) \ F_x = \overline{\text{Int}F_x}, \quad (iii) \ \text{Int}F_x = (\text{Int}F)_x \quad (T)$$

where $\text{Int}F_x$ denotes interior with respect to the fibre.

Note that \mathcal{P} is *not* a subequation. However, when discussing pure second-order subequations, it is sometimes used as an abbreviation for $\mathbf{R} \times \mathbf{R}^n \times \mathcal{P}$, which is a subequation. (see 2.1.1 and 2.1.2).

REMARK 2.2.2. (Splitting the 2-Jet Bundle) Let ∇ be a torsion-free connection on X . Then each $u \in C^2(X)$ has an associated hessian $\text{Hess } u \in \Gamma(\text{Sym}^2(T^*X))$ defined on vector fields V, W by

$$(\text{Hess } u)(V, W) = VWu - WVu - (\nabla_V W)u. \quad (2.2.3)$$

Since $\nabla_V W - \nabla_W V = [V, W]$, one easily sees that $\text{Hess } u$ is a symmetric tensor. If X is riemannian and ∇ is the Levi-Civita connection, then $\text{Hess } u$ is called the *riemannian hessian* of u .

The hessian in (2.2.3) depends only on the 2-jet of u at each point, and so it gives a splitting of the short exact sequence (2.2.1). That is, we can write

$$J^2(X) = \mathbf{R} \oplus T^*X \oplus \text{Sym}^2(T^*X) \quad (2.2.4)$$

by the association

$$J_x^2 u = (u(x), (du)_x, \text{Hess}_x u).$$

REMARK 2.2.3. (Universal Subequations) Each of the subequations given in Example 2.1.2 carries over to any riemannian manifold X by using the splitting (2.2.4) (determined by the riemannian hessian). For instance, Example 2.1.2(a) gives the Laplace-Beltrami operator. More generally, any constant coefficient subequation $\mathbf{F} \subset \mathbf{J}^2$ which is invariant under the action of the group $O(n)$, transplants to every riemannian manifold. In the case of $\mathbf{C}^n = (\mathbf{R}^{2n}, J)$, each $U(n)$ -invariant subequation transplants to every hermitian almost complex manifold.

There is, in fact, a very general principle:

Let $\mathbf{F} \subset \mathbf{J}^2$ be a constant coefficient subequation which is invariant under a subgroup $G \subset O(n)$ acting naturally on \mathbf{J}^2 . Then \mathbf{F} carries over to a subequation F on every manifold X with a topological G -structure.

See [HL6] and §8.2 below for definitions and many examples.

The concepts of the previous section carry over to this general setting.

DEFINITION 2.2.4. Given a subequation $F \subset J^2(X)$, the **dual** subequation \tilde{F} is defined by

$$\tilde{F} \equiv \sim(-\text{Int}F) = -(\sim \text{Int}F).$$

LEMMA 2.2.5.

$$F \text{ is a subequation} \iff \tilde{F} \text{ is a subequation},$$

and in this case

$$\tilde{\tilde{F}} = F \quad \text{and} \quad \widetilde{F + S} = \tilde{F} - S$$

for any section S of $J^2(X)$.

The proof can be found in [HL6 §3]. The dual of a universal subequation associated to $\mathbf{F} \subset \mathbf{J}^2$ is the universal subequation associated to $\tilde{\mathbf{F}}$. As before we have the following.

DEFINITION 2.2.6. Let F be a subequation on a manifold X . A function $u \in \text{USC}(X)$ is said to be **F -subharmonic** on X if for each $x \in X$ and each test function φ for u at x ,

$$J_x^2 \varphi \equiv (\varphi(x), (D\varphi)_x, (D^2\varphi)_x) \in F. \quad (2.2.5)$$

The set of such functions is denoted by $F(X)$.

DEFINITION 2.2.7. Let F be a subequation on a manifold X . A function $u \in \text{USC}(X)$ is said to be **F -harmonic** on X if

$$u \in F(X) \quad \text{and} \quad -u \in \tilde{F}(X) \quad (2.2.6)$$

As before, positivity ensures that a function $u \in C^2(X)$ is F -subharmonic on X iff $J_x^2 u \in F$ for all x , and it is F -harmonic iff

$$J_x^2 u \in \partial F \quad \text{for all } x.$$

2.3. Properties of F -Subharmonic Functions. The F -subharmonic functions share many of the important properties of classical subharmonic functions.

THEOREM 2.3.1. (Elementary Properties of F -Subharmonic Functions) Let F be an arbitrary closed subset of $J^2(X)$.

- (i) (*Maximum Property*) If $u, v \in F(X)$, then $w = \max\{u, v\} \in F(X)$.
- (ii) (*Coherence Property*) If $u \in F(X)$ is twice differentiable at $x \in X$, then $J_x^2 u \in F_x$.
- (iii) (*Decreasing Sequence Property*) If $\{u_j\}$ is a decreasing ($u_j \geq u_{j+1}$) sequence of functions with all $u_j \in F(X)$, then the limit $u = \lim_{j \rightarrow \infty} u_j \in F(X)$.
- (iv) (*Uniform Limit Property*) Suppose $\{u_j\} \subset F(X)$ is a sequence which converges to u uniformly on compact subsets to X , then $u \in F(X)$.
- (v) (*Families Locally Bounded Above*) Suppose $\mathcal{F} \subset F(X)$ is a family of functions which are locally uniformly bounded above. Then the upper semicontinuous regularization v^* of the upper envelope

$$v(x) = \sup_{f \in \mathcal{F}} f(x)$$

belongs to $F(X)$.

A proof can be found, for example, in Appendix B in [HL6]. For parts (i) and (ii), even the closure hypothesis on F can be weakened (op. cit.).

3. Jet Equivalence of Subequations

Many important nonlinear equations that occur in geometry can be transformed locally to constant coefficient equations. This technique allows one to apply standard arguments from viscosity theory to prove local comparison results.

3.1. Affine Automorphisms of the Jet Bundle $J^2(X)$. The transformations we shall use are the affine automorphisms of $J^2(X)$ which we now introduce. To begin, note that there is a canonical direct sum decomposition

$$J^2(X) = \mathbf{R} \oplus J_{\text{red}}^2(X) \quad (3.1.1)$$

where the trivial \mathbf{R} -factor corresponds to the value of the function. For the reduced 2-jet bundle there is a short exact sequence

$$0 \longrightarrow \text{Sym}^2(T^*X) \longrightarrow J_{\text{red}}^2(X) \longrightarrow T^*X \longrightarrow 0 \quad (3.1.2)$$

coming from (2.2.1) above.

DEFINITION 3.1.1. A linear isomorphism of $J^2(X)$ is an **automorphism** if, with respect to the splitting (3.1.1) it has the form $\text{Id} \oplus \Phi$ where $\Phi : J_{\text{red}}^2(X) \rightarrow J_{\text{red}}^2(X)$ has the following properties. We first require that

$$\Phi(\text{Sym}^2(T^*X)) = \text{Sym}^2(T^*X), \quad (3.1.3)$$

so by (3.1.2) there is an induced bundle automorphism

$$g = g_\Phi : T^*X \longrightarrow T^*X. \quad (3.1.4)$$

We further require that there exist a second bundle automorphism

$$h = h_\Phi : T^*X \longrightarrow T^*X \quad (3.1.5)$$

such that on $\text{Sym}^2(T^*X)$, Φ has the form $\Phi(A) = hAh^t$, i.e.,

$$\Phi(A)(v, w) = A(h^tv, h^tw) \quad \text{for } v, w \in TX. \quad (3.1.6)$$

The automorphisms of $J^2(X)$ form a group. They are the sections of the bundle of groups $\text{Aut}(J^2(X))$ whose fibre at $x \in X$ is the group of automorphisms of $J_x^2(X)$ defined by (3.1.3) - (3.1.6) above. See [HL₆, §6.2] for this and the following.

PROPOSITION 3.1.2. *With respect to any splitting*

$$J^2(X) = \mathbf{R} \oplus T^*X \oplus \text{Sym}^2(T^*X)$$

of the short exact sequence (2.2.1), a bundle automorphism has the form

$$\Phi(r, p, A) = (r, gp, hAh^t + L(p)) \quad (3.1.7)$$

*where $g, h : T^*X \rightarrow T^*X$ are bundle isomorphisms and L is a smooth section of the bundle $\text{Hom}(T^*X, \text{Sym}^2(T^*X))$.*

EXAMPLE 3.1.3. Given a local coordinate system (ξ_1, \dots, ξ_n) on an open set $U \subset X$, the *canonical trivialization*

$$J^2(U) = U \times \mathbf{R} \times \mathbf{R}^n \times \text{Sym}^2(\mathbf{R}^n) \quad (3.1.8)$$

is determined by $J_x^2 u = (u, Du, D^2 u)$ where $Du = (u_{\xi_1}, \dots, u_{\xi_n})$ and $D^2 u = ((u_{\xi_i \xi_j}))$ evaluated at the point $\xi(x) \in \mathbf{R}^n$. With respect to this splitting, every automorphism is of the form

$$\Phi(u, Du, D^2 u) = (u, gDu, h \cdot D^2 u \cdot h^t + L(Du)) \quad (3.1.9)$$

where $g_x, h_x \in \text{GL}_n$ and $L_x : \mathbf{R}^n \rightarrow \text{Sym}^2(\mathbf{R}^n)$ is linear for each point $x \in U$.

EXAMPLE 3.1.4. The trivial 2-jet bundle on \mathbf{R}^n has fibre

$$\mathbf{J}^2 = \mathbf{R} \times \mathbf{R}^n \times \text{Sym}^2(\mathbf{R}^n).$$

with automorphism group

$$\text{Aut}(\mathbf{J}^2) \equiv \text{GL}_n \times \text{GL}_n \times \text{Hom}(\mathbf{R}^n, \text{Sym}^2(\mathbf{R}^n))$$

where the action is given by

$$\Phi_{(g,h,L)}(r,p,A) = (r, gp, hAh^t + L(p)).$$

Note that the group law is

$$(\bar{g}, \bar{h}, \bar{L}) \cdot (g, h, L) = (\bar{g}g, \bar{h}h, \bar{h}L\bar{h}^t + \bar{L} \circ g)$$

Automorphisms at a point, with $g = h$, appear naturally when one considers the action of diffeomorphisms. Namely, if φ is a diffeomorphism fixing a point x_0 , then in local coordinates (as in Example 3.1.3 above) the right action on $J^2_{x_0}$, induced by the pull-back φ^* on 2-jets, is an automorphism.

REMARK 3.1.5. Despite this last remark, automorphisms of the 2-jet bundle $J^2(X)$, even those with $g = h$, have little to do with global diffeomorphisms or global changes of coordinates. In fact an automorphism radically restructures $J^2(X)$ in that the image of an integrable section (one obtained by taking $J^2 u$ for a fixed smooth function u on X) is essentially never integrable.

The automorphism group $\text{Aut}(J^2(X))$ can be naturally extended by the fibre-wise translations. Recall that the group of affine transformations of a vector space V is the product $\text{Aff}(V) = \text{GL}(V) \times V$ acting on V by $(g, v)(u) = g(u) + v$. The group law is $(g, v) \cdot (h, w) = (gh, v + g(w))$. There is a short exact sequence

$$0 \rightarrow V \rightarrow \text{Aff}(V) \xrightarrow{\pi} \text{GL}(V) \rightarrow \{I\}.$$

DEFINITION 3.1.6. The **affine automorphism group** of $J^2(X)$ is the space of smooth sections of

$$\pi^{-1}\{\text{Aut}(J^2(X))\} \subset \text{Aff}(J^2(X))$$

where π is the surjective bundle map $\pi : \text{Aff}(J^2(X)) \rightarrow \text{GL}(J^2(X))$.

Note that any affine automorphism can be written in the form

$$\Psi = \Phi + S \tag{3.1.10}$$

where Φ is a (linear) automorphism and S is a section of the bundle $J^2(X)$.

3.2. Jet-Equivalence.

DEFINITION 3.2.1. Two subequations $F, F' \subset J^2(X)$ are said to be **jet-equivalent** if there exists an automorphism $\Phi : J^2(X) \rightarrow J^2(X)$ with $\Phi(F) = F'$. If this holds for an affine automorphism $\Psi = \Phi + S$, they are said to be **affinely jet-equivalent**.

REMARK 3.2.2. A jet-equivalence $\Phi : F \rightarrow F'$ does not take F -subharmonic functions to F' -subharmonic functions. In fact as mentioned above, for $u \in C^2$, $\Phi(J^2 u)$ is almost never the 2-jet of a function. It happens if and only if $\Phi(J^2 u) = J^2 u$. Nevertheless, if $\Psi = \Phi + S$ is an affine automorphism of $J^2(X)$ and $F \subset J^2(X)$ is a closed set, then

$$F \text{ is a subequation} \iff \Psi(F) \text{ is a subequation},$$

and furthermore, by 2.2.5,

$$\widetilde{\Psi(F)} = \Phi(\tilde{F}) - S,$$

which is basic in establishing comparison.

DEFINITION 3.2.3. We say that a subequation $F \subset J^2(X)$ is *locally affinely jet-equivalent to a constant coefficient subequation \mathbf{F}* if each point x has a local coordinate neighborhood U such that, in the canonical trivialization (3.1.8) of $J^2(U)$ determined by those coordinates, F is affinely jet-equivalent to the constant coefficient subequation $U \times \mathbf{F}$.

This concept is robust as shown by the following lemma, whose proof is a straightforward calculation.

LEMMA 3.2.4. *If F is affinely jet-equivalent to \mathbf{F} in some local coordinate trivialization of $J^2(U)$, then this is true in every local coordinate trivialization of $J^2(U)$.*

A basic reason for introducing this concept is the following (see [HL6, Prop. 6.9]). Let X be a riemannian manifold with topological G -structure for a subgroup $G \subset O(n)$ (see (8.2.1)).

PROPOSITION 3.2.5. *Suppose that $F \subset J^2(X)$ is the subequation determined by a G -invariant constant coefficient subequation $\mathbf{F} \subset \mathbf{J}^2$ (cf. 2.2.3 and 8.2). Then F is locally jet-equivalent to \mathbf{F} on X .*

EXAMPLE 3.2.6. (Universal Equations) Basic examples come from universal riemannian equations ($G = O(n)$) such as those given in Example 2.1.2 (a), (b), (c), and their complex analogues on almost complex hermitian manifolds ($G = U(n)$) or the analogues on almost quaternionic hermitian manifolds ($G = Sp(n)$). There are also the other branches of these equations as discussed in Note 2.1.10. There are also the many geometric examples coming from Lagrangian geometry and calibrated geometry which are discussed below.

EXAMPLE 3.2.7. (Inhomogeneous Equations) Another important fact about affine jet equivalence is that it can transform inhomogeneous equations into constant coefficient ones and vice versa. We present several illustrative examples here (and more in 8.5). They each have the structure $F = \Psi(H)$, $H = \Psi^{-1}(F)$ where F is a pure second-order, universal riemannian subequation, and

$$\Psi(A) \equiv hAh^t + S = \eta^2 A + S$$

where $h(x) = \eta(x)\text{Id}$, for $\eta : X \rightarrow \mathbf{R}$, and $S : X \rightarrow \text{Sym}^2 T^*(X)$ is a translation term.

- (i) Let F correspond to the k^{th} branch $\{\lambda_k(\text{Hess } u) = 0\}$ of the homogeneous Monge-Ampère equation (see 2.1.10). Taking $\eta \equiv 1$ and $S = -f(x)\text{Id}$ shows that F is affinely jet-equivalent to the inhomogeneous equation

$$\lambda_k(\text{Hess } u) = f(x)$$

for any smooth function f . This includes the Monge-Ampère equation from 2.1.2(b) when written as $\lambda_{\min}(\text{Hess } u) = 0$.

- (ii) Let F correspond to the universal equation $\det(\text{Hess } u) = 1$ with $\text{Hess } u \geq 0$. One can transform this to the inhomogeneous equation

$$\det(\text{Hess } u) = f(x) \quad \text{with } \text{Hess } u \geq 0$$

for any smooth $f > 0$ by choosing $\eta = f^{-\frac{1}{2n}}$ and $S = 0$.

- (iii) More generally, one can transform the universal subequation: $\sigma_k(\text{Hess } u) = 1$ and $\sigma_\ell(\text{Hess } u) \geq 0$, $1 \leq \ell < k$, into the inhomogeneous equation

$$\sigma_k(\text{Hess } u) = f(x) \quad \text{and} \quad \sigma_\ell(\text{Hess } u) \geq 0, \quad 1 \leq \ell < k$$

for any smooth $f > 0$ by choosing $\eta = f^{-\frac{1}{2k}}$ and $S = 0$.

EXAMPLE 3.2.8. (The Calabi-Yau Equation) Let X be an almost complex hermitian manifold (a Riemannian U_n -manifold), and consider the subequation $F \subset J^2(X)$ determined by the euclidean subequation:

$$\det_{\mathbf{C}}\{A_{\mathbf{C}} + I\} \geq 1 \quad \text{and} \quad A_{\mathbf{C}} + I \geq 0$$

where $A_{\mathbf{C}} \equiv \frac{1}{2}(A - JAJ)$ is the hermitian symmetric part of A . Let $f > 0$ be a smooth positive function on X and write $f = h^{-2n}$. Consider the global affine automorphism of $J^2(X)$ given by

$$\Psi(r, p, A) = (r, p, h^2 A + (h^2 - 1)I)$$

and set $F_f = \Psi^{-1}(F)$. Then

$$\begin{aligned} (r, p, A) \in F_f &\iff \det_{\mathbf{C}}\{h^2(A_{\mathbf{C}} + I)\} \geq 1 \quad \text{and} \quad h^2(A_{\mathbf{C}} + I) \geq 0 \\ &\iff \det_{\mathbf{C}}\{(A_{\mathbf{C}} + I)\} \geq f \quad \text{and} \quad (A_{\mathbf{C}} + I) \geq 0 \end{aligned}$$

so we see that the F_f -harmonic functions are functions u with $\det_{\mathbf{C}}\{\text{Hess}_{\mathbf{C}} u + I\} = f$ and $\text{Hess}_{\mathbf{C}} u + I \geq 0$ (quasi-plurisubharmonic).

If X is actually a complex manifold of dimension n with Kähler form ω , this last equation can be written in the more familiar form

$$(i\partial\bar{\partial}u + \omega)^n = f\omega^n$$

with u quasi-plurisubharmonic.

One can similarly treat the equation

$$(i\partial\bar{\partial}u + \omega)^n = e^u f\omega^n.$$

or the same equation with e^u replaced by any non-decreasing positive function $F(u)$.

The concept of affine jet equivalence plays a critical role in the study of intrinsically subharmonic functions on almost complex manifolds [HL10].

4. Monotonicity.

A concept of fundamental importance here is that of a *monotonicity cone* for a given subequation. It is the key to establishing comparison and removable singularity theorems for equations which are highly non-convex.

4.1. The Constant Coefficient Case. Let $\mathbf{F}, \mathbf{M} \subset \mathbf{J}^2$ be constant coefficient subequations.

DEFINITION 4.1.1. We say that \mathbf{M} is a **monotonicity subequation** for \mathbf{F} if

$$\mathbf{F} + \mathbf{M} \subset \mathbf{F}. \quad (4.1.1)$$

It follows directly from 2.1.6 that the sum of an \mathbf{F} -subharmonic function and an \mathbf{M} -subharmonic function is again \mathbf{F} -subharmonic, provided that one of them is smooth. Thus, the reader can see that monotonicity is related to approximation whenever \mathbf{M} has the *cone property*

$$t\mathbf{M} \subset \mathbf{M} \quad \text{for } 0 \leq t \leq 1.$$

When this holds M can be expanded so that each fibre is a convex cone with vertex at the origin (cf. 4.1.4). Under this added assumption \mathbf{M} is called a **monotonicity cone**.

LEMMA 4.1.2. *If \mathbf{M} is a monotonicity cone for \mathbf{F} , then*

$$\mathbf{F} + \mathbf{M} \subset \widetilde{\mathbf{F}} \quad \text{and} \quad (4.1.2)$$

$$\mathbf{F} + \widetilde{\mathbf{F}} \subset \widetilde{\mathbf{M}}. \quad (4.1.3)$$

These elementary facts are basic. The first states that:

\mathbf{M} is a monotonicity cone for $\mathbf{F} \iff \mathbf{M}$ is a monotonicity cone for $\widetilde{\mathbf{F}}$.

The second is the algebraic precursor to proving that:

The sum of an \mathbf{F} -subharmonic function and an $\widetilde{\mathbf{F}}$ -subharmonic function is $\widetilde{\mathbf{M}}$ -subharmonic.

If one of the two functions is smooth, this last result follows easily from the definitions. It is important, because in most cases, the $\widetilde{\mathbf{M}}$ -subharmonic functions satisfy the following:

Zero Maximum Principle: For any compact set K in the domain of u ,

$$u \leq 0 \text{ on } \partial K \quad \Rightarrow \quad u \leq 0 \text{ on } K. \quad (\text{ZMP})$$

EXAMPLE 4.1.3. The (ZMP) holds for $\widetilde{\mathbf{M}}$ -subharmonic functions when

$$\mathbf{M} = \{(r, p, A) \in \mathbf{J}^2 : r \leq -\gamma|p|, p \in \mathcal{D} \text{ and } A \geq 0\}$$

where $\gamma > 0$ and $\mathcal{D} \subset \mathbf{R}^n$ is a convex cone with non-empty interior (and vertex at 0). See Appendix B for a proof and further discussion of Examples. Note incidentally that the smaller M is, the easier it is to be a monotonicity cone for F , while the larger \widetilde{M} is, the harder it is to satisfy (ZMP).

NOTE 4.1.4. Associated to any subequation \mathbf{F} is the set $\mathbf{M}_\mathbf{F}$ of all $J \in \mathbf{J}^2$ such that $\mathbf{F} + tJ \subset \mathbf{F}$ for $0 \leq t \leq 1$. One checks easily that $\mathbf{M}_\mathbf{F}$ is a closed convex cone which satisfies (P) and (N). Thus, if $\text{Int}\mathbf{M}_\mathbf{F} \neq \emptyset$, it is the maximal monotonicity cone for \mathbf{F} .

4.2. The General Case. Let $F \subset J^2(X)$ be a subequation on a manifold X .

DEFINITION 4.2.1. A **monotonicity cone** for F is a convex cone subequation $M \subset J^2(X)$ (each fibre is a convex cone with vertex at the origin) satisfying the condition

$$F + M \subset F \quad (4.2.1)$$

LEMMA 4.2.2. *If M is a monotonicity cone for F , then*

$$\widetilde{F} + M \subset \widetilde{F} \quad \text{and} \quad (4.2.2)$$

$$F + \widetilde{F} \subset \widetilde{M}. \quad (4.2.3)$$

NOTE 4.2.3. Suppose $\mathbf{F} \subset \mathbf{J}^2$ is a constant coefficient subequation invariant under a subgroup $G \subset \text{O}(n)$. Then $\mathbf{M}_\mathbf{F}$ is also G -invariant. Thus if $\text{Int}\mathbf{M}_\mathbf{F} \neq \emptyset$, it determines a monotonicity cone M_F for every subequation F canonically determined on any manifold with a topological G -structure (cf. Remark 2.2.3).

4.3. Examples. (Branches of Polynomial Equations) Many subequations have naturally associated monotonicity cones. The most basic case is the following.

EXAMPLE 4.3.1. (Homogeneous Monge Ampère Equations) Let $K = \mathbf{R}, \mathbf{C}$ or \mathbf{H} and let $K^n = \mathbf{R}^N$ for $N/n = 1, 2$, or 4 . Then any quadratic form $A \in \text{Sym}^2(\mathbf{R}^N)$ has a K -hermitian symmetric part A_K defined in Remark 2.1.3. Let $\lambda_1^K(A) \leq \dots \leq \lambda_n^K(A)$ be the ordered eigenvalues of A_K

(where we ignore the natural multiplicities 2 in the complex case and 4 in the quaternion case). Let

$$\Lambda_k^K \equiv \{\lambda_k^K(A) \geq 0\}$$

denote the k^{th} branch of the homogeneous Monge-Ampère equation (cf. Note 2.1.10). The dual subequation is $\tilde{\Lambda}_k^K = \Lambda_{n-k+1}^K$. These subequations carry over to any riemannian manifold with orthogonal almost complex or quaternionic structures.

The smallest, most basic branch is $\Lambda_1^K = \{A^K \geq 0\} = \mathbf{F}(G(1, K^n))$, which will be denoted by \mathcal{P}^K , $K = \mathbf{R}, \mathbf{C}$ or \mathbf{H} . The monotonicity of ordered eigenvalues: $\lambda_k^K(A) \leq \lambda_k^K(A + P)$ for $P \in \mathcal{P}^K$ implies that

$$\Lambda_k^K + \mathcal{P}^K \subset \Lambda_k^K,$$

i.e., the top branch \mathcal{P}^K is a monotonicity cone for each branch Λ_k^K of the Monge-Ampère equation.

EXAMPLE 4.3.2. (p -Convexity) Fix p , $1 \leq p \leq n$. For each $A \in \text{Sym}^2(\mathbf{R}^n)$ and each p -tuple $I = \{i_1 < i_2 < \dots < i_p\}$, set $\lambda_I(A) = \lambda_{i_1}(A) + \dots + \lambda_{i_p}(A)$. Consider the second-order polynomial differential equation determined by

$$\text{MA}_p(A) \equiv \prod_I \lambda_I(A) = \det \{D_A : \Lambda^p \mathbf{R}^n \rightarrow \Lambda^p \mathbf{R}^n\} = 0$$

where D_A denotes A acting as a derivation on the exterior power $\Lambda^p \mathbf{R}^n$. This equation splits into branches $\Lambda_k(p)$, $k = 1, \dots, \binom{n}{p}$, obtained by ordering the eigenvalues $\{\lambda_I(A)\}$. The *principle branch* $\Lambda_1(p)$, which is denoted by

$$\mathcal{P}(p) \equiv \{A : \lambda_1(A) + \dots + \lambda_p(A) \geq 0\} = \mathbf{F}(G(p, \mathbf{R}^n)),$$

is exactly the one considered in 2.1.11(c). In particular, the $\mathcal{P}(p)$ -subharmonic functions are just the *p -plurisubharmonic* functions—those which are harmonic on all affine p -planes. The monotonicity of eigenvalues shows that $\mathcal{P}(p)$ is a monotonicity cone for every branch of this equation, that is,

$$\Lambda_k(p) + \mathcal{P}(p) \subset \Lambda_k(p).$$

More generally, let $K = \mathbf{R}, \mathbf{C}$ or \mathbf{H} and, using the notation of 4.3.1, set

$$\text{MA}_p^K(A) \equiv \prod_I \lambda_I^K(A).$$

This defines a polynomial differential equation with principal branch $\mathcal{P}^K(p) = \mathbf{F}(G(p, K^n))$. The other branches, obtained as above by ordering the eigenvalues $\{\lambda_I^K(A)\}$, are subequations for which $\mathcal{P}^K(p)$ is a monotonicity cone.

The cone $\mathcal{P}(p)$ can be defined for any real number p , $1 \leq p \leq n$ by

$$\mathcal{P}(p) \equiv \{A : \lambda_1(A) + \dots + \lambda_{[p]}(A) + (p - [p])\lambda_{p+1}(A) \geq 0\}. \quad (4.3.1)$$

This extension plays an important role in removable singularity theorems (see Section 6.2 below). We note that this extended $\mathcal{P}(p)$ is the principal

branch of the polynomial operator $\text{MA}_p(A) = \prod(\lambda_I(A) + (p - [p])\lambda_k(A))$ where the product is over $|I| = [p] - 1$ and $k \notin I$.

EXAMPLE 4.3.3. (δ -Uniform Ellipticity) A basic family of monotonicity subequations is given by

$$\mathcal{P}(\delta) \equiv \{A \in \text{Sym}^2(\mathbf{R}^n) : A \geq -\delta \text{tr}A \cdot I\}$$

for $\delta > 0$. Any subequation \mathbf{F} , for which $\mathcal{P}(\delta)$ is a monotonicity cone, is uniformly elliptic in the usual sense. This subequation is the principal branch of the pure second-order polynomial differential equation:

$$\prod_{i=1}^n (\lambda_k(\text{Hess } u) + \delta \Delta u) = 0.$$

This equation has n branches

$$\lambda_k(\text{Hess } u) + \delta \Delta u \geq 0 \quad \text{for } k = 1, \dots, n,$$

and $\mathcal{P}(\delta)$ is a monotonicity cone for each of these branches, so in particular, each branch is uniformly elliptic.

This is easily generalized as follows. Suppose $\mathbf{F} \subset \text{Sym}^2(\mathbf{R}^n)$ is any pure second-order subequation. Then for each $\delta > 0$, the δ -elliptic regularization $\mathbf{F}(\delta)$ is defined by requiring that $A + \delta(\text{tr}A) \cdot I \in \mathbf{F}$. Now if \mathbf{M} is a monotonicity cone for \mathbf{F} , it follows immediately from the definitions that $\mathbf{M}(\delta)$ is a monotonicity cone for $\mathbf{F}(\delta)$. Also, $\mathcal{P} \subset \mathbf{M}$ implies that $\mathcal{P}(\delta) \subset \mathbf{M}(\delta)$, which ensures that each $\mathbf{F}(\delta)$ is uniformly elliptic.

EXAMPLE 4.3.4. Gårding Hyperbolic Polynomials) The examples above, and several below, fall into a general class of equations where monotonicity cones appear naturally. A homogeneous polynomial $Q : \text{Sym}^2(\mathbf{R}^n) \rightarrow \mathbf{R}$ of degree m is said to be *Gårding hyperbolic with respect to the identity* if $Q(I) = 1$ and for each $A \in \text{Sym}^2(\mathbf{R}^n)$ the polynomial $q_A(t) \equiv Q(tI + A)$ has m real roots. Thus we can write

$$Q(tI + A) = \prod_{k=1}^m (t + \lambda_k(A))$$

where the $\lambda_1(A) \leq \dots \leq \lambda_m(A)$ are the ordered eigenvalues (the negatives of the roots) of $q_A(t)$. Such a polynomial has m branches

$$\Lambda_{Q,k} \equiv \{\lambda_k(A) \geq 0\}, \quad k = 1, \dots, m,$$

which correspond to m constant coefficient pure second-order subequations in \mathbf{R}^n . The principal branch

$$\mathbf{M}_Q \equiv \Lambda_{Q,1}$$

is called the *Gårding cone*. Gårding's beautiful theory of hyperbolic polynomials [G] applies here to give the following.

PROPOSITION 4.3.5. *The Gårding cone \mathbf{M}_Q is a convex cone containing the identity I . It satisfies the property*

$$\Lambda_{Q,k} + \mathbf{M}_Q \subset \Lambda_{Q,k} \text{ for all } k = 1, \dots, m,$$

that is, \mathbf{M}_Q gives a monotonicity cone for each of the subequations $\Lambda_{Q,k}$. In particular, as long as \mathbf{M}_Q contains \mathcal{P} , each branch $\Lambda_{Q,k}$ of Q is a subequation.

One of the simplest examples comes by taking $Q(A) = \sigma_m(A)$, the m^{th} elementary symmetric function in the eigenvalues. Here the Gårding cone \mathbf{M}_Q is the set $\{\sigma_1 \geq 0, \dots, \sigma_m \geq 0\}$ (cf. Example 2.1.2(c)).

In general, for any hyperbolic polynomial Q as above, one can construct large families of associated subequations, equipped with monotonicity cones, by using the eigenvalues of Q . For a discussion of this as well as an elementary introduction to Gårding's theory, see [HL7,8].

4.4. Monotonicity and Duality. The key algebraic fact that the dual of a translated subequation $F + J$ is just $\tilde{F} - J$ (see 2.1.5) easily proves the following result, which in turn proves the basic algebraic lemmas 4.1.2 and 4.2.2.

LEMMA 4.4.1. *Given three subequations $G, M, F \subset J^2(X)$, the fibre-wise sums satisfy:*

$$G + M \subset F \iff G + \tilde{F} \subset \tilde{M}. \quad (4.4.1)$$

PROOF. Note that $J + M \subset F \iff M \subset -J + F \iff J + \tilde{F} \subset \tilde{M}$. \square

Later on, (4.4.1) will be implemented with $G = F^c \subset F$ (cf. (5.1.1)) to obtain weak comparison (see Remark 5.1.4).

4.5. Uniform Ellipticity as Monotonicity. As noted in Example 4.4.3 the classical notion of uniform ellipticity can be reformulated in terms of monotonicity. We now examine this in greater detail. Suppose that F is a subequation defined on an open set $X \subset \mathbf{R}^n$, in the classical way, by $F \equiv \{f(x, r, p, A) \geq 0\}$ for a function $f : J^2(X) \rightarrow \mathbf{R}$ (cf. Appendix A). Then uniform ellipticity (with constants $0 < \lambda < \Lambda$) is the condition that for $A, P \in \text{Sym}^2(\mathbf{R}^n)$ with $P \geq 0$,

$$\lambda \text{tr}(P) \leq f(x, r, p, A + P) - f(x, r, p, A) \leq \Lambda \text{tr}(P) \quad (4.5.1)$$

(and is usually combined with Lipschitz continuity in p). This condition can be reformulated in terms of a monotonicity subequation for F . To see this it suffices to consider the simplest case $f : \text{Sym}^2(\mathbf{R}^n) \rightarrow \mathbf{R}$. The condition (4.5.1) is equivalent to requiring that for all A, B (not just $B \geq 0$),

$$\mathcal{P}_{\lambda, \Lambda}^-(B) \leq f(A + B) - f(A) \leq \mathcal{P}_{\lambda, \Lambda}^+(B) \quad (4.5.1)'$$

where $\mathcal{P}_{\lambda, \Lambda}^\pm$ are the *Pucci operators* defined by

$$\mathcal{P}_{\lambda, \Lambda}^-(B) \equiv \lambda \text{tr}(B^+) + \Lambda \text{tr}(B^-) \quad \text{and} \quad \mathcal{P}_{\lambda, \Lambda}^+(B) \equiv -\mathcal{P}_{\lambda, \Lambda}^-(B)$$

and where $B = B^+ + B^-$ is the decomposition into $B^+ \geq 0$ and $B^- \leq 0$. It is easy to see that the left hand inequality in (4.5.1)' for all A, B is equivalent to the right hand inequality for all A, B . The desired monotonicity is given by the **Pucci cone**

$$\mathbf{P}_{\lambda,\Lambda} \equiv \{B \in \text{Sym}^2(\mathbf{R}^n) : \mathcal{P}_{\lambda,\Lambda}^-(B) \geq 0\}. \quad (4.5.2)$$

Note that the left hand inequality in (4.5.1)' implies the monotonicity:

$$F + \mathbf{P}_{\lambda,\Lambda} \subset F. \quad (4.5.3)$$

The equivalence of $F + \mathbf{P}_{\lambda,\Lambda} \subset F$ and $\tilde{F} + \mathbf{P}_{\lambda,\Lambda} \subset \tilde{F}$ corresponds to the equivalence of the right and left hand inequalities in (4.5.1)'.

The Pucci cones are convex. One way to see this is to compute that $\mathbf{P}_{\lambda,\Lambda}$ is the polar of the convex cone on the set $\{B \in \text{Sym}^2(\mathbf{R}^n) : \lambda I \leq B \leq \Lambda I\}$.

We point out that Pucci cones provide just one of many choices of a family of monotonicity subequations (convex cones) which form a “fundamental” neighborhood system of $\mathcal{P} = \{A \geq 0\}$, e.g. Example 4.3.3 above. All such families give equivalent notions of uniform ellipticity.

5. Comparison and Strict Approximation

Let $F \subset J^2(X)$ be a subequation on a manifold X and for each compact set $K \subset X$ set $F(K) = \text{USC}(K) \cap F(\text{Int } K)$.

DEFINITION 5.0.1. We say that **comparison holds** for F on X if for every compact subset K , the Zero Maximum Principle

$$u + v \leq 0 \text{ on } \partial K \quad \Rightarrow \quad u + v \leq 0 \text{ on } K \quad (\text{ZMP})$$

holds for all

$$u \in F(K) \quad \text{and} \quad v \in \tilde{F}(K).$$

One sees easily that comparison implies **uniqueness for the Dirichlet problem**:

If u and v are F -harmonic on $\text{Int } K$ and $u = v$ on ∂K , then $u = v$ on K

5.1. Weak Comparison. A C^2 function u on X is said to be **strictly** F -subharmonic if $J_x^2 u \in \text{Int } F_x$ for all x . This notion has the following useful extension to functions which are not C^2 . For $c > 0$ let F^c be the subequation with fibres

$$F_x^c \equiv \{J \in F_x : \text{dist}(J, \sim F_x) \geq c\} \quad (5.1.1)$$

where dist denotes distance in the fibre $J_x^2(X)$. This set satisfies conditions (P) and (N). A function $u \in \text{USC}(X)$ is called **strictly** F -subharmonic if each x has a neighborhood U and $c > 0$ such that u is F^c -subharmonic on U .

DEFINITION 5.1.1. We say that **weak comparison holds** for F on X if for every compact subset K ,

$$u + v \leq 0 \text{ on } \partial K \quad \Rightarrow \quad u + v \leq 0 \text{ on } K$$

holds for all

$$u \in F^c(K), \quad v \in \tilde{F}(K) \quad \text{and} \quad c > 0.$$

We say that **local weak comparison holds** for F on X if every point has a neighborhood in which weak comparison holds. This weakened form of comparison has several advantages. The first is the following.

THEOREM 5.1.2. (*Local implies Global*) *If local weak comparison holds on X , then weak comparison holds on X .*

A second important advantage is the following.

THEOREM 5.1.3. *Suppose F is a subequation on X which is locally jet-equivalent to a constant coefficient subequation. Then local weak comparison holds for F on X .*

REMARK 5.1.4. F^c is exactly the subset of F which satisfies the “weak monotonicity”

$$F^c + M^c \subset F \quad \text{and hence} \quad F^c + \tilde{F} \subset \tilde{M}^c$$

where M^c is the universal subequation corresponding to the constant coefficient subequation

$$\mathbf{M}^c \equiv (-\infty, 0] \times \overline{B(0, c)} \times (\mathcal{P} - c \cdot I).$$

The smaller subequation $\mathbf{M}_c \subset \mathbf{M}^c$ defined by

$$\mathbf{M}_c \equiv (-\infty, 0] \times \overline{B(0, c)} \times \mathcal{P}$$

has dual $\tilde{\mathbf{M}}_c \supset \tilde{\mathbf{M}}^c$ which satisfies the (ZMP). It is the union of three subequations:

$$\begin{aligned} \mathbf{R}_- \times \mathbf{R}^n \times \text{Sym}^2(\mathbf{R}^n) &\quad (\text{zeroth order}) \\ \mathbf{R} \times (\sim B(0, c)) \times \text{Sym}^2(\mathbf{R}^n) &\quad (\text{dual Eikonal}) \\ \mathbf{R} \times \mathbf{R}^n \times \tilde{\mathcal{P}} &\quad (\text{subaffine}), \end{aligned}$$

5.2. Strict Approximation. We say that **strict approximation** holds for F on X if for each compact set $K \subset X$, each function $u \in F(K)$ can be uniformly approximated by functions in $F(K)$ which are strict on $\text{Int}K$.

THEOREM 5.2.1. *If weak comparison and strict approximation hold for F on X , then comparison holds for F on X .*

THEOREM 5.2.2. *Let F be a subequation on X with a monotonicity cone subequation M . Suppose X carries a C^2 -function which is strictly M -subharmonic. Then local weak comparison implies global comparison for F on X .*

The idea is to approximate $u \in F(K)$ by $u + \epsilon\psi$, $\epsilon > 0$, where ψ is the strictly M -subharmonic function. (The proofs of these theorems can be found in [HL6].)

Thus we see that monotonicity subequations are of central importance in solving the Dirichlet Problem for nonlinear equations which are degenerate and highly non-convex.

There are times when strict approximation can be achieved by other means. One example is given by the Eikonal subequation $|\nabla u| \leq 1$. Here the family of functions $u_\epsilon = (1 - \epsilon)u$ for $\epsilon > 0$ gives strict approximation.

5.3. Addition Theorems. In [HL4] the following results were proved for pure second-order, constant coefficient subequations on an open subset $X \subset \mathbf{R}^n$. We recall that a function u on an open set in \mathbf{R}^n is *quasi-convex* if the function $u(x) + c|x|^2$ is convex for some $c > 0$. Local quasi-convexity is invariant under coordinate changes and therefore makes sense on manifolds.

Suppose u is locally quasi-convex on X . Then

$$u \in \mathbf{F}(X) \iff D_x^2 u \in \mathbf{F} \text{ a.e. on } X.$$

If $\mathbf{F} + \mathbf{G} \subset \mathbf{H}$, then for quasi-convex functions u and v ,

$$u \in \mathbf{F}(X) \text{ and } v \in \mathbf{G}(X) \Rightarrow u + v \in \mathbf{H}(X).$$

Both of these results hold in much greater generality.

THEOREM 5.3.1. (AE Theorem) *Suppose F is a subequation (in the sense of Definition 2.2.1) on a manifold X , and suppose u is locally quasi-convex on X . Then*

$$u \in F(X) \iff J_x^2 u \in F_x \text{ a.e. on } X.$$

THEOREM 5.3.2. (Quasi-Convex Addition) *Given three subequations F , G and H (as in 5.3.1) with $F + G \subset H$, one has that*

$$u \in F(X) \text{ and } v \in G(X) \Rightarrow u + v \in H(X).$$

for locally quasi-convex functions u and v .

Theorem 5.3.1 follows in an elementary manner from either Jensen's Lemma [J₁] or Slodkowski's Lemma [S₁] (in fact, they are equivalent). Theorem 5.3.2 is immediate from the first. These results will be elaborated in a forthcoming paper.

Of course, quasi-convex approximation can be used in the constant coefficient case to obtain the full Addition Theorem:

$$u \in \mathbf{F}(X) \text{ and } v \in \mathbf{G}(X) \Rightarrow u + v \in \mathbf{H}(X). \quad (5.3.1)$$

APPLICATION 5.3.3. (Comparison via Monotonicity for Constant Coefficient Equations) Suppose \mathbf{F} satisfies

$$\mathbf{F} + \mathbf{M} \subset \mathbf{F} \quad (5.3.2)$$

where $\widetilde{\mathbf{M}}$ -subharmonic functions satisfy the Zero Maximum Principle. From (5.3.2) we have $\mathbf{F} + \widetilde{\mathbf{F}} \subset \widetilde{\mathbf{M}}$. Therefore

$$u \in \mathbf{F}(X) \text{ and } v \in \widetilde{\mathbf{F}}(X) \Rightarrow u + v \in \widetilde{\mathbf{M}}(X),$$

and so comparison holds for \mathbf{F} .

Note that \mathbf{M} can be any of the monotonicity cones discussed in Appendix B. For example, the cone $\mathbf{M} = \mathbf{R}_- \times \mathbf{R}^n \times \mathcal{P}$ implies comparison for all gradient independent subequations.

6. Removable Singularities

Monotonicity cones lend themselves nicely to the question of removable singularities for F -subharmonic and F -harmonic functions.

6.1. M -Polar Sets. Suppose $M \subset J^2(X)$ is a convex cone subequation, i.e., one for which the fibres are convex cones with vertex at the origin.

DEFINITION 6.1.1. A closed subset $E \subset X$ is called $C^\infty M$ -**polar** if $E = \{x : \psi(x) = -\infty\}$ for some M -subharmonic function ψ which is smooth on $X - E$.

Examples.

- (a) Consider the pure second-order constant coefficient equation $\mathbf{M} = \mathcal{P}$ on \mathbf{R}^n . The \mathcal{P} -subharmonic functions are convex (See Proposition 2.1.7), and so there do not exist any $C^\infty \mathcal{P}$ -polar sets.
- (b) Consider the complex analogue \mathcal{P}^C on \mathbf{C}^n . Then \mathcal{P}^C -subharmonic functions are the standard plurisubharmonic functions and \mathcal{P}^C -polar sets are standard pluripolar sets. These exist in abundance. They include, for example, $\log|f|$ with f holomorphic.
- (c) For the quaternionic analogue \mathcal{P}^H on \mathbf{H}^n there is a 2-sphere of complex structures coming from unit imaginary quaternions. A plurisubharmonic function in any one of these structures is \mathcal{P}^H -subharmonic, and so any pluripolar set for that structure is \mathcal{P}^H -polar.
- (d) Consider the constant coefficient subequation $\mathcal{P}(p)$ defined in (4.3.1) and equal to $\mathbf{F}(G(p, \mathbf{R}^n))$ for integer p (cf. 2.1.11(c)). The following result is proved in [HL12] using the theory of classical Riesz potentials (see [L] for example).

THEOREM 6.1.2. *Any closed set of locally finite Hausdorff $(p-2)$ -measure is $\mathcal{P}(p)$ -polar.*

6.2. Removability Results. The following removable singularity results on manifolds are proved in [HL12]. Recall that M is a monotonicity cone for F if and only if it is a monotonicity cone for \tilde{F} (see 4.2.2).

THEOREM 6.2.1. *Suppose F is a subequation on X with monotonicity cone M , and $E \subset X$ is locally $C^\infty M$ -polar with no interior. Then E is removable for F -subharmonic functions which are locally bounded above across E . More precisely, if $u \in F(X - E)$ is locally bounded across E , then its canonical upper semi-continuous extension U to X is F -subharmonic on X .*

THEOREM 6.2.2. *Suppose F is a subequation on X with monotonicity cone M , and $E \subset X$ is locally C^∞ M -polar with no interior. Then for $u \in C(X)$*

$$u \text{ is } F\text{-harmonic on } X - E \quad \Rightarrow \quad u \text{ is } F\text{-harmonic on } X.$$

More generally, Theorem 6.2.1 remains true when E has interior if the extension U is defined to be $\equiv -\infty$ on $\text{Int}E$.

Theorems 6.2.1 and 6.2.2 can be applied to the many subequations given in Section 4.3. For example, this gives removable singularity results for all branches of the homogeneous complex Monge-Ampère equation on a complex hermitian manifold. Here E can be any pluripolar set (not just a C^∞ pluripolar set). The result also applies to the intrinsic notion of maximal functions on an almost complex manifold (see [HL10]).

These general results combined with Theorem 6.1.2 above give the following. We restrict attention to constant coefficient pure second-order subequations in \mathbf{R}^n .

COROLLARY 6.2.3. If F is a subequation for which $\mathcal{P}(p)$ is a monotonicity cone, then any closed set of locally finite Hausdorff $(p-2)$ -measure is removable for F - and \tilde{F} -subharmonics and F -harmonics as in the two theorems above.

This applies immediately to all branches of the equation MA_p in Example 4.3.2. It also applies to all subequations geometrically defined by a subset \mathbb{G} of the Grassmannian $G(p, \mathbf{R}^n)$. (See Example 2.1.11 and also example (c) following Theorem 2.1.12.). These include the Lagrangian and Special Lagrangian subequations in \mathbf{C}^n , the associative and coassociative subequations in \mathbf{R}^7 , and the Cayley subequations in \mathbf{R}^8 (where the appropriate value of p is clear in each case).

For the general applicability of this result we introduce the following invariant, which is studied in [HL15].

DEFINITION 6.2.4. Suppose M is a convex cone subequation. The **Riesz characteristic p_M of M** is defined to be

$$p_M \equiv \sup\{p \in \mathbf{R} : I - pP_e \in M \ \forall |e| = 1\}.$$

It has the important property that

$$\mathcal{P}(p) \subset M \iff p \leq p_M. \quad (6.2.1)$$

and hence: For any subequation F which is M -monotone, closed sets of locally finite Hausdorff $(p_M - 2)$ -measure are F -removable as above.

EXAMPLE 6.2.5. For $M = \mathbf{P}_{\lambda, \Lambda}$, the Pucci cone defined in (4.5.2), the Riesz characteristic is

$$p_M = \frac{\lambda}{\Lambda}(n-1) + 1.$$

As a consequence one retrieves the removable singularity results in [AGV]. In fact Corollary 6.2.3 is stronger since it applies to interesting equations which are not uniformly elliptic.

For $M = \mathcal{P}(\delta)$, another choice for defining uniform ellipticity, the Riesz characteristic is

$$p_M = \frac{\delta n + 1}{\delta + 1}$$

Final Remark. In the special case of convex subequations (in the general setting of manifolds) there are many interesting removability results [HL12]. They come from combining the Strong Bellman Principle (see §10) and known results ([Le], [HP_{1,2}], [H], [Shi]) for linear elliptic equations. See [HL13] for details.

7. Boundary Convexity

Fix a subequation F on a manifold X and a domain $\Omega \subset\subset X$ with smooth boundary. We shall be interested in the Dirichlet problem for F -harmonic functions on Ω . In this chapter we present geometric conditions on $\partial\Omega$ which guarantee the existence of solutions for all continuous boundary functions. These conditions are based on the following concept.

7.1. The Asymptotic Interior of a Reduced Subequation. Throughout this section we assume that F is a subequation which is “independent of the r -variable” or “reduced”. This means that with respect to the splitting

$$J^2(X) = \mathbf{R} \oplus J_{\text{red}}^2(X)$$

in (3.1.1), F is of the form $F = \mathbf{R} \times F_0$. For simplicity we just take $F \subset J_{\text{red}}^2(X)$.

DEFINITION 7.1.1. The **asymptotic interior** \overrightarrow{F} of F is the set of all $J \in J_{\text{red}}^2(X)$ for which there exists a neighborhood $\mathcal{N}(J)$ in the total space of $J_{\text{red}}^2(X)$ and a number $t_0 > 0$ such that

$$t \cdot \mathcal{N}(J) \subset F \text{ for all } t \geq t_0$$

The set \overrightarrow{F} is an open cone in $J_{\text{red}}^2(X)$ which satisfies Condition (P). If F is itself a cone, then $\overrightarrow{F} = \text{Int}F$. Otherwise, \overrightarrow{F} is smaller than $\text{Int}F$ and may be empty.

DEFINITION 7.1.2. A function $u \in C^2(X)$ is called **strictly \overrightarrow{F} -subharmonic** if $J_{\text{red},x}^2 u \in \overrightarrow{F}$ for all x .

Let $\Omega \subset X$ be a domain with smooth boundary $\partial\Omega$. By a *defining function* for $\partial\Omega$ we mean a smooth function ρ defined on a neighborhood of $\partial\Omega$ such that $\partial\Omega = \{x : \rho(x) = 0\}$, $d\rho \neq 0$ on $\partial\Omega$, and $\rho < 0$ on Ω .

DEFINITION 7.1.3. Suppose F is a reduced subequation. The boundary $\partial\Omega$ is said to be **strictly F -convex at $x \in \partial\Omega$** if there exists a strictly \overrightarrow{F} -subharmonic defining function for $\partial\Omega$ on some neighborhood of x .

This is equivalent to either of the following two conditions.

- (i) For some local defining function ρ , $J_{\text{red},x}^2\rho \in \overrightarrow{F}$.
- (ii) For any local defining function ρ , $J_{\text{red},x}^2\rho + t(d\rho)_x \circ (d\rho)_x \in \overrightarrow{F}$ for all $t \geq$ some t_0 .

7.2. General F -Convexity. Suppose now that $F \subset J^2(X)$ is a general subequation on X . For each $\lambda \in \mathbf{R}$ there is a reduced subequation $F_\lambda \subset J_{\text{red}}^2(X)$ obtained by fixing the r -variable to be λ , that is

$$F_\lambda \equiv F \cap (\{\lambda\} \times J_{\text{red}}^2(X)).$$

As above we fix a domain $\Omega \subset X$ with smooth boundary $\partial\Omega$.

DEFINITION 7.2.1. Suppose F is a general subequation. The boundary $\partial\Omega$ is said to be **strictly F -convex at $x \in \partial\Omega$** if it is strictly $\overrightarrow{F}_\lambda$ -convex at x for all $\lambda \in \mathbf{R}$.

For example, consider the universal riemannian subequation F given by $\text{Hess } u \geq 0$ and $\det\{\text{Hess } u\} \geq e^u$. Then F_λ is given by the condition that $\text{Hess } u \geq 0$ and $\det\{\text{Hess } u\} \geq e^\lambda$. One easily checks that for every λ , $\overrightarrow{F}_\lambda$ is the open cone $\{\text{Hess } u > 0\}$, and so in this case the strictly F -convex boundaries are just the classical strictly convex boundaries.

Strict F - and \overrightarrow{F} -convexity of $\partial\Omega$ at each point are sufficient for the construction of barriers used in the proof of the existence of solutions to the Dirichlet problem.

7.3. F -Convexity in Terms of the Second Fundamental Form.

For a reduced subequation F on a riemannian manifold X , the F -convexity of a boundary $\partial\Omega$ can be characterized in terms of its second fundamental form $II_{\partial\Omega}$ with respect to the outward-pointing unit normal ν . We use the decomposition given by (2.2.4):

$$J_{\text{red}}^2(X) = T^*X \oplus \text{Sym}^2(T^*X).$$

PROPOSITION 7.3.1. *The boundary $\partial\Omega$ is strictly F -convex at $x \in \partial\Omega$ if and only if*

$$(\nu, tP_\nu \oplus II_{\partial\Omega}) \in \overrightarrow{F}_x \quad \text{for all } t \geq \text{some } t_0. \quad (7.3.1)$$

where P_ν denotes orthogonal projection onto the normal line $\mathbf{R}\nu$ at x .

Note. Blocking with respect to the decomposition $T_x X = \mathbf{R}\nu \oplus T_x(\partial\Omega)$, (7.3.1) can be rewritten

$$\left((1, 0), \begin{pmatrix} t & 0 \\ 0 & II_{\partial\Omega} \end{pmatrix} \right) \in \overrightarrow{F}_x \quad \text{for all } t \geq \text{some } t_0. \quad (7.3.2)$$

7.4. Examples. (a) **k -Laplacians.** There are many examples where every boundary is strictly F -convex. The simplest one is the subequation $\Delta u \geq 0$ or more generally $\Delta u \geq f(x, u)$ where f is non-decreasing in u .

Other examples come from the constant coefficient k -Laplace subequation, defined by

$$\mathbf{F}_k^{\text{Lap}} \equiv \text{Closure} \{ (p, A) : |p|^2 \text{tr } A + (k - 2)p^t A p > 0 \} \quad (7.3.3)$$

where $k \geq 1$. These equations are self-dual. Since $\mathbf{F}_k^{\text{Lap}}$ is a cone, $\overrightarrow{\mathbf{F}}_k^{\text{Lap}} = \text{Int } \mathbf{F}_k^{\text{Lap}}$. One can check directly from (7.3.2) that for $k > 1$ every boundary is $\overrightarrow{\mathbf{F}}_k^{\text{Lap}}$ -convex.

When $k = 1$ this equation is the implicit minimal surface equation studied by De Giorgi and his school [Giu]. Here one sees that a boundary $\partial\Omega$ is strictly $\mathbf{F}_1^{\text{Lap}}$ -convex if and only if it is strictly mean convex, i.e., $\text{tr}(II_{\partial\Omega}) > 0$ at all points.

At the other extreme is the infinity Laplacian (cf. [CIL], [J₂], [ESm])

$$\mathbf{F}_{\infty}^{\text{Lap}} \equiv \text{Closure} \{ (p, A) : p^t A p > 0 \} \quad (7.3.4)$$

where again all boundaries are strictly $\mathbf{F}_{\infty}^{\text{Lap}}$ -convex.

(b) **Elementary Symmetric Functions of $\text{Hess}(u)$.** Consider Example 2.1.2(c)

$$\mathbf{F}_{\sigma_k} \equiv \{ \sigma_k(A) \geq 0, \sigma_{k-1}(A) \geq 0, \dots, \sigma_1(A) \geq 0 \} \quad (7.3.5)$$

which can be extended to the complex and quaternionic cases, and carried over to riemannian manifolds. One finds that $\partial\Omega$ is strictly \mathbf{F}_{σ_k} -convex if and only if

$$\sigma_{k-1}(II_{\partial\Omega}) > 0, \sigma_{k-2}(II_{\partial\Omega}) > 0, \dots, \sigma_1(II_{\partial\Omega}) > 0.$$

Moreover, if $\partial\Omega$ is strictly \mathbf{F}_{σ_k} -convex, then it is $\mathbf{F}_{\sigma_k,i}$ -convex for every branch $\mathbf{F}_{\sigma_k,i}$ of the equation $\sigma_k(\text{Hess } u) = 0$ (see Section 4.3). This includes the dual subequation $\widetilde{\mathbf{F}}_{\sigma_k}$, which is the bottom branch.

(c) **Geometrically Defined Subequations.** Consider now the subequations discussed in Example 2.1.11. Here the boundary convexity is particularly nice. Fix a compact subset $\mathbf{G} \subset G(p, \mathbf{R}^n)$ and define $\mathbf{F}(\mathbf{G})$ as in (2.1.6). Then a boundary $\partial\Omega$ is strictly $\mathbf{F}(\mathbf{G})$ -convex if and only if

$$\text{tr}_W \{ II_{\partial\Omega} \} > 0 \text{ for all } \mathbf{G} \text{ planes } W \text{ which are tangent to } \partial\Omega. \quad (7.3.6)$$

This condition holds automatically at $x \in \partial\Omega$ if there are no \mathbf{G} -planes tangent to $\partial\Omega$ at x .

On the other hand, if $\mathbf{G} = G(p, \mathbf{R}^n)$, then $\partial\Omega$ is strictly $\mathbf{F}(\mathbf{G})$ -convex if and only if $II_{\partial\Omega}$ has positive trace on all tangent p -planes, i.e., $\partial\Omega$ is p -convex as in [Wu], [Sha_{1,2}].

For example, suppose $\mathbf{G} \subset G(1, \mathbf{R}^2)$ is the single point $\mathbf{G} = \{x\text{-axis}\}$. Then a domain $\Omega \subset \subset \mathbf{R}^2$ with smooth boundary is strictly \mathbf{G} -convex iff the curvature vector of $\partial\Omega$ points strictly inward at every horizontal tangent.

This implies that all horizontal slices of Ω are connected. Thus, one can see directly that the Dirichlet problem for \mathbf{G} -harmonic functions ($u_{xx} = 0$) is uniquely solvable for all continuous boundary data.

A classical example comes from the set $\mathbf{G} = G_{\mathbf{C}}(1, \mathbf{C}^n) \subset G(2, \mathbf{R}^{2n})$ of complex lines in \mathbf{C}^n . A domain $\Omega \subset \mathbf{C}^n$ is strictly \mathbf{G} -convex iff it is strictly pseudo-convex in the usual sense in complex analysis (cf. [Ho1]). This is the boundary convexity required to solve the Dirichlet problem for $\mathcal{P}^{\mathbf{C}} = \mathbf{F}(\mathbf{G})$ -harmonic functions, i.e., for solutions to the homogeneous complex Monge-Ampère equation.

We note that in all cases $F(\mathbf{G}) \subset \widetilde{F(\mathbf{G})}$, so that a strictly $F(\mathbf{G})$ -convex boundary is automatically strictly $\widetilde{F(\mathbf{G})}$ -convex.

(d) **p-Plurisubharmonic Functions.** Consider now the p^{th} branch of the homogeneous complex Monge-Ampère equation. This is the pure second-order subequation given by $\Lambda_p^{\mathbf{C}} \equiv \{A : \lambda_p^{\mathbf{C}}(A) \geq 0\}$ where $\lambda_1^{\mathbf{C}}(A) \leq \dots \leq \lambda_n^{\mathbf{C}}(A)$ are the ordered eigenvalues of the hermitian symmetric part of A (see 2.1.3 and 2.1.10). The $\Lambda_p^{\mathbf{C}}$ -subharmonic functions are the classical $(p-1)$ -plurisubharmonic functions in complex analysis – those for which the complex hessian has at least $n-p+1$ non-negative eigenvalues. The Dirichlet problem for $\Lambda_p^{\mathbf{C}}$ -harmonic functions was studied by Hunt and Murray [HM] and then solved by Slodkowski [S1]. A smooth boundary $\partial\Omega \subset \mathbf{C}^n$ is strictly $\Lambda_p^{\mathbf{C}}$ -convex iff

$$\lambda_p^{\mathbf{C}}(II_{\partial\Omega}) \geq 0, \quad \text{or equivalently} \tag{7.3.7}$$

the Levi form of $\partial\Omega$ has $n-p-1$ eigenvalues ≥ 0 at each point.

(e) **Calabi-Yau-Type Equations.** Let X be a complex hermitian manifold. Consider the subequation F on X corresponding to $\det_{\mathbf{C}}(I + \text{Hess}_{\mathbf{C}}u) \geq f(x, u)$ for a continuous $f > 0$ which is non-decreasing in u and $I + \text{Hess } u \geq 0$. For $\lambda \in \mathbf{R}$ the subequation F_λ given in Section 7.2 corresponds to $\det_{\mathbf{C}}(I + \text{Hess}_{\mathbf{C}}u) \geq f(x, \lambda)$ at each point. One checks that F_λ -convexity of a boundary $\partial\Omega$ amounts to the statement that $(II_{\partial\Omega})_{\mathbf{C}} > -I$ at each point (a condition independent of λ). Levi convexity of the boundary $((II_{\partial\Omega})_{\mathbf{C}} > 0)$ will certainly suffice.

(f) **Principal curvatures of the graph.** Other equations of interest are those which impose conditions on the principal curvatures of the graph of the function u in $X \times \mathbf{R}$. See [HL6, §11.5] for a complete discussion of this case.

8. The Dirichlet Problem

Throughout this chapter $F \subset J^2(X)$ will be a subequation on a manifold X and $\Omega \subset\subset X$ will be a domain with smooth boundary $\partial\Omega$. We shall say that *existence holds for the Dirichlet Problem* for F -harmonic functions on Ω if for each continuous function $\varphi \in C(\partial\Omega)$ there exists a function $u \in C(\overline{\Omega})$ such that

- (i) u is F -harmonic on Ω , and
- (ii) $u|_{\partial\Omega} = \varphi$.

We say that *uniqueness holds for this problem* if for each $\varphi \in C(\partial\Omega)$, there exists at most one such function u .

8.1. General Theorems. It is an elementary fact that if comparison holds for F on X (see Definition 5.1), then uniqueness holds for the Dirichlet problem. Under appropriate boundary convexity comparison also implies existence.

THEOREM 8.1.1. *Suppose comparison holds for F on X . Then existence and uniqueness hold for the Dirichlet problem for F -harmonic functions on any domain $\Omega \subset\subset X$ whose boundary is both strictly F -convex and strictly \tilde{F} -convex.*

Note that u is F -harmonic if and only if $-u$ is \tilde{F} -harmonic. Thus, it is expected that both conditions, strict F and \tilde{F} convexity, should be required, if one of them is. Often one of these convexity conditions implies the other. This is clearly the case for $F = \mathcal{P}$ in \mathbf{R}^n where strict \mathcal{P} -convexity is the usual strict convexity and $\tilde{\mathcal{P}}$ -convexity is much weaker. It also holds in the case of q -plurisubharmonic functions (Example 7.4(d)) where by (7.3.7) \mathcal{P}_q^C -convexity implies $\mathcal{P}_{q'}^C$ -convexity if $q < q'$. This is reflected in the work of Hunt and Murray [HM] who noted the failure of the statement when only one convexity condition is required.

Theorems 5.1.2 and 5.2.1 imply that

If local weak comparison and strict approximation hold for F on X , then comparison holds for F on X .

THEOREM 8.1.2. *Let F be a subequation with monotonicity cone M . Suppose that:*

- (i) F is locally affinely jet-equivalent to a constant coefficient subequation, and
- (ii) X carries a strictly M -subharmonic function.

Then existence and uniqueness hold for the Dirichlet problem for F -harmonic functions on any domain $\Omega \subset\subset X$ whose boundary is both strictly F - and \tilde{F} -convex.

Comparison and therefore uniqueness follow from Theorems 5.1.3 and 5.2.2. It is then proved, using comparison and barriers constructed from boundary convexity, that existence also holds. Further details are given in §8.

Assumption (ii) is always true for pure second-order equations in \mathbf{R}^n (and in any complete simply-connected manifold of non-positive sectional curvature) since the subequation \mathcal{P} is always a monotonicity cone by the positivity condition (P) and $|x|^2$ is strictly \mathcal{P} -convex.

On the other hand something like assumption (ii) must be required in the general case. For example, suppose F is a universal riemannian equation as in 2.2.3. One could completely change the geometry (and topology) of the interior of a domain $\Omega \subset X$ without changing the F -convexity of the boundary. Take the subequation \mathcal{P} on the euclidean ball, and change the interior so that it is not contractible. Then there are no \mathcal{P} -subharmonic (riemannian convex) functions on the resulting space, and certainly no \mathcal{P} -harmonic ones.

In homogeneous spaces one can apply a trick of Walsh [W] to establish existence without uniqueness.

THEOREM 8.1.3. *Let $X = G/H$ be a riemannian homogeneous space and suppose that $F \subset J^2(X)$ is a subequation which is invariant under the natural action of G on $J^2(X)$. Let $\Omega \subset\subset X$ be a connected domain whose boundary is both F and \tilde{F} strictly convex. Then existence holds for the Dirichlet problem for F -harmonic functions on Ω .*

This theorem applies to give (the known) existence for the k -Laplacian, $1 < k \leq \infty$ on arbitrary domains, and for the 1-Laplacian on mean convex domains in G/H . The literature on these equations in \mathbf{R}^n is vast. See [JLM], [CIL], [J₂], [ESm] and references therein, for example. We note that even in \mathbf{R}^n , uniqueness for the 1-Laplacian fails catastrophically. For a generic smooth function on the boundary of the unit disk in \mathbf{R}^2 there are families of solutions to the Dirichlet problem parameterized by \mathbf{R} (and often \mathbf{R}^m for large m)!

The proof of existence in the theorems above uses the standard Perron method based on the properties in Theorem 2.3.1. Given $\varphi \in C(\partial\Omega)$, consider the family

$$\mathcal{F}(\varphi) \equiv \{u \in \text{USC}(\bar{\Omega}) \cap F(\Omega) : u \leq \varphi \text{ on } \partial\Omega\},$$

and define the *Perron function* to be the upper envelope of this family:

$$U(x) \equiv \sup_{u \in \mathcal{F}(\varphi)} u(x). \quad (8.1.1)$$

PROPOSITION 8.1.4. *Suppose that F satisfies weak comparison and that $\partial\Omega$ is both F and \tilde{F} strictly convex. Then the upper and lower semi-continuous regularizations U^* and U_* of U on $\bar{\Omega}$ satisfy:*

- (i) $U^* = U_* = U = \varphi$ on $\partial\Omega$,
- (ii) $U = U^*$ on $\bar{\Omega}$
- (iii) U is F -subharmonic and $-U_*$ is \tilde{F} -subharmonic on Ω .

The classical barrier argument, used by Bremermann [B] for the case $F = \mathcal{P}^C$, establishes (i), while weak comparison is used in (ii). Part (iii) relies on a “bump argument” found in Bedford and Taylor [BT₁] and also in [I].

When one can ultimately establish comparison, as in Theorem 8.1.2, the Perron function is the unique solution. When this is not necessarily possible,

as in Theorem 8.1.3, arguments of Walsh [W] can be applied to show that the Perron function is a solution.

In this latter case one can say more. Fix F and Ω as in Theorem 8.1.3.

Suppose

U is the Perron function for F on Ω with boundary values φ , and
 $-\tilde{U}$ is the Perron function for \tilde{F} on Ω with boundary values $-\varphi$.

Both U and \tilde{U} solve the Dirichlet problem for F -harmonic functions on Ω with boundary values φ , and if u is any other such solution,

$$\tilde{U} \leq u \leq U. \quad (8.1.2)$$

Theorems 8.1.2 and 8.1.3 have wide applications. In the following sections we will examine some specific examples.

8.2. Manifolds with Reduced Structure Group. Fix a constant coefficient subequation $\mathbf{F} \subset \mathbf{J}^2$, and let

$$G \equiv G_{\mathbf{F}} \equiv \{g \in O(n) : g(\mathbf{F}) = \mathbf{F}\} \quad (8.2.1)$$

where g acts naturally on \mathbf{J}^2 by $g(r, p, A) = (r, gp, g^t Ag)$.

DEFINITION 8.2.1. Let X be a riemannian n -manifold and $G \subset O(n)$ a subgroup. A **topological G -structure** on X is a family $\{(U_\alpha, e_\alpha)\}_\alpha$ where $\{U_\alpha\}_\alpha$ is an open covering of X and each $e_\alpha = (e_\alpha^1, \dots, e_\alpha^n)$ is a continuous tangent frame field on U_α , such that for all α, β the change of framing $g : U_\alpha \cap U_\beta \rightarrow O(n)$ takes values in G .

Each constant coefficient subequation \mathbf{F} canonically determines a subequation F on any riemannian manifold X equipped with a topological $G_{\mathbf{F}}$ -structure. (Use the splitting (2.2.4) and then the trivializations induced by the local tangent frames. The subequation determined by \mathbf{F} in these trivializations is preserved under the change of framings.) By Proposition 3.2.5, F is locally jet-equivalent to \mathbf{F} .

If \mathbf{M} is a $G_{\mathbf{F}}$ -invariant monotonicity cone for \mathbf{F} , then the corresponding subequation M on X is a monotonicity cone for F . Note that the maximal monotonicity cone for \mathbf{F} is always $G_{\mathbf{F}}$ -invariant.

THEOREM 8.2.2. *Let F be a subequation with monotonicity cone M canonically determined by \mathbf{F} and \mathbf{M} on a riemannian manifold X with a topological $G_{\mathbf{F}}$ -structure. Let $\Omega \subset\subset X$ be a domain with smooth boundary which is both F and \tilde{F} strictly convex. Assume there exists a strictly M -subharmonic function on $\bar{\Omega}$. Then the Dirichlet Problem for F -harmonic functions is uniquely solvable for all $\varphi \in C(\partial\Omega)$.*

EXAMPLE 8.2.3. (a) Universal Riemannian Subequations: As noted in Remark 2.2.3, if $G_{\mathbf{F}} = O(n)$, then \mathbf{F} universally determines a subequation on every riemannian manifold by choosing the framings e_α to be orthonormal. In particular this covers all branches of the homogeneous

Monge-Ampère equation. In fact, it covers all pure second-order subequations which depend only on the ordered eigenvalues of the Hessian. The subequation $\mathcal{P} = \{\text{Hess } u \geq 0\}$ is a monotonicity cone for all such equations. Thus Theorem 8.2.2 applies to all such F 's in any region of X where there exists a smooth strictly convex function.

Other interesting examples are given by the branches of the p -convex Monge-Ampère equation MA_p given in example 4.3.2. Here the monotonicity cone is $\mathcal{P}(p)$, and the appropriate boundary convexity is the p -convexity discussed in 7.4 (c).

Further examples come from elementary symmetric functions of $\text{Hess } u$ (see 7.4 (b) and the discussion after 4.3.5.), and functions of eigenvalues of the graph (7.4 (f)).

(b) Universal Hermitian Subequations: If $G_{\mathbf{F}} = \text{U}(n)$, then \mathbf{F} universally determines a subequation on every almost complex hermitian manifold. For example, this covers all pure second-order subequations which depend only on the ordered eigenvalues of the hermitian symmetric part $\text{Hess}_{\mathbf{C}} u$ of $\text{Hess } u$. For such equations, $\mathcal{P}^{\mathbf{C}} = \{\text{Hess}_{\mathbf{C}} u \geq 0\}$ is a monotonicity cone. Thus, for example, one has the following consequence of Theorem 8.2.2. *Let X be an almost complex hermitian manifold, and $\Omega \subset\subset X$ a smoothly bounded domain with a strictly plurisubharmonic ($\mathcal{P}^{\mathbf{C}}$ -subharmonic) defining function. Then the Dirichlet problem for every branch of the homogeneous complex Monge-Ampère equation is uniquely solvable on Ω .*

A similar result holds for branches of the equation $\text{MA}_p^{\mathbf{C}}$ where p -convexity of the Levi form on the boundary plays a role (see 7.4 (d)).

The discussion of elementary symmetric functions also carries over to this case.

Theorem 8.2.2 can similarly be applied to Calabi-Yau type equations (7.4 (e)).

All of this discussion can be replicated for almost quaternionic hermitian manifolds.

(c) Geometrically Defined Subequations: Theorem 8.2.2 applies directly to all subequations geometrically defined by a compact subset $\mathbf{G} \subset G(p, \mathbf{R}^n)$ (see 2.1.11, 2.1.12 and 7.4 (b)). Suppose X has a topological G -structure where $G = \{g \in \text{O}(n) : g(\mathbf{G}) = \mathbf{G}\}$ and let $F(\mathbf{G})$ be the corresponding subequation on X . Suppose $\Omega \subset X$ is a domain with a global defining function which is strictly \mathbf{G} -plurisubharmonic. Then the Dirichlet problem for \mathbf{G} -harmonic functions is uniquely solvable on Ω .

Thus, one can solve the Dirichlet problem for (in fact, all branches of) the Lagrangian harmonic equation (see 2.1.11 (d)) on domains with a strictly Lagrangian-plurisubharmonic defining function.

One can also solve for $G(\varphi)$ -harmonic functions on strictly $G(\varphi)$ -convex domains in a manifold with a topological calibration φ . A typical example is the following. Let X be a riemannian 7-manifold with a topological G_2 -structure determined by a global associative 3-form φ of constant comass 1.

(Such structures exist on X if and only if X is a spin manifold.) Then the Dirichlet problem for $G(\varphi)$ -harmonic functions is uniquely solvable on any domain with a strictly $G(\varphi)$ -plurisubharmonic defining function.

8.3. Inhomogeneous Equations. Since Theorem 8.1.2 assumes *affine* jet-equivalence, it applies to inhomogeneous equations as in Examples 3.2.7–8. In these cases boundary convexity and monotonicity cones are the same as in the homogeneous case.

8.4. Existence Without Uniqueness. Theorem 8.1.3 applies in cases where monotonicity cones do not exist, such as the 1-laplacians in 7.4 (a). As previously noted, solutions of the Dirichlet problem for the 1-laplacian are highly non-unique. However, they are all caught between the Perron functions U and \tilde{U} (see (8.1.2) above).

8.5. Parabolic Equations. The methods and results above carry over effectively to parabolic equations. Let X be a riemannian n -manifold with a topological G -structure for $G \subset O(n)$, and consider a constant coefficient subequation of the form

$$\mathbf{F} = \{J \in \mathbf{J}^2 : f(J) \geq 0\}$$

where $f : J^2(X) \rightarrow \mathbf{R}$ is G -invariant, \mathcal{P} - and \mathcal{N} -monotone, and Lipschitz in the reduced variables (p, A) . This induces a subequation F on X . The associated constant coefficient parabolic subequation \mathbf{H}_F on $\mathbf{R} \times X$ is defined by

$$f(J) - p_0 \geq 0$$

(where p_0 denotes the u_t component of the 2-jet of u), and it induces the associated *parabolic subequation* H_F on the riemannian product $\mathbf{R} \times X$. The H_F -harmonic functions are solutions of the equation

$$u_t = f(u, Du, D^2u).$$

Examples which can be treated include:

- (i) $f = \text{tr}A$, the standard heat equation $u_t = \Delta u$ for the Laplace-Beltrami operator.
- (ii) $f = \lambda_q(A)$, the q th ordered eigenvalue of A . This is the natural parabolic equation associated to the q th branch of the Monge-Ampère equation.
- (iii) $f = \text{tr}A + \frac{k}{|p|^2+\epsilon^2}p^tAp$ for $k \geq -1$ and $\epsilon > 0$. When $X = \mathbf{R}^n$ and $k = -1$, the solutions $u(x, t)$ of the associated parabolic equation, in the limit as $\epsilon \rightarrow 0$, have the property that the associated level sets $\Sigma_t \equiv \{x \in \mathbf{R}^n : u(x, t) = 0\}$ are evolving by mean curvature flow (cf. [ES*], [CGG*], [E] and [Gi].)
- (iv) $f = \text{tr}\{\arctan A\}$. When $X = \mathbf{R}^n$, solutions $u(x, t)$ have the property that the graphs of the gradients: $\Gamma_t \equiv \{(x, y) \in \mathbf{R}^n \times \mathbf{R}^n = \mathbf{C}^n : y = D_x u(x, t)\}$ are Lagrangian submanifolds which evolve the initial data by mean curvature flow. (See [CCH].)

Techniques discussed above show that:

Comparison holds for the subequation H_F on $X \times \mathbf{R}$.

Applying standard viscosity techniques for parabolic equations, one can prove more. Consider a compact subset $K \subset \{t \leq T\} \subset X \times \mathbf{R}$ and let $K_T \equiv K \cap \{t = T\}$ denote the terminal time slice of K . Let $\partial_0 K \equiv \partial K - \text{Int}K_T$ denote the *parabolic boundary* of K . Here $\text{Int}K$ denotes the relative interior in $\{t = T\} \subset X \times \mathbf{R}$. We say that *parabolic comparison holds for H_F* if for all such K (and T)

$$u + v \leq c \quad \text{on } \partial_0 K \quad \Rightarrow \quad u + v \leq c \quad \text{on } \text{Int}K$$

for all $u \in H_F(K)$ and $v \in \tilde{H}_F(K)$. Then one has that:

Parabolic comparison holds for the subequation H_F on $X \times \mathbf{R}$.

Under further mild assumptions on f which are satisfied in the examples above, one also has existence results. Consider a domain $\Omega \subset X$ whose boundary is strictly F - and \tilde{F} -convex. Set $K = \bar{\Omega} \times [0, T]$. Then

For each $\varphi \in C(\partial_0 K)$ there exists a unique function $u \in C(K)$ such that $u|_{\text{Int}K}$ is H_F -harmonic and $u|_{\partial_0 K} = \varphi$.

One then obtains corresponding long-time existence results.

8.6. Obstacle Problems. The methods discussed here lend themselves easily to solving boundary value problems with obstacles. Suppose that $F = \mathbf{R} \times F_0$ is a reduced subequation, i.e., independent of the r -variable. Given $g \in C(X)$, the associated obstacle subequation is defined to be

$$H \equiv (\mathbf{R}_- + g) \times F_0 \quad \text{where } \mathbf{R}_- \equiv \{r \leq 0\} \subset \mathbf{R}.$$

The following facts are easy to prove.

- The H -subharmonic functions are the F -subharmonic functions u which satisfy $u \leq g$.
- If F has a monotonicity cone $M = \mathbf{R} \times M_0$, then $M_- \equiv \mathbf{R}_- \times M_0$ is a monotonicity cone for H .
- If X carries a strictly M -subharmonic function ψ , then on any given compact set, the function $\psi - c$ is strictly (M_-) -subharmonic for $c > 0$ sufficiently large.
- If F is locally affinely jet-equivalent to a constant coefficient reduced subequation $\mathbf{R} \times F_0$, then H is locally affinely jet-equivalent to the subequation $\mathbf{R}_- \times F_0$.

Consequently, under the assumptions in Theorem 8.1.2 on a reduced subequation $F = \mathbf{R} \times F_0$ with monotonicity cone $M = \mathbf{R} \times M_0$, **comparison holds for each associated obstacle subequation $H \equiv (\mathbf{R}_- + g) \times F_0$** .

However, existence fails for a boundary function $\varphi \in C(\partial\Omega)$ unless $\varphi \leq g|_{\partial\Omega}$. Nevertheless, **if $\partial\Omega$ is both F and \tilde{F} strictly convex as in**

Theorem 8.1.2, then existence holds for each boundary function $\varphi \leq g|_{\partial\Omega}$.

To see that this is true, note the following. The Perron family for a boundary function $\varphi \in C(\partial\Omega)$ consists of those F -subharmonic functions u on Ω with $u|_{\partial\Omega} \leq \varphi$ (the usual family for F) subject to the additional constraint $u \leq g$ on Ω . The dual subequation to H is $\tilde{H} = [(\mathbf{R}_- - g) \times J^2_{\text{red}}(X)] \cup \tilde{F}$ so that the boundary $\partial\Omega$ is strictly \tilde{H} -convex if it is strictly \tilde{F} -convex. Although $\partial\Omega$ can never be strictly H -convex (since $(\vec{F}_\lambda)_x = \emptyset$ for $\lambda > g(x)$), the only place that this hypothesis is used in proving Theorem 8.1.2 for H is in the barrier construction which appears in the proof of Proposition F in [HL₆]. However, if $\varphi(x_0) \leq g(x_0)$, then the barrier $\beta(x)$ as defined in (12.1) in [HL₆] is not only F -strict near x_0 but also automatically H -strict since $\beta < g$.

The obstacle problem for the basic subequation \mathcal{P} is related to convex envelopes. This was discovered by Oberman [O] and developed by Oberman-Silvestre [OS].

9. Restriction Theorems

Let $F \subset J^2(Z)$ be a subequation on a manifold Z , and suppose $i : X \subset Z$ is a submanifold. Then there is a natural induced subequation i^*F on X given by restriction of 2-jets. For functions $u \in C^2(Z)$ one has directly that

$$u \text{ is } F\text{-subharmonic on } Z \quad \Rightarrow \quad u|_X \text{ is } i^*F\text{-subharmonic on } X.$$

Generically this induced subequation i^*F is trivial, i.e., all of $J^2(X)$. The first problem is to determine the class of submanifolds for which the restriction is interesting. In such cases we then have the following

Question: When does the implication above hold for all $u \in \text{USC}(Z)$?

Example. The situation is illustrated by the basic subequation \mathcal{P} in \mathbf{R}^n whose subharmonics are the convex functions. The restriction of a smooth convex function $u \in C^\infty(\mathbf{R}^n)$ to the unit circle in \mathbf{R}^2 obeys no proper subequation, while the restriction of u to a *minimal* submanifold $M \subset \mathbf{R}^n$, of any dimension, is subharmonic for the Laplace-Beltrami operator on M . This assertion carries over to general convex functions u .

9.1. The First General Theorem. The paper [HL₉] establishes two restriction theorems of a general nature, each of which has interesting applications. The first entails the following technical hypothesis. Fix coordinates $z = (x, y)$ on Z so that locally $X \cong \{y = y_0\}$.

The Restriction Hypothesis: Given $x_0 \in X$ and $(r_0, p_0, A_0) \in \mathbf{J}_n^2$ and given $z_\epsilon = (x_\epsilon, y_\epsilon)$ and r_ϵ for a sequence of real numbers ϵ converging to 0:

$$\text{If } \left(r_\epsilon, \left(p_0 + A_0(x_\epsilon - x_0), \frac{y_\epsilon - y_0}{\epsilon} \right), \begin{pmatrix} A_0 & 0 \\ 0 & \frac{1}{\epsilon} I \end{pmatrix} \right) \in F_{z_\epsilon}$$

$$\text{and } x_\epsilon \rightarrow x_0, \quad \frac{|y_\epsilon - y_0|^2}{\epsilon} \rightarrow 0, \quad r_\epsilon \rightarrow r_0,$$

then

$$(r_0, p_0, A_0) \in (i^*F)_{x_0}.$$

THEOREM 9.1.1. *Suppose $u \in \text{USC}(Z)$. Assume the restriction hypothesis and suppose that (i^*F) is closed. Then*

$$u \in F(Z) \quad \Rightarrow \quad u|_X \in (i^*F)(X).$$

If (i^*F) is not closed, the conclusion holds with (i^*F) replaced by $\overline{(i^*F)}$.

9.2. Applications of the First General Theorem. Theorem 9.1.1 applies to several interesting cases. In the following, the term *restriction holds* refers to the conclusion of Theorem 9.1.1. The reader is referred to [HL9] for full statements and proofs.

THEOREM 9.2.1. *Let \mathbf{F} be a constant coefficient subequation in \mathbf{R}^n . Then restriction holds for all affine subspaces X for which $i^*\mathbf{F}$ is closed.*

More generally, if u is \mathbf{F} -subharmonic, then $u|_X$ is $\overline{i^*\mathbf{F}}$ -subharmonic.

Consider now a second-order linear operator \mathbb{L} with smooth coefficients on \mathbf{R}^n . Fix linear coordinates $z = (x, y)$ and suppose $X \cong \{y = y_0\}$ as above. Using the summation convention, write

$$\mathbb{L}(u) = A_{ij}(z)u_{x_i x_j} + a_i(z)u_{x_i} + \alpha(z)u + B_{k\ell}(z)u_{y_k y_\ell} + b_k(z)u_{y_k} + C_{ik}(z)u_{x_i y_k}$$

Suppose the subequation L corresponding to $\mathbb{L}u \geq 0$ satisfies positivity. If any one of the coefficients $B(x_0, y_0)$, $b(x_0, y_0)$ or $C(x_0, y_0)$ is non-zero, restriction is trivial locally since i^*L is everything for x near x_0 . Hence, we assume the following

$$B(x, y_0), b(x, y_0), \text{ and } C(x, y_0) \text{ vanish identically on } X \quad (9.2.1)$$

THEOREM 9.2.2. *Assuming (9.2.1), restriction holds for the linear operator L to X .*

This result for linear operators proves to be quite useful.

The next result concerns geometric subequations (see Example 2.1.11) on general riemannian manifolds Z .

THEOREM 9.2.3. *Let $\mathbb{G} \subset G(p, TZ)$ be a closed subset of the bundle of tangent p -planes on Z , which admits a fibre-wise neighborhood retract (a subbundle for example). Let $F(\mathbb{G})$ be the induced subequation on Z , defined as in (2.1.6) using the riemannian hessian. Then restriction holds for all minimal \mathbb{G} -submanifolds $X \subset Z$, i.e., minimal submanifolds with $T_x X \in \mathbb{G}_x$ for all $x \in X$.*

9.3. The Second General Theorem. Let F be a subequation on a manifold Z and fix a submanifold $i : X \subset Z$ as above. In 3.2.3 we defined the notion of F being locally jet-equivalent to a constant coefficient subequation \mathbf{F} . In our current situation there is a notion of F being locally jet-equivalent to \mathbf{F} relative to the submanifold X . This entails i^*F being locally jet-equivalent to a constant coefficient subequation (assumed closed) on X . For details, see [HL₉, §§9 and 10].

THEOREM 9.3.1. *If F is locally jet-equivalent to a constant coefficient subequation relative to X , then restriction holds for F to X .*

9.4. Applications of the Second General Theorem. A nice application of Theorem 9.3.1 is the following.

THEOREM 9.4.1. *Let Z be a riemannian manifold of dimension n and $F \subset J^2(Z)$ a subequation canonically determined by an $O(n)$ -invariant constant coefficient subequation $\mathbf{F} \subset \mathbf{J}^2$. Then restriction holds for F to any totally geodesic submanifold $X \subset Z$.*

Suppose now that Z is a riemannian manifold with a topological G -structure and $F \subset J^2(Z)$ is determined by a G -invariant constant coefficient subequation as in Section 8.2. The local framings e_α appearing in Definition 8.2.1 are called **admissible**. So also is any framing of the form $e'_\alpha = ge_\alpha$ for a smooth map $g : U_\alpha \cap U_\beta \rightarrow G$. A submanifold $X \subset Z$ is said to be **compatible** with the G -structure if at every point $z \in X$ there is an admissible framing e on a neighborhood U of z such that on $X \cap U$

$$\begin{aligned} e_1, \dots, e_n &\text{ are tangent to } X \cap U \quad \text{and} \\ e_{n+1}, \dots, e_N &\text{ are normal to } X \cap U. \end{aligned}$$

For example, if $G = U(N/2)$, then any submanifold of constant CR-rank is compatible.

THEOREM 9.4.2. *Let Z be a riemannian manifold with topological G -structure, and $F \subset J^2(Z)$ a subequation canonically determined by a G -invariant constant coefficient subequation $\mathbf{F} \subset \mathbf{J}^2$. Then restriction holds for F to any totally geodesic submanifold $X \subset Z$ which is compatible with the G -structure.*

There is a further application of Theorem 9.3.1 to almost complex manifolds, which is discussed in §11.

10. Convex Subequations and the Strong Bellman Principle

An elementary fact, known to all, is that a closed convex set in a vector space V is the intersection of the closed half-spaces containing it. Put this into a family and you have a fundamental principle, which we call the *Bellman Principle*, for dealing with nonlinear pde's which are convex. Specifically, suppose $F \subset J^2(X)$ is a convex subequation—one with the property that every fibre F_x is convex. Then, under mild assumptions, F

can be written locally as the intersection of a family of **linear** subequations. These are subequations of the form

$$Lu = \langle a, D^2u \rangle + \langle b, Du \rangle + cu \geq \lambda, \quad (10.1)$$

where, from the Conditions (P) and (N) for F , one can deduce that the matrix function a and the scalar function c satisfy

$$a \geq 0 \quad \text{and} \quad c \leq 0. \quad (10.2)$$

The introduction of these local linear equations goes back to Richard Bellman and his work in dynamic programming. These equations can be found in many areas of mathematics. Examples close in spirit to those above appear in work of Bedford-Taylor [BT_{*}] and Krylov [K].

It is obviously a big improvement if all the linear equations in (10.1) needed to carve out F can be taken to have

$$a > 0, \quad (10.3)$$

for then the machinery of uniformly elliptic linear equations can be brought to bear.

More specifically: any F -subharmonic function u is locally a viscosity subsolution of $Lu \geq \lambda$. From this one sees that u is a classical subsolution (see [HL10, Thm. A.5]), and if $a > 0$, the results of [HH] apply to prove that u is L^1_{loc} . It can then be shown that u is a distributional subsolution to $Lu \geq \lambda$, and the full linear elliptic theory ([Ho₂] or [G] for example) applies.

This naturally raises the question: What assumptions on F will guarantee that it is cut out by linear equations with $a > 0$?

This question has two parts. The first concerns only the convex geometry of the fibres F_x at each point x ; in other words, the question for a convex constant coefficient subequation $\mathbf{F} \subset \mathbf{J}^2$. The second only involves the mild regularity condition that a containing half-space for F_x extends locally to a linear (variable coefficient) subequation containing F .

These questions have been discussed in [K], and an account has also been given in [HL13], where the answer to the first question is given as follows. We say that a subset $C \subset \text{Sym}^2(\mathbf{R}^n)$ depends on all the variables if there is no proper subspace $W \subset \mathbf{R}^n$ and subset $C' \subset \text{Sym}^2(W)$ such that $A \in C \iff A|_W \in C'$. Then a (constant coefficient) subequation $\mathbf{F} \subset \mathbf{J}^2 = \mathbf{R} \times \mathbf{R}^n \times \text{Sym}^2(\mathbf{R}^n)$ is said to depend weakly on all the second-order variables if for each $(r, p) \in \mathbf{R} \times \mathbf{R}^n$, the fibre $\mathbf{F}_{(r,p)} = \{A \in \text{Sym}^2(\mathbf{R}^n) : (r, p, A) \in \mathbf{F}\}$ depends on all the variables.

THEOREM 10.1. *If \mathbf{F} depends weakly on all the second-order variables, then \mathbf{F} can be written as the intersection of a family of half-space subequations $\langle a, A \rangle + \langle b, p \rangle + cr \geq \lambda$ with $a > 0$.*

NOTE 10.2. For subequations which do not depend on all the second order variables, the conclusions above fail. Consider the (geometrically determined) subequation

$$\mathbf{F} \cong \{u_{xx} \geq 0\}$$

in the (x, y) -plane. Any continuous function $u(y)$ is \mathbf{F} -subharmonic, in fact, \mathbf{F} -harmonic, but not in general L^1_{loc} .

See [HL13] for a full discussion of these matters.

11. Applications to Almost Complex Manifolds

In this section we consider completely general almost complex manifolds (X, J) where $J : TX \rightarrow TX$ is smooth bundle map with $J^2 \equiv -\text{Id}$. On any such manifold there is an *intrinsically defined* subequation

$$F(J) \subset J^2(X),$$

for which, when the structure is integrable, the $F(J)$ -subharmonic functions are exactly the standard plurisubharmonic functions. Hence, the results and techniques discussed in this paper apply to give a full-blown potential theory on almost complex manifolds, which extends the classical theory. The consequences are worked out in detail in [HL10]. Here are a few highlights.

11.1. J -Holomorphic Curves. A submanifold $Y \subset X$ is an *almost complex submanifold* if $J(T_y Y) = T_y Y$ for all $y \in Y$. In general dimensions such submanifolds exist only rarely. However, when the real dimension of Y is two, Y is called a **J -holomorphic curve**, and we have the following important classical result.

THEOREM 11.1.1. (*Nijenhuis and Woolf [NW]*) *For each point $x \in X$ and each complex tangent line $\ell \subset T_x X$, there exists a J -holomorphic curve passing through x with tangent direction ℓ .*

The restriction result 9.3.1 applies in this case to prove the following. For historical compatibility we replace the term “ $F(J)$ -subharmonic” with “ $F(J)$ -plurisubharmonic”.

THEOREM 11.1.2. *Let (Y, J_Y) be an almost complex submanifold of (X, J_X) . Then the restriction of any $F(J_X)$ -plurisubharmonic function to Y is $F(J_Y)$ -plurisubharmonic.*

This leads to the following result equating two natural definitions of plurisubharmonicity. We recall that an almost complex structure J on a 2-dimensional manifold S is always integrable, and all notions of (usc) subharmonic functions on (S, J) coincide.

THEOREM 11.1.3. *A function $u \in \text{USC}(X)$ is $F(J)$ -plurisubharmonic if and only if its restriction to every J -holomorphic curve is subharmonic.*

11.2. Completion of the Pali Conjecture. There is a third definition of J -plurisubharmonic functions on an almost complex manifold (X, J) , which makes sense for any distribution $u \in \mathcal{D}'(X)$. Any such distribution u is known to be L^1_{loc} . By work of Nefton Pali [P] we know that any $u \in \text{USC}(X)$ which is J -plurisubharmonic in the sense of Section 11.1, is L^1_{loc} on X and J -plurisubharmonic as a distribution. In the converse direction he showed

that if a J -plurisubharmonic distribution u has a continuous representative (as a $[-\infty, \infty)$ -valued function), then it is J -plurisubharmonic as above. He further conjectured that the converse should hold in general. This was proved in [HL₁₀].

The proof used the Strong Bellman Principle and involved showing that the upper semi-continuous representative of the L^1_{loc} -class obtained for each of the associated linear equations, is independent of the linear equation. It is, in fact, given by the *essential upper-semi-continuous regularization*

$$u_{\text{ess}}^*(x) \equiv \lim_{r \searrow 0} \left\{ \text{ess sup}_{B_x(r)} u \right\}$$

which depends only on the L^1_{loc} -class of u .

11.3. The Dirichlet Problem for Maximal Functions. Theorem 8.12 applies in this case to prove existence and uniqueness for the Dirichlet problem for J -maximal functions. One can show that the more classical notion of a function u being **J -maximal** (going back to [B], [W]), is the same as u being $F(J)$ -harmonic, i.e., u is $F(J)$ -(pluri)subharmonic and $-u$ is $\tilde{F}(J)$ -subharmonic. A domain $\Omega \subset\subset X$ with smooth boundary is strictly J -convex if it has a strictly $F(J)$ -plurisubharmonic defining function.

THEOREM 11.3.1. *Let $\Omega \subset\subset X$ be a strictly J -convex domain in an almost complex manifold (X, J) . Then the Dirichlet problem for J -maximal functions is uniquely solvable on Ω for all continuous boundary values $\varphi \in C(\partial\Omega)$.*

NOTE 11.3.2. Recently Szymon Plis has also studied the Dirichlet problem on almost complex manifolds [Pl]. His result is the almost-complex analogue of a main result in [CKNS]. It treats the inhomogeneous Monge-Ampère equation with positive right hand side. All data are assumed to be smooth, and complete regularity is established for the solution.

Appendix A. A Pocket Dictionary

The conventions adopted in this paper (and related ones) are not common in the literature, but they have several advantages, particularly for applications to calibrated geometry and to branches of polynomial operators. In the case of comparison the advantage is discussed in Comment 3 below.

For readers hard-wired to standard notation (as in, say, [CIL]), we give here a concise translation of concepts to serve as a guide.

Classically, a fully nonlinear partial differential equation for a smooth function $u(x)$ on an open set $X \subset \mathbf{R}^n$ is written in the form

$$f(x, u, Du, D^2u) = 0$$

for a given continuous function $f : X \times \mathbf{R} \times \mathbf{R}^n \times \text{Sym}^2(\mathbf{R}^n) \rightarrow \mathbf{R}$.

Here the function f is typically replaced by the closed set

$$F \equiv \{(x, r, p, A) : f(x, r, p, A) \geq 0\}.$$

For C^2 -functions $u(x)$ we have the following translations. Set $J_x^2 u \equiv (x, u, Du, D^2 u)$.

$$\begin{array}{lll}
u \text{ is a } \mathbf{subsolution} & \longleftrightarrow & u \text{ is } F \mathbf{subharmonic}, \text{ i.e.,} \\
f(x, u, Du, D^2 u) \geq 0 & \longleftrightarrow & J_x^2 u \in F \quad \forall x \in X. \\
\\
u \text{ is a } \mathbf{supersolution} & \longleftrightarrow & -u \text{ is } \tilde{F} \mathbf{subharmonic}, \text{ i.e.,} \\
f(x, u, Du, D^2 u) \leq 0 & \longleftrightarrow & -J_x^2 u \in \tilde{F} \quad \forall x \in X. \\
\\
u \text{ is a } \mathbf{solution} & \longleftrightarrow & u \text{ is } F \mathbf{harmonic}, \text{ i.e.,} \\
f(x, u, Du, D^2 u) = 0 & \longleftrightarrow & J_x^2 u \in \partial F \quad \forall x \in X \\
& \longleftrightarrow & u \text{ is } F \text{ subharmonic and} \\
& & -u \text{ is } \tilde{F} \text{ subharmonic}
\end{array}$$

These same translations apply to any upper semi-continuous function u by applying them to test functions at each point x .

We also have the following translations between some of the standard structural conditions placed on the function f and conditions on the set F . Let $\mathcal{P} \equiv \{(0, 0, A) : A \geq 0\}$ and $\mathcal{N} \equiv \{(r, 0, 0) : r \leq 0\}$.

$$\begin{array}{lll}
f \text{ is } \mathbf{degenerate elliptic} & \longleftrightarrow & F \text{ satisfies } \mathbf{positivity}, \text{ i.e.,} \\
f(x, r, p, A + P) \geq f(x, r, p, A) \quad \forall P \geq 0 & \longleftrightarrow & F + \mathcal{P} \subset F. \\
\\
f \text{ is } \mathbf{monotone in the dependent variable} & \longleftrightarrow & \\
& F \text{ satisfies } \mathbf{negativity}, \text{ i.e.,} \\
& f(x, r - s, p, A) \geq f(x, r, p, A) \quad \forall s \geq 0 & \longleftrightarrow & F + \mathcal{N} \subset F.
\end{array}$$

$$f \text{ is } \mathbf{proper} \text{ if both conditions hold} \quad \longleftrightarrow \quad F + \mathcal{P} \subset F \text{ and } F + \mathcal{N} \subset F$$

$$f \text{ is } \mathbf{uniformly elliptic} \quad \longleftrightarrow \quad \begin{cases} F + \mathbf{P}_{\lambda, \Lambda} \subset F \text{ for some } 0 < \lambda < \Lambda, \\ \text{or equivalently,} \\ F + \mathbf{P}(\delta) \subset F \text{ for some } \delta > 0. \end{cases}$$

Here $\mathbf{P}_{\lambda, \Lambda}$ is the Pucci cone discussed in §4.5, and $\mathbf{P}(\delta)$ is the cone defined in Example 4.3.3.

It is important to realize that these translations are not precise equivalences (although there is an implication). In passing from the function f to the set $F \equiv \{f \geq 0\}$, the behavior of f away from its zero-set is lost. Matters become simpler, and this can be an advantage (See Comment 3). There are also natural examples where the set $\{f \geq 0\}$ is not really what one wants to take for the set F , and the topological condition required in the “set” point of view easily corrects matters (see Comment 2 below).

Comment 1. As noted above, these translations are not equivalences in general. For example, the positivity condition $F + \mathcal{P} \subset F$ is equivalent to the assumption that

$$f(x, r, p, A) \geq 0 \quad \Rightarrow \quad f(x, r, p, A + P) \geq 0 \quad \forall P \geq 0.$$

which is weaker than the inequality on f required for degenerate ellipticity. The negativity condition $F + \mathcal{N} \subset F$ is equivalent to the assumption that

$$f(x, r, p, A) \geq 0 \quad \Rightarrow \quad f(x, r - s, p, A) \geq 0 \quad \forall s \geq 0.$$

which is weaker than the properness condition placed on f above.

Comment 2. The Topological Condition (T) that $F = \overline{\text{Int}F}$, holds for most classical equations of interest. However, there are cases where it fails, such as the infinite Laplacian $f(p, A) = \langle Ap, p \rangle$ or the k -Laplacian $|p|^2 + (k-2)\langle Ap, p \rangle$, ($1 \leq k \neq 2$). When it fails, it is condition (T) that selects the “correct” subequation F .

Comment 3 (Supersolutions versus \tilde{F} -subharmonicity). There is an important difference between u being a supersolution and $-u$ being \tilde{F} -subharmonic, which arises when $\text{Int}F \neq \{f > 0\}$. However, since we have $\{f > 0\} \subset \text{Int}F$ (equivalently $\sim \text{Int}F \subset \{f \leq 0\}$) we deduce

$$-v \text{ is } \tilde{F} \text{-subharmonic} \quad \Rightarrow \quad v \text{ is an } f \text{ supersolution.} \quad (\text{A.1})$$

The fact that the converse is not true is important. For a constant coefficient, pure second-order subequation $F \subset \text{Sym}^2(\mathbf{R}^n)$, the more restrictive condition on v in (A.1) ensures that comparison holds. That is, with u F -subharmonic and $-v$ \tilde{F} -subharmonic,

$$u \leq v \text{ on } \partial K \quad \Rightarrow \quad u \leq v \text{ on } K$$

(See [HL4] for a proof.) One can show that (A.1) is an equivalence if and only if whenever $F(A) = 0$, the function $F(A + \epsilon I)$ has an isolated zero at $\epsilon = 0$.

Appendix B. Examples of Basic Monotonicity Cones

The following is a list of constant-coefficient convex cone subequations \mathbf{M} such that the Zero Maximum Principle (see §4.1) holds for $\widetilde{\mathbf{M}}$ -subharmonic functions. In cases (1), (5) and (6) the full maximum principle holds, since these equations are independent of the r -variable.

(1) $\mathbf{M} = \mathbf{R} \times \mathbf{R}^n \times \mathcal{P}$. Here the $\widetilde{\mathbf{M}}$ -subharmonic functions are the subaffine functions (see Proposition 2.1.7). This is a monotonicity subequation for any pure second-order subequation $\mathbf{F} = \mathbf{R} \times \mathbf{R}^n \times \mathbf{F}_0$.

(2) $\mathbf{M} = \mathbf{R}_- \times \mathbf{R}^n \times \mathcal{P}$. Here one can characterize the $\widetilde{\mathbf{M}}$ -subharmonics as being “sub” the functions of the form $\max\{0, a(x)\}$ with $a(x)$ affine (the *affine-plus functions*). This is a monotonicity subequation for any gradient-independent subequation.

(3) $\mathbf{M} = \mathbf{R}_- \times \mathcal{D} \times \mathcal{P}$ with $\mathcal{D} \subset \mathbf{R}^n$ a “directional” convex cone with vertex at the origin and non-empty interior.

(4) $\mathbf{M} = \{(r, p, A) \in \mathbf{J}^2 : r \leq -\gamma|p|, p \in \mathcal{D} \text{ and } A \geq 0\}$ with $\gamma > 0$ and $\mathcal{D} \subset \mathbf{R}^n$ as above.

(5) $\mathbf{M} = \mathbf{R} \times \mathbf{M}_0$ with $(p, A) \in \mathbf{M}_0 \iff \langle Ae, e \rangle - \lambda|\langle p, e \rangle| \geq 0 \forall |e| = 1$

For the next example the Maximum Principle only holds for compact sets $K \subset \mathbf{R}^n$ which are contained in a ball of radius R .

(6) $\mathbf{M} = \mathbf{R} \times \mathbf{M}_0$ with $(p, A) \in \mathbf{M}_0 \iff A - \frac{|p|}{R}\text{Id} \geq 0$

The proofs depend on the following elementary result.

THEOREM B.2. *Suppose \mathbf{M} is a constant coefficient convex subequation and $K \subset \mathbf{R}^n$ is compact. If K admits a smooth function ψ which is strictly \mathbf{M} -subharmonic on $\text{Int}K$, then the Zero Maximum Principle holds for the dual subequation $\widetilde{\mathbf{M}}$ on K .*

PROOF. Suppose that the (ZMP) fails for $u \in \text{USC}(K)$. We will show that there exists a point $\bar{x} \in \text{Int}K$ and $\epsilon > 0$ such that $\varphi \equiv -\epsilon\psi$ is a test function for u at \bar{x} . This proves that u is not $\widetilde{\mathbf{M}}$ -subharmonic near \bar{x} because $J_{\bar{x}}^2\psi \in \text{Int}\mathbf{M}$ implies that $J_{\bar{x}}^2\varphi = -\epsilon J_{\bar{x}}^2\psi \notin \widetilde{\mathbf{M}}$.

By assumption, $u \leq 0$ on ∂K but $\sup_K u > 0$. The negativity condition (N) for $\widetilde{\mathbf{M}}$ allows us to subtract a small number from u and assume that $u < 0$ on ∂K with $\sup_K u > 0$. Set $v \equiv u + \epsilon\psi$. Then with $\epsilon > 0$ sufficiently small, $v < 0$ on ∂K but $\sup_K v > 0$. Now let \bar{x} denote a maximum point for v on K . Since $\bar{x} \in \text{Int}K$, this proves that $\varphi \equiv -\epsilon\psi$ is a test function for u at \bar{x} as desired. \square

Proof of (1)–(4). Since the \mathbf{M} in (4) is contained in the other three \mathbf{M} 's, it suffices to find a strictly \mathbf{M} -subharmonic function for \mathbf{M} defined as in (4). Choose $\psi(x) \equiv \frac{1}{2}\delta|x - x_0|^2 - c$ with $\delta, c > 0$. Denote the jet coordinates of ψ at $x \in K$ by $r = \psi(x)$, $p = \delta(x - x_0)$ and $A = \delta I$. Choose $x_0 \in \mathbf{R}^n$ so that $K \subset x_0 + \text{Int}\mathcal{D}$. Then $A \in \text{Int}\mathcal{P}$, $p \in \text{Int}\mathcal{D}$ and $r + \gamma|p| = \frac{1}{2}\epsilon|x - x_0|^2 - c + \gamma\delta|x - x_0| < 0$ if c is large. \square

Proof of (5). Consider $\psi(x) \equiv \frac{1}{N+1}|x|^{N+1}$. Then one computes that

$$p = D\psi = |x|^N \frac{x}{|x|} \quad \text{and} \quad A = D^2\psi = |x|^{N-1} (I + (N-1)P_{[x]})$$

where $P_{[x]}$ is orthogonal projection onto the x -line. Then with $|e| = 1$ we have

$$\frac{1}{|x|^{N-1}}(\langle Ae, e \rangle - \lambda|\langle p, e \rangle|) = 1 - \lambda|x|t + (N-1)t^2 \equiv g(t).$$

with $t \equiv |\langle \frac{x}{|x|}, e \rangle|$. We can assume that $0 \notin K$ and $x \in K$ implies $|x| \leq R$.

The quadratic $g(t)$ has a minimum at $t_0 = \frac{\lambda|x|}{2(N-1)}$ with the minimum value

$g(t_0) = 1 - \frac{\lambda^2|x|^2}{4(N-1)} \geq 1 - \frac{\lambda^2 R^2}{4(N-1)}$. Choose N large enough so that this is > 0 . \square

Proof of (6). This is similar to the proof of (5). It reduces to showing that $g(t) = 1 - \frac{|x|}{R} + (N-1)t^2 > 0$. Now the minimum value (at $t=0$) is $1 - \frac{|x|}{R}$. For the counterexample, consider

$$u(x) \equiv \begin{cases} -(R-|x|)^3 & |x| \leq R \\ 0 & |x| \geq R \end{cases}$$

References

- [A₁] S. Alesker, *Non-commutative linear algebra and plurisubharmonic functions of quaternionic variables*, Bull. Sci. Math., **127** (2003), 1–35. also ArXiv:math.CV/0104209.
- [A₂] ———, *Quaternionic Monge-Ampère equations*, J. Geom. Anal., **13** (2003), 205–238. ArXiv:math.CV/0208805.
- [AV] S. Alesker and M. Verbitsky, *Plurisubharmonic functions on hypercomplex manifolds and HKT-geometry*, arXiv: math.CV/0510140 Oct.2005
- [Al] A. D. Alexandrov, *The Dirichlet problem for the equation $\text{Det}\|z_{i,j}\| = \psi(z_1, \dots, z_n, x_1, \dots, x_n)$* , I. Vestnik, Leningrad Univ. **13** No. 1, (1958), 5–24.
- [AGV] M. E. Amendola, G. Galise and A. Vitolo, *Riesz capacity, maximum principle and removable sets of fully nonlinear second-order elliptic operators*, Preprint, University of Salerno.
- [AFS] D. Azagra, J. Ferrera and B. Sanz, *Viscosity solutions to second order partial differential equations on riemannian manifolds*, ArXiv:math.AP/0612742v2, Feb. 2007.
- [BT₁] E. Bedford and B. A. Taylor, The Dirichlet problem for a complex Monge-Ampère equation, Inventiones Math. **37** (1976), no.1, 1–44.
- [BT₂] ———, Variational properties of the complex Monge-Ampère equation, I. Dirichlet principle, Duke Math. J. **45** (1978), no. 2, 375–403.
- [BT₃] ———, A new capacity for plurisubharmonic functions, Acta Math. **149** (1982), no.1–2, 1–40.
- [B] H. J. Bremermann, *On a generalized Dirichlet problem for plurisubharmonic functions and pseudo-convex domains*, Trans. A. M. S. **91** (1959), 246–276.
- [BH] R. Bryant and F. R. Harvey, *Submanifolds in hyper-Kähler geometry*, J. Amer. Math. Soc. **1** (1989), 1–31.
- [CKNS] L. Caffarelli, J. J. Kohn, L. Nirenberg, and J. Spruck, *The Dirichlet problem for non-linear second order elliptic equations II: Complex Monge-Ampère and uniformly elliptic equations*, Comm. on Pure and Applied Math. **38** (1985), 209–252.
- [CLN] L. Caffarelli, Y.Y. Li and L. Nirenberg *Some remarks on singular solutions of nonlinear elliptic equations, III: viscosity solutions, including parabolic operators*. ArXiv:1101.2833.
- [CNS₁] L. Caffarelli, L. Nirenberg and J. Spruck, *The Dirichlet problem for nonlinear second order elliptic equations. I: Monge-Ampère equation*, Comm. Pure Appl. Math. **37** (1984), 369–402.
- [CNS₂] ———, *The Dirichlet problem for nonlinear second order elliptic equations, III: Functions of the eigenvalues of the Hessian*, Acta Math. **155** (1985), 261–301.
- [CNS₃] ———, *The Dirichlet problem for the degenerate Monge-Ampère equation*, Rev. Mat. Iberoamericana **2** (1986), 19–27.

- [CNS₄] ———, *Correction to: “The Dirichlet problem for nonlinear second order elliptic equations. I: Monge-Ampère equation”*, Comm. Pure Appl. Math. **40** (1987), 659–662.
- [CCH] A. Chau, J. Chen and W. He, *Lagrangian mean curvature flow for entire Lipschitz graphs*, ArXiv:0902.3300 Feb, 2009.
- [CGG₁] Y.-G. Chen, Y. Giga and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, Proc. Japan Acad. Ser. A. Math. Sci **65** (1989), 207–210.
- [CGG₂] ———, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Diff. Geom. **33** (1991), 749–789.
- [CY₁] S.-Y. Cheng and S.-T. Yau, *On the regularity of the Monge-Ampère equation $\det(\partial^2 u / \partial x_i \partial x_j) = F(x, u)$* , Comm. Pure Appl. Math. **30** (1977), no. 1, 41–68.
- [CY₂] ———, *The real Monge-Ampère equation and affine flat structures*, Proceedings of the 1980 Beijing Symposium on Differential Geometry and Differential Equations, Vol. 1, 2, 3 (Beijing, 1980), 339–370, Science Press, Beijing, 1982.
- [C] M. G. Crandall, *Viscosity solutions: a primer*, pp. 1–43 in “Viscosity Solutions and Applications” Ed.’s Dolcetta and Lions, SLMN **1660**, Springer Press, New York, 1997.
- [CIL] M. G. Crandall, H. Ishii and P. L. Lions *User’s guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N. S.) **27** (1992), 1–67.
- [EGZ] P. Eyssidieux, V. Guedj, and A. Zeriahi, *Viscosity solutions to degenerate Monge-Ampère equations*, ArXiv:1007.0076.
- [E] L. C. Evans, *Regularity for fully nonlinear elliptic equations and motion by mean curvature*, pp. 98–133 in “Viscosity Solutions and Applications” Ed.’s Dolcetta and Lions, SLMN **1660**, Springer Press, New York, 1997.
- [ESm] L. C. Evans and C. K. Smart, *Everywhere differentiability of infinite harmonic functions*, Berkeley preprint, 2012
- [ES₁] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature, I*, J. Diff. Geom. **33** (1991), 635–681.
- [ES₂] ———, *Motion of level sets by mean curvature, II*, Trans. A. M. S. **330** (1992), 321–332.
- [ES₃] ———, *Motion of level sets by mean curvature, III*, J. Geom. Anal. **2** (1992), 121–150.
- [ES₄] ———, *Motion of level sets by mean curvature, IV*, J. Geom. Anal. **5** (1995), 77–114.
- [G] P. Garabedian, *Partial Differential Equations*, J. Wiley and Sons, New York, 1964.
- [G] L. Gårding, *An inequality for hyperbolic polynomials*, J. Math. Mech. **8** no. 2 (1959), 957–965.
- [Gi] Y. Giga, *Surface Evolution Equations—A level set approach*, Birkhäuser, 2006.
- [Giu] E. Giusti, *Minimal surfaces and functions of bounded variation*, Monographs in Mathematics, 80. Birkhäuser Verlag, Basel, 1984.
- [Gr] M. Gromov, *Pseudoholomorphic curves in symplectic manifolds*, Invent. Math. **82** (1985), no. 2, 307–347.
- [H] F. R. Harvey, *Removable singularities and structure theorems for positive currents*. Partial differential equations (Proc. Sympos. Pure Math., Vol. XXIII, Univ. California, Berkeley, Calif., 1971), pp. 129–133. Amer. Math. Soc., Providence, R.I., 1973.
- [HL₁] F. R. Harvey and H. B. Lawson, Jr, *Calibrated geometries*, Acta Mathematica **148** (1982), 47–157.
- [HL₂] ———, *An introduction to potential theory in calibrated geometry*, Amer. J. Math. **131** no. 4 (2009), 893–944. ArXiv:math.0710.3920.

- [HL₃] ———, *Duality of positive currents and plurisubharmonic functions in calibrated geometry*, Amer. J. Math. **131** no. 5 (2009), 1211–1240. ArXiv:math.0710.3921.
- [HL₄] ———, *Dirichlet duality and the non-linear Dirichlet problem*, Comm. on Pure and Applied Math. **62** (2009), 396–443. ArXiv:math.0710.3991
- [HL₅] ———, *Plurisubharmonicity in a general geometric context*, Geometry and Analysis **1** (2010), 363–401. ArXiv:0804.1316.
- [HL₆] ———, *Dirichlet duality and the nonlinear Dirichlet problem on Riemannian manifolds*, J. Diff. Geom. **88** (2011), 395–482. ArXiv:0912.5220.
- [HL₇] ———, *Hyperbolic polynomials and the Dirichlet problem*, ArXiv:0912.5220.
- [HL₈] ———, *Gårding’s theory of hyperbolic polynomials*, to appear in *Communications in Pure and Applied Mathematics*.
- [HL₉] ———, *The restriction theorem for fully nonlinear subequations*, Ann. Inst. Fourier (to appear). ArXiv:1101.4850.
- [HL₁₀] ———, *Potential Theory on almost complex manifolds*, Ann. Inst. Fourier (to appear). ArXiv:1107.2584.
- [HL₁₁] ———, *Foundations of p -convexity and p -plurisubharmonicity in riemannian geometry*, ArXiv: 1111.3895.
- [HL₁₂] ———, *Removable singularities for nonlinear subequations*. (Stony Brook Preprint).
- [HL₁₃] ———, *The equivalence of viscosity and distributional subsolutions for convex subequations—the strong Bellman principle*, Bol. Soc. Bras. de Mat. (to appear). ArXiv:1301.4914.
- [HL₁₄] ———, *Lagrangian plurisubharmonicity and convexity*, Stony Brook Preprint.
- [HL₁₅] ———, *Radial subequations, isolated singularities and tangent functions*, Stony Brook Preprint.
- [HP₁] F. R. Harvey and J. Polking, *Removable singularities of solutions of linear partial differential equations*, Acta Math. **125** (1970), 39–56.
- [HP₂] ———, *Extending analytic objects*, Comm. Pure Appl. Math. **28** (1975), 701–727.
- [HH] M. Hervé and R.M. Hervé, *Les fonctions surharmoniques dans l’axiomatique de M. Brelot associées à un opérateur elliptique dégénéré*, Annals de l’institut Fourier, **22**, no. 2 (1972), 131–145.
- [Ho₁] L. Hörmander, An introduction to complex analysis in several variables, Third edition. North-Holland Mathematical Library, 7. North-Holland Publishing Co., Amsterdam, 1990.
- [Ho₂] ———, The analysis of linear partial differential operators. III. Pseudodifferential operators. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], 274. Springer-Verlag, Berlin, 1985.
- [HM] L. R. Hunt and J. J. Murray, *q -plurisubharmonic functions and a generalized Dirichlet problem*, Michigan Math. J., **25** (1978), 299–316.
- [I] H. Ishii, *On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic pde’s*, Comm. Pure and App. Math. **42** (1989), 14–45.
- [IL] H. Ishii and P. L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Diff. Eq. **83** (1990), 26–78.
- [J₁] R. Jensen, *Uniqueness criteria for viscosity solutions of fully nonlinear elliptic partial differential equations*, Indiana Univ. Math. J. **38** (1989), 629–667.
- [J₂] ———, *Uniqueness of Lipschitz extensions: minimizing the sup norm of the gradient*, Arch. Rational Mech. Analysis 123 (1993), 51–74.
- [JLM] P. Juutinen, P. Lindqvist, and J. Manfredi, *On the equivalence of viscosity solutions and weak solutions for a quasi-linear equation*, SIAM J. Math. Anal. 33 (2001), no. 3, 699–717.
- [K] N. V. Krylov, *On the general notion of fully nonlinear second-order elliptic equations*, Trans. Amer. Math. Soc. (3) **347** (1979), 30–34.

- [La₁] D.Labutin, *Isolated singularities for fully nonlinear elliptic equations*, J. Differential Equations **177** (2001), No. 1, 49–76.
- [La₂] D.Labutin, *Singularities of viscosity solutions of fully nonlinear elliptic equations*, Viscosity Solutions of Differential Equations and Related Topics, Ishii ed., RIMS Kôkyûroku No. 1287, Kyoto University, Kyoto (2002), 45–57
- [La₃] ———, *Potential estimates for a class of fully nonlinear elliptic equations*, Duke Math. J. **111** No. 1 (2002), 1–49.
- [L] N. S. Landkof, Foundations of Modern Potential Theory, Springer-Verlag, New York, 1972.
- [Le] P. Lelong, Fonctions plurisousharmoniques et formes différentielles positives, Gordon and Breach, Paris-London-New York (Distributed by Dunod Éditeur, Paris) 1968.
- [LE] Y. Luo and A. Eberhard, An application of $C^{1,1}$ approximation to comparison principles for viscosity solutions of curvatures equations, Nonlinear Analysis **64** (2006), 1236–1254.
- [NTV] N. Nadirashvili, V. Tkachev and S. Vlăduță *Non-classical Solution to Hessian Equation from Cartan Isoparametric Cubic*, ArXiv:1111. 0329.
- [NW] A. Nijenhuis and W. Woolf, *Some integration problems in almost -complex and complex manifolds*, Ann. of Math., **77** (1963), 424–489.
- [O] A. Oberman, *The convex envelope is the solution of a nonlinear obstacle problem*, Proc. A.M.S. **135** (2007), no. 6, 1689–1694.
- [OS] A. Oberman and L. Silvestre, *The Dirichlet problem for the convex envelope*, Trans. A.M.S. **363** (2011), no. 11, 5871–5886.
- [P] N. Pali, *Fonctions plurisousharmoniques et courants positifs de type (1,1) sur une variété presque complexe*, Manuscripta Math. **118** (2005), no. 3, 311–337.
- [PZ] S. Peng and D. Zhou, *Maximum principle for viscosity solutions on riemannian manifolds*, ArXiv:0806.4768, June 2008.
- [Pl] S. Pliś, *The Monge-Ampère equation on almost complex manifolds*, ArXiv:1106.3356, June, 2011.
- [Po₁] A. V. Pogorelov, *On the regularity of generalized solutions of the equation $\det(\partial^2 u / \partial x_i \partial x_j) = \phi(x_1, \dots, x_n) > 0$* , Dokl. Akad. Nauk SSSR 200, 1971, pp. 534–537.
- [Po₂] ———, *The Dirichlet problem for the n-dimensional analogue of the Monge-Ampère equation*, Dokl. Akad. Nauk SSSR 201, 1971, pp. 790–793.
- [RT] J. B. Rauch and B. A. Taylor, *The Dirichlet problem for the multidimensional Monge-Ampère equation*, Rocky Mountain J. Math **7** (1977), 345–364.
- [Sh₁] J.-P. Sha, *p -convex riemannian manifolds*, Invent. Math. **83** (1986), 437–447.
- [Sh₂] ———, *Handlebodies and p -convexity*, J. Diff. Geom. **25** (1987), 353–361.
- [Shi] B. Shiffman, *Extension of positive line bundles and meromorphic maps.*, Invent. Math. **15** (1972), no. 4, 332–347.
- [S₁] Z. Slodkowski, *The Bremermann-Dirichlet problem for q -plurisubharmonic functions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **11** (1984), 303–326.
- [S₂] ———, *Pseudoconvex classes of functions. I. Pseudoconcave and pseudoconvex sets*, Pacific J. of Math. **134** no. 2 (1988), 343–376.
- [S₃] ———, *Pseudoconvex classes of functions. II. Affine pseudoconvex classes on \mathbf{R}^N* , Pacific J. of Math. **141** no. 1 (1990), 125–163.
- [S₄] ———, *Pseudoconvex classes of functions. III. Characterization of dual pseudoconvex classes on complex homogeneous spaces*, Trans. A. M. S. **309** no.1 (1988), 165–189.
- [S₅] ———, *Complex interpolation of normed and quasinormed spaces in several dimensions*, I. Trans. Amer. Math. Soc. 308 (1988), no. 2, 685–711.
- [SYZ₁] A. Strominger, S.-T. Yau and E. Zaslow, *Mirror symmetry is T-duality*, Winter School on Mirror Symmetry, Vector Bundles and Lagrangian Submanifolds

- (Cambridge, MA, 1999), 333–347, AMS/IP Stud. Adv. Math., 23, Amer. Math. Soc., Providence, RI, 2001.
- [SYZ₂] ———, *Mirror symmetry is T-duality*, Nuclear Phys. B 479 (1996), no. 1–2, 243–259.
- [So] P. Soravia, *On nonlinear convolution and uniqueness of viscosity solutions*, Analysis **20** (2000), 373–386.
- [TU] N. S. Trudinger and J.n I. E. Urbas, *Second derivative estimates for equations of Monge-Ampère type*, Bull. Austral. Math. Soc. **30** (1984), 321–334.
- [W] J. B. Walsh, *Continuity of envelopes of plurisubharmonic functions*, J. Math. Mech. **18** (1968–69), 143–148.
- [WY] D. Wang and Y. Yuan, *Hessian estimates for special Lagrangian equation with critical and supercritical phases in general dimensions*, ArXiv:1110.1417.
- [Wu] H. Wu, *Manifolds of partially positive curvature*, Indiana Univ. Math. J. **36** No. 3 (1987), 525–548.
- [Yau] S.-T. Yau, *On the Ricci curvature of a compact Kähler manifold and the complex Monge-Ampère equation. I*, Comm. Pure Appl. Math. 31 (1978), no. 3, 339–411.
- [Y] Yu Yuan, *A priori estimates for solutions of fully nonlinear special lagrangian equations*, Ann Inst. Henri Poincaré **18** (2001), 261–270.

MATHEMATICS DEPARTMENT, RICE UNIVERSITY, HOUSTON, TX 77005-1982, USA
E-mail address: harvey@rice.edu

MATHEMATICS DEPARTMENT, STONY BROOK UNIVERSITY, STONY BROOK, NY 11790-3651, USA
E-mail address: blaine@math.sunysb.edu

Links of complex analytic singularities

János Kollár

Let X be a complex algebraic or analytic variety. Its local topology near a point $x \in X$ is completely described by its *link* $L(x \in X)$, which is obtained as the intersection of X with a sphere of radius $0 < \epsilon \ll 1$ centered at x . The intersection of X with the closed ball of radius ϵ centered at x is homeomorphic to the cone over $L(x \in X)$; cf. [GM88, p.41].

If $x \in X$ is a smooth point then its link is a sphere of dimension $2 \dim_{\mathbb{C}} X - 1$. Conversely, if X is a normal surface and $L(x \in X)$ is a sphere then x is a smooth point [Mum61], but this fails in higher dimensions [Bri66].

The aim of this survey is to study in some sense the opposite question: we are interested in the “most complicated” links. In its general form, the question is the following.

PROBLEM 1. Which topological spaces can be links of complex algebraic or analytic singularities?

If $\dim X = 1$, then the possible links are disjoint unions of circles. The answer is much more complicated in higher dimensions and we focus on isolated singularities from now on, though many results hold for non-isolated singularities as well. Thus the link $L(x \in X)$ is a (differentiable) manifold of (real) dimension $2 \dim_{\mathbb{C}} X - 1$.

Among the simplest singularities are the cones over smooth projective varieties. Let $Z \subset \mathbb{P}^N$ be a smooth projective variety and $X := \text{Cone}(Z) \subset \mathbb{C}^{N+1}$ the cone over Z with vertex at the origin. Then $L(0 \in X)$ is a circle bundle over Z whose first Chern class is the hyperplane class. Thus the link of the vertex of $\text{Cone}(Z)$ is completely described by the base Z and by the hyperplane class $[H] \in H^2(Z, \mathbb{Z})$.

Note that a singularity $0 \in X \subset \mathbb{C}^N$ is a cone iff it can be defined by homogeneous equations. One gets a much larger class of singularities if we consider homogeneous equations where different variables have different degree (or weight).

For a long time it was believed that links of isolated singularities are “very similar” to links of cones and weighted cones. The best illustration of

this is given by the complete description of links of surface singularities given in [Neu81]. Cones give circle bundles over Riemann surfaces and weighted cones give Seifert bundles over Riemann surfaces. General links are more complicated but they are all obtained by gluing Seifert bundles over Riemann surfaces with boundary. These are definitely more complicated than Seifert bundles, but much simpler than general 3–manifolds. In particular, hyperbolic 3–manifolds – which comprise the largest and most complicated class – do not occur as links.

Important examples of the similarity of general links to smooth projective varieties are given by the local Lefschetz theorems, initiated by Grothendieck [Gro68] and developed much further subsequently; see [GM88] for a detailed treatment.

As another illustration, the weights of the mixed Hodge structure on the cohomology groups of links also follow the same pattern for general links as for links of cones, see [DH88] or [PS08, Sec.6.3].

These and many other examples led to a viewpoint that was best summarized in [GM88, p.26]: “*Philosophically, any statement about the projective variety or its embedding really comes from a statement about the singularity at the point of the cone. Theorems about projective varieties should be consequences of more general theorems about singularities which are no longer required to be conical.*”

Recently this belief was called into question by [KK11] which proved that fundamental groups of general links are very different from fundamental groups of links of cones. The aim of this paper is to summarize the results, present several new theorems and review the problems that arise.

Philosophically, the main long term question is to understand the limits of the above principle. We know that it fails for the fundamental group but it seems to apply to cohomology groups. It is unclear if it applies to simply connected links or not.

The new results rely on a method, considered in [Kol11], to construct singularities using their resolution. By Hironaka’s resolution theorem, for every isolated singularity ($x \in X$) there is a proper, birational morphism $f : Y \rightarrow X$ such that $E := f^{-1}(x)$ is a simple normal crossing divisor and $Y \setminus E \rightarrow X \setminus \{x\}$ is an isomorphism. The method essentially reverses the resolution process. That is, we start with a (usually reducible) simple normal crossing variety E , embed E into a smooth variety Y and then contract $E \subset Y$ to a point to obtain ($x \in X$). If E is smooth, this is essentially the cone construction.

This approach has been one of the standard ways to construct surface singularities but it has not been investigated in higher dimensions until recently. There were probably two reason for this. First, if $\dim X \geq 3$ then there is no “optimal” choice for the resolution $f : Y \rightarrow X$. Thus the exceptional set $E = f^{-1}(x)$ depends on many arbitrary choices and it is not easy to extract any invariant of the singularity from E ; see, however, Definition 6. Thus any construction starting with E seemed rather arbitrary.

Second, the above philosophy suggested that one should not get anything substantially new this way.

The first indication that this method is worth exploring was given in [Kol11] where it was used to construct new examples of terminal and log canonical singularities that contradicted earlier expectations.

A much more significant application was given in [KK11]. Since in higher dimensions a full answer to Problem 1 may well be impossible to give, it is sensible to focus on some special aspects. A very interesting question turned out to be the following.

PROBLEM 2. Which groups occur as fundamental groups of links of complex algebraic or analytic singularities?

Note that the fundamental groups of smooth projective varieties are rather special; see [ABC⁺96] for a survey. Even the fundamental groups of smooth quasi projective varieties are quite restricted [Mor78, KM98a, CS08, DPS09]. By contrast fundamental groups of links are arbitrary.

THEOREM 3. [KK11] *For every finitely presented group G there is an isolated, complex singularity $(0 \in X_G)$ with link L_G such that $\pi_1(L_G) \cong G$.*

Note that once such a singularity exists, a local Lefschetz–type theorem (cf. [GM88, Sec.II.1.2]) implies that the link of a general 3-dimensional hyperplane section has the same fundamental group.

There are two natural directions to further develop this result: one can connect properties of the fundamental group of a link to algebraic or analytic properties of a singularity and one can investigate further the topology of the links or of the resolutions.

In the first direction, the following result answers a question of Wahl.

THEOREM 4. *For a finitely presented group G the following are equivalent.*

- (1) *G is \mathbb{Q} -perfect, that is, its largest abelian quotient is finite.*
- (2) *G is the fundamental group of the link of an isolated Cohen–Macaulay singularity (46) of dimension ≥ 3 .*

One can study the local topology of X by choosing a resolution of singularities $\pi : Y \rightarrow X$ such that $E_x := \pi^{-1}(x) \subset Y$ is a simple normal crossing divisor and then relating the topology of E_x to the topology of the link $L(x \in X)$.

The topology of a simple normal crossing divisor E can in turn be understood in 2 steps. First, the E_i are smooth projective varieties, and their topology is much studied. A second layer of complexity comes from how the components E_i are glued together. This gluing process can be naturally encoded by a finite cell complex $\mathcal{D}(E)$, called the *dual complex* or *dual graph* of E .

DEFINITION 5 (Dual complex). Let E be a variety with irreducible components $\{E_i : i \in I\}$. We say that E is a *simple normal crossing* variety (abbreviated as *snc*) if the E_i are smooth and every point $p \in E$ has an open (Euclidean) neighborhood $p \in U_p \subset E$ and an embedding $U_p \hookrightarrow \mathbb{C}^{n+1}$ such that the image of U_p is an open subset of the union of coordinate hyperplanes ($z_1 \cdots z_{n+1} = 0$). A *stratum* of E is any irreducible component of an intersection $\cap_{i \in J} E_i$ for some $J \subset I$.

The combinatorics of E is encoded by a cell complex $\mathcal{D}(E)$ whose vertices are labeled by the irreducible components of E and for every stratum $W \subset \cap_{i \in J} E_i$ we attach a $(|J| - 1)$ -dimensional cell. Note that for any $j \in J$ there is a unique irreducible component of $\cap_{i \in J \setminus \{j\}} E_i$ that contains W ; this specifies the attaching map. $\mathcal{D}(E)$ is called the *dual complex* or *dual graph* of E . (Although $\mathcal{D}(E)$ is not a simplicial complex in general, it is an unordered Δ -complex in the terminology of [Hat02, p.534].)

DEFINITION 6 (Dual complexes associated to a singularity). Let X be a normal variety and $x \in X$ a point. Choose a resolution of singularities $\pi : Y \rightarrow X$ such that $E_x := \pi^{-1}(x) \subset Y$ is a simple normal crossing divisor. Thus it has a dual complex $\mathcal{D}(E_x)$.

The dual graph of a normal surface singularity has a long history. Higher dimensional versions appear in [Kul77, Per77, Gor80, FM83] but systematic investigations were started only recently; see [Thu07, Ste08, Pay09, Pay11].

It is proved in [Thu07, Ste08, ABW11] that the homotopy type of $\mathcal{D}(E_x)$ is independent of the resolution $Y \rightarrow X$. We denote it by $\mathcal{DR}(x \in X)$.

The proof of Theorem 3 gives singularities for which the fundamental group of the link is isomorphic to the fundamental group of $\mathcal{DR}(x \in X)$. In general, it seems easier to study $\mathcal{DR}(x \in X)$ than the link and the next theorem shows that not just the fundamental group but the whole homotopy type of $\mathcal{DR}(0 \in X)$ can be arbitrary. The additional properties (7.2–3) follow from the construction as in [Kol11, KK11].

THEOREM 7. *Let T be a connected, finite cell complex. Then there is a normal singularity ($0 \in X$) such that*

- (1) *the complex $\mathcal{DR}(0 \in X)$ is homotopy equivalent to T ,*
- (2) *$\pi_1(L(0 \in X)) \cong \pi_1(T)$ and*
- (3) *if $\pi : Y \rightarrow X$ is any resolution then $R^i \pi_* \mathcal{O}_Y \cong H^i(T, \mathbb{C})$ for $i > 0$.*

The fundamental groups of the dual complexes of rational singularities (52) were determined in [KK11, Thm.42]. The next result extends this by determining the possible homotopy types of $\mathcal{DR}(0 \in X)$.

THEOREM 8. *Let T be a connected, finite cell complex. Then there is a rational singularity ($0 \in X$) whose dual complex $\mathcal{DR}(0 \in X)$ is homotopy equivalent to T iff T is \mathbb{Q} -acyclic, that is, $H^i(T, \mathbb{Q}) = 0$ for $i > 0$.*

As noted in [Pay11], the dual complex $\mathcal{DR}(0 \in X)$ can be defined even up-to simple-homotopy equivalence [Coh73]. The proofs given in

[KK11] use Theorem 25, which in turn relies on some general theorems of [Cai61, Hir62] that do not seem to give simple-homotopy equivalence.¹

Content of the Sections.

Cones, weighted cones and the topology of the corresponding links are discussed in Section 1.

The plan for the construction of singularities from their resolutions is outlined in Section 2 and the rest of the paper essentially fleshes out the details.

In Section 3 we show that every finite cell complex is homotopy equivalent to a Voronoi complex. These Voronoi complexes are then used to construct simple normal crossing varieties in Section 4.

The corresponding singularities are constructed in Section 5 where we prove Theorem 7 except for an explicit resolution of the resulting singularities which is accomplished in Section 6.

The proof of Theorem 4 is given in Section 7 where several other equivalent conditions are also treated. Theorem 8 on rational singularities is reviewed in Section 8.

Open questions and problems are discussed in Section 9.

Acknowledgments. I thank I. Dolgachev, T. de Fernex, T. Jarvis, M. Kapovich, L. Maxim, A. Némethi, P. Ozsváth, S. Payne, P. Popescu-Pampu, M. Ramachandran, J. Shaneson, T. Szamuely, D. Toledo, B. Totaro, J. Wahl, and C. Xu for comments and corrections. Partial financial support was provided by the NSF under grant number DMS-07-58275 and by the Simons Foundation. Part of this paper was written while the author visited the University of Utah.

1. Weighted homogeneous links

DEFINITION 9 (Weighted homogeneous singularities). Assign positive weights to the variables $w(x_i) \in \mathbb{Z}$, then the weight of a monomial $\prod_i x_i^{a_i}$ is

$$w(\prod_i x_i^{a_i}) := \sum_i a_i w(x_i).$$

A polynomial f is called *weighted homogeneous* of weighted-degree $w(f)$ iff every monomial that occurs in f with nonzero coefficient has weight $w(f)$.

Fix weights $\mathbf{w} := (w(x_1), \dots, w(x_N))$ and let $\{f_i : i \in I\}$ be weighted homogeneous polynomials. They define both a projective variety in a weighted projective space

$$Z(f_i : i \in I) \subset \mathbb{P}(\mathbf{w})$$

and an affine *weighted cone*

$$C(f_i : i \in I) \subset \mathbb{C}^N.$$

Somewhat loosely speaking, a singularity is called *weighted homogeneous* if it is isomorphic to a singularity defined by a weighted cone for some weights

¹This problem is settled in [Kol13a].

$w(x_i)$. (In the literature these are frequently called *quasi-homogeneous* singularities.)

In many cases the weights are uniquely determined by the singularity (up to rescaling) but not always. For instance, the singularity $(xy = z^n)$ is weighted homogeneous for any weights that satisfy $w(x) + w(y) = n \cdot w(z)$.

If $C \subset \mathbb{C}^N$ is a weighted cone then it has a \mathbb{C}^* -action given by

$$(x_1, \dots, x_N) \mapsto (t^{m_1}x_1, \dots, t^{m_N}x_N) \quad \text{where } m_i = \frac{1}{w(x_i)} \prod_j w(x_j).$$

Conversely, let X be a variety with a \mathbb{C}^* action and $x \in X$ a fixed point that is attractive as $t \rightarrow 0$. Linearizing the action shows that $x \in X$ is a weighted homogeneous singularity.

10 (Links of weighted homogeneous singularities). The \mathbb{C}^* -action on a weighted homogeneous singularity ($x \in X$) induces a fixed point free \mathbb{S}^1 -action on its link L . If we think of X as a weighted cone over the corresponding projective variety $Z \subset \mathbb{P}(\mathbf{w})$ then we get a projection $\pi : L \rightarrow Z$ whose fibers are exactly the orbits of the \mathbb{S}^1 -action, that is, the link of a weighted homogeneous singularity has a *Seifert bundle* structure. (For our purposes we can think that a Seifert bundle is the same as a fixed point free \mathbb{S}^1 -action.) If $(x \in X)$ is an isolated singularity then Z is an orbifold.

It is thus natural to study the topology of links of weighted homogeneous singularities in two steps.

- (1) Describe all $2n - 1$ -manifolds with a fixed point free \mathbb{S}^1 -action.
- (2) Describe which among them occur as links of weighted homogeneous singularities.

11 (Homology of a weighted homogeneous link). [OW75] Let $\pi : L \rightarrow Z$ be the Seifert bundle structure. The cohomology of L is computed by a spectral sequence

$$H^i(Z, R^j \pi_* \mathbb{Q}_L) \Rightarrow H^{i+j}(L, \mathbb{Q}). \quad (11.1)$$

All the fibers are oriented circles, thus $R^0 \pi_* \mathbb{Q}_L \cong R^1 \pi_* \mathbb{Q}_L \cong \mathbb{Q}_Z$ and $R^j \pi_* \mathbb{Q}_L = 0$ for $j > 1$. Thus the E_2 -term of the spectral sequence is

$$\begin{array}{ccccccc} H^0(Z, \mathbb{Q}) & & H^1(Z, \mathbb{Q}) & & H^2(Z, \mathbb{Q}) & & \cdots \\ & \searrow & & \searrow & & & \\ H^0(Z, \mathbb{Q}) & & H^1(Z, \mathbb{Q}) & & H^2(Z, \mathbb{Q}) & & \cdots \end{array} \quad (11.2)$$

where the differentials are cup product with the (weighted) hyperplane class

$$c_1(\mathcal{O}_Z(1)) \cup : H^i(Z, R^1 \pi_* \mathbb{Q}_L) \cong H^i(Z, \mathbb{Q}) \rightarrow H^{i+2}(Z, \mathbb{Q}). \quad (11.3)$$

Since Z is an orbifold, these are injective if $i + 2 \leq \dim Z$ and surjective if $i \geq \dim Z$. Thus we conclude that

$$\begin{aligned} h^i(L, \mathbb{Q}) &= h^i(Z, \mathbb{Q}) - h^{i-2}(Z, \mathbb{Q}) & \text{if } i \leq \dim Z \text{ and} \\ h^{i+1}(L, \mathbb{Q}) &= h^i(Z, \mathbb{Q}) - h^{i+2}(Z, \mathbb{Q}) & \text{if } i \geq \dim Z \end{aligned} \quad (11.4)$$

where we set $h^i(Z, \mathbb{Q}) = 0$ for $i < 0$ or $i > 2\dim Z$. In particular we see that L is a rational homology sphere iff Z is a rational homology complex projective space.

By contrast, the spectral sequence computing the integral cohomology of L is much more complicated. We have a natural injection $R^1\pi_*\mathbb{Z}_L \hookrightarrow \mathbb{Z}_Z$ which is, however, rarely an isomorphism. The computations were carried out only for $\dim L \leq 5$ [Kol05].

12 (Weighted homogeneous surface singularities). This is the only case that is fully understood.

The classification of fixed point free circle actions on 3-manifolds was considered by Seifert [Sei32]. If M is a 3-manifold with a fixed point free circle action then the quotient space $F := M/\mathbb{S}^1$ is a surface (without boundary in the orientable case). The classification of these *Seifert fibered* 3-manifolds $f : M \rightarrow F$ is thus equivalent to the classification of fixed point free circle actions. It should be noted that already in this classical case, it is conceptually better to view the base surface F not as a 2-manifold but as a 2-dimensional *orbifold*, see [Sco83] for a detailed survey from this point of view.

Descriptions of weighted homogeneous surface singularities are given in [Pin77, Dol83, Dem88, FZ03].

Weighted homogeneous 3-fold singularities.

There is a quite clear picture about the simply connected case since simply connected 5-manifolds are determined by their homology.

By a theorem of [Sma62, Bar65], a simply connected, compact 5-manifold L is uniquely determined by $H_2(L, \mathbb{Z})$ and the second Stiefel–Whitney class, which we view as a map $w_2 : H_2(L, \mathbb{Z}) \rightarrow \mathbb{Z}/2$. Furthermore, there is such a 5-manifold iff there is an integer $k \geq 0$ and a finite Abelian group A such that either $H_2(L, \mathbb{Z}) \cong \mathbb{Z}^k + A + A$ and $w_2 : H_2(L, \mathbb{Z}) \rightarrow \mathbb{Z}/2$ is arbitrary, or $H_2(L, \mathbb{Z}) \cong \mathbb{Z}^k + A + A + \mathbb{Z}/2$ and w_2 is projection on the $\mathbb{Z}/2$ -summand.

The existence of Seifert bundles on simply connected compact 5-manifolds was treated in [Kol06]. The answer mostly depends on the torsion subgroup of $H_2(L, \mathbb{Z})$, but there is a subtle interplay with w_2 .

DEFINITION 13. Let M be any manifold. Write its second homology as a direct sum of cyclic groups of prime power order

$$H_2(M, \mathbb{Z}) = \mathbb{Z}^k + \sum_{p,i} (\mathbb{Z}/p^i\mathbb{Z})^{c(p^i)} \quad (13.1)$$

for some $k = \dim H_2(M, \mathbb{Q})$ and $c(p^i) = c(p^i, M)$. The numbers $k, c(p^i)$ are determined by $H_2(M, \mathbb{Z})$ but the subgroups $(\mathbb{Z}/p^i\mathbb{Z})^{c(p^i)} \subset H_2(M, \mathbb{Z})$ are usually not unique. One can choose the decomposition (13.1) such that $w_2 : H_2(M, \mathbb{Z}) \rightarrow \mathbb{Z}/2$ is zero on all but one summand $\mathbb{Z}/2^n$. This value n is unique and it is denoted by $i(M)$ [Bar65]. This invariant can take up any value n for which $c(2^n) \neq 0$, besides 0 and ∞ . Alternatively, $i(M)$ is the

smallest n such that there is an $\alpha \in H_2(M, \mathbb{Z})$ such that $w_2(\alpha) \neq 0$ and α has order 2^n .

The existence of a fixed point free differentiable circle action puts strong restrictions on H_2 and on w_2 .

THEOREM 14. [Kol06, Thm.3] *Let L be a compact, simply connected 5-manifold. Then L admits a fixed point free differentiable circle action if and only if $H_2(L, \mathbb{Z})$ and w_2 satisfy the following conditions.*

- (1) *For every p , we have at most $\dim H_2(M, \mathbb{Q}) + 1$ nonzero $c(p^i)$ in (13.1).*
- (2) *One can arrange that $w_2 : H_2(L, \mathbb{Z}) \rightarrow \mathbb{Z}/2$ is the zero map on all but the $\mathbb{Z}^k + (\mathbb{Z}/2)^{c(2)}$ summands in (13.1). That is, $i(L) \in \{0, 1, \infty\}$.*
- (3) *If $i(L) = \infty$ then $\#\{i : c(2^i) > 0\} \leq \dim H_2(M, \mathbb{Q})$.*

REMARK 15. Note that while Theorem 14 tells us which compact, simply connected 5-manifolds admit a fixed point free differentiable circle action, the proof does not classify all circle actions. In particular, the classification of all circle actions on \mathbb{S}^5 is not known.

By contrast very little is known about which compact, simply connected 5-manifolds occur as links of weighted homogeneous singularities. It is known that not every Seifert bundle occurs [Kol06, Lem.49] but a full answer seems unlikely.

Nothing seems to be known in higher dimensions.

16 (Einstein metrics on weighted homogeneous links). By a result of [Kob63], the link of a cone over a smooth projective variety $Z \subset \mathbb{P}^N$ carries a natural Einstein metric iff $-K_Z$ is a positive multiple of the hyperplane class and Z carries a Kähler–Einstein metric. This was generalized by [BG00] to weighted cones. Here one needs to work with an orbifold canonical class $K_X + \Delta$ and a suitable orbifold Kähler–Einstein metric on (X, Δ) .

This approach was used to construct new Einstein metrics on spheres and exotic spheres [BGK05, BGKT05] and on many 5-manifolds [Kol05, Kol07a, Kol09].

See [BG08] for a comprehensive treatment.

2. Construction of singularities

The construction has 5 main steps, none of which is fully understood at the moment. After summarizing them, we discuss the difficulties in more detail. Although the steps can not be carried out in full generality, we understand enough about them to obtain the main theorems.

17 (Main steps of the construction).

Step.17.1. For a simplicial complex C construct projective simple normal crossing varieties $V(C)$ such that $\mathcal{D}(V(C)) \cong C$.

Step.17.2. For a projective simple normal crossing variety V construct a smooth variety $Y(V)$ that contains V as a divisor.

Step.17.3. For a smooth variety Y containing a simple normal crossing divisor D construct an isolated singularity $(x \in X)$ such that $(D \subset Y)$ is a resolution of $(x \in X)$.

Step.17.4. Describe the link $L(x \in X)$ in terms of the topology of D and the Chern class of the normal bundle of D .

Step.17.5. Describe the relationship between the properties of the singularity $(x \in X)$ and the original simplicial complex C .

18 (Discussion of Step 17.1). I believe that for every simplicial complex C there are many projective simple normal crossing varieties $V(C)$ such that $\mathcal{D}(V(C)) \cong C$.²

There seem to be two main difficulties of a step-by-step approach.

First, topology would suggest that one should build up the skeletons of $V(C)$ one dimension at a time. It is easy to obtain the 1-skeleton by gluing rational curves. The 2-skeleton is still straightforward since rational surfaces do contain cycles of rational curves of arbitrary length. However, at the next step we run into a problem similar to Step 17.2 and usually a 2-skeleton can not be extended to a 3-skeleton. Our solution in [KK11] is to work with triangulations of n -dimensional submanifolds with boundary in \mathbb{R}^n . The ambient \mathbb{R}^n gives a rigidification and this makes it possible to have a consistent choice for all the strata.

Second, even if we construct a simple normal crossing variety V , it is not easy to decide whether it is projective. This is illustrated by the following example of “triangular pillows” [KK11, Exmp.34].

Let us start with an example that is not simple normal crossing.

Take 2 copies $\mathbb{P}_i^2 := \mathbb{P}^2(x_i : y_i : z_i)$ of \mathbb{CP}^2 and the triangles $C_i := (x_i y_i z_i = 0) \subset \mathbb{P}_i^2$. Given $c_x, c_y, c_z \in \mathbb{C}^*$, define $\phi(c_x, c_y, c_z) : C_1 \rightarrow C_2$ by $(0 : y_1 : z_1) \mapsto (0 : y_1 : c_z z_1)$, $(x_1 : 0 : z_1) \mapsto (c_x x_1 : 0 : z_1)$ and $(x_1 : y_1 : 0) \mapsto (x_1 : c_y y_1 : 0)$ and glue the 2 copies of \mathbb{P}^2 using $\phi(c_x, c_y, c_z)$ to get the surface $S(c_x, c_y, c_z)$.

We claim that $S(c_x, c_y, c_z)$ is projective iff the product $c_x c_y c_z$ is a root of unity.

To see this note that $\text{Pic}^0(C_i) \cong \mathbb{C}^*$ and $\text{Pic}^r(C_i)$ is a principal homogeneous space under \mathbb{C}^* for every $r \in \mathbb{Z}$. We can identify $\text{Pic}^3(C_i)$ with \mathbb{C}^* using the restriction of the ample generator L_i of $\text{Pic}(\mathbb{P}_i^2) \cong \mathbb{Z}$ as the base point.

The key observation is that $\phi(c_x, c_y, c_z)^* : \text{Pic}^3(C_2) \rightarrow \text{Pic}^3(C_1)$ is multiplication by $c_x c_y c_z$. Thus if $c_x c_y c_z$ is an r th root of unity then L_1^r and L_2^r glue together to an ample line bundle but otherwise $S(c_x, c_y, c_z)$ carries only the trivial line bundle.

²This is now proved in [Kol13a].

We can create a similar simple normal crossing example by smoothing the triangles C_i . That is, we take 2 copies $\mathbb{P}_i^2 := \mathbb{P}^2(x_i : y_i : z_i)$ of \mathbb{CP}^2 and smooth elliptic curves $E_i := (x_i^3 + y_i^3 + z_i^3 = 0) \subset \mathbb{P}_i^2$.

Every automorphism $\tau \in \text{Aut}(x^3 + y^3 + z^3 = 0)$ can be identified with an isomorphism $\tau : E_1 \cong E_2$, giving a simple normal crossing surface $S(\tau)$. The above argument then shows that $S(\tau)$ is projective iff $\tau^m = 1$ for some $m > 0$.

These examples are actually not surprising. One can think of the surfaces $S(c_x, c_y, c_z)$ and $S(\tau)$ as degenerate K3 surfaces of degree 2 and K3 surfaces have non-projective deformations. Similarly, $S(c_x, c_y, c_z)$ and $S(\tau)$ can be non-projective. One somewhat unusual aspect is that while a smooth K3 surface is projective iff it is a scheme, the above singular examples are always schemes yet many of them are non-projective.

19 (Discussion of Step 17.2). This is surprisingly subtle. First note that not every projective simple normal crossing variety V can be realized as a divisor on a smooth variety Y . A simple obstruction is the following.

Let Y be a smooth variety and $D_1 + D_2$ a simple normal crossing divisor on Y . Set $Z := D_1 \cap D_2$. Then $N_{Z, D_2} \cong N_{D_1, Y}|_Z$ where $N_{X, Y}$ denotes the normal bundle of $X \subset Y$.

Thus if $V = V_1 \cup V_2$ is a simple normal crossing variety with $W := V_1 \cap V_2$ such that N_{W, V_2} is not the restriction of any line bundle from V_1 then V is not a simple normal crossing divisor in a nonsingular variety.

I originally hoped that such normal bundle considerations give necessary and sufficient conditions, but recent examples of [Fuj12a, Fuj12b] show that this is not the case.

For now, no necessary and sufficient conditions of embeddability are known. In the original papers [Kol11, KK11] we went around this problem by first embedding a simple normal crossing variety V into a singular variety Y and then showing that for the purposes of computing the fundamental group of the link the singularities of Y do not matter.

We improve on this in Section 6.

20 (Discussion of Step 17.3). By a result of [Art70], a compact divisor contained in a smooth variety $D = \cup_i D_i \subset Y$ can be contracted to a point if there are positive integers m_i such that $\mathcal{O}_Y(-\sum_i m_i D_i)|_{D_j}$ is ample for every j .

It is known that this condition is not necessary and no necessary and sufficient characterizations are known. However, it is easy to check the above condition in our examples.

21 (Discussion of Step 17.4). This approach, initiated in [Mum61], has been especially successful for surfaces.

In principle the method of [Mum61] leads to a complete description of the link, but it seems rather difficult to perform explicit computations. Computing the fundamental group of the links seems rather daunting in general. Fortunately, we managed to find some simple conditions that ensure

that the natural maps

$$\pi_1(L(x \in X)) \rightarrow \pi_1(\mathcal{R}(X)) \rightarrow \pi_1(\mathcal{DR}(X))$$

are isomorphisms. However, these simple conditions force D to be more complicated than necessary, in particular we seem to lose control of the canonical class of X .

22 (Discussion of Step 17.5). For surfaces there is a very tight connection between the topology of the link and the algebro-geometric properties of a singularity. In higher dimension, one can obtain very little information from the topology alone. As we noted, there are many examples where X is a topological manifold yet very singular as a variety.

There is more reason to believe that algebro-geometric properties restrict the topology. For example, the results of Section 7 rely on the observation that if $(x \in X)$ is a rational (or even just 1-rational) singularity then $H_1(L(x \in X), \mathbb{Q}) = 0$.

3. Voronoi complexes

DEFINITION 23. A (convex) *Euclidean polyhedron* is a subset P of \mathbb{R}^n given by a finite collection of linear inequalities (some of which may be strict and some not). A *face* of P is a subset of P which is given by converting some of these non-strict inequalities to equalities.

A *Euclidean polyhedral complex* in \mathbb{R}^n is a collection of closed Euclidean polyhedra \mathcal{C} in \mathbb{R}^n such that

- (1) if $P \in \mathcal{C}$ then every face of P is in \mathcal{C} and
- (2) if $P_1, P_2 \in \mathcal{C}$ then $P_1 \cap P_2$ is a face of both of the P_i (or empty).

The union of the faces of a Euclidean polyhedral complex \mathcal{C} is denoted by $|\mathcal{C}|$.

For us the most important examples are the following.

DEFINITION 24 (Voronoi complex). Let $Y = \{y_i : i \in I\} \subset \mathbb{R}^n$ be a finite subset. For each $i \in I$ the corresponding *Voronoi cell* is the set of points that are closer to y_i than to any other y_j , that is

$$V_i := \{x \in \mathbb{R}^n : d(x, y_i) \leq d(x, y_j), \forall j \in I\}$$

where $d(x, y)$ denotes the Euclidean distance. Each cell V_i is a closed (possibly unbounded) polyhedron in \mathbb{R}^n .

The Voronoi cells and their faces give a Euclidean polyhedral complex, called the *Voronoi complex* or *Voronoi tessellation* associated to Y .

For a subset $J \subset I$ let H_J denote the linear subspace

$$H_J := \{x \in \mathbb{R}^n : d(x, y_i) = d(x, y_j) \forall i, j \in J\}.$$

The affine span of each face of the Voronoi complex is one of the H_J . If J has 2 elements $\{i, j\}$ then H_{ij} is a hyperplane $H_{ij} = \{x \in \mathbb{R}^n : d(x, y_i) = d(x, y_j)\}$.

A Voronoi complex is called *simple* if for every k , every codimension k face is contained in exactly $k + 1$ Voronoi cells. Not every Voronoi complex is simple, but it is easy to see that among finite subsets $Y \subset \mathbb{R}^n$ those with a simple Voronoi complex $\mathcal{C}(Y)$ form an open and dense set.

Let \mathcal{C} be a simple Voronoi complex. For each face $F \in \mathcal{C}$, let V_i for $i \in I_F$ be the Voronoi cells containing F . The vertices $\{y_i : i \in I_F\}$ form a simplex whose dimension equals the codimension of F . These simplices define the *Delaunay triangulation* dual to \mathcal{C} .

THEOREM 25. [KK11, Cor.21] *Let T be a finite simplicial complex of dimension n . Then there is an embedding $j : T \hookrightarrow \mathbb{R}^{2n+1}$, a simple Voronoi complex \mathcal{C} in \mathbb{R}^{2n+1} and a subcomplex $\mathcal{C}(T) \subset \mathcal{C}$ of pure dimension $2n + 1$ containing $j(T)$ such that the inclusion $j(T) \subset |\mathcal{C}(T)|$ is a homotopy equivalence.*

Outline of the proof. First we embed T into \mathbb{R}^{2n+1} . This is where the dimension increase comes from. (We do not need an actual embedding, only an embedding up-to homotopy, which is usually easier to get.)

Then we first use a result of [Hir62] which says that if T is a finite simplicial complex in a smooth manifold \mathbf{R} then there exists a codimension 0 compact submanifold $M \subset \mathbf{R}$ with smooth boundary containing T such that the inclusion $T \subset M$ is a homotopy equivalence.

Finally we construct a Voronoi complex using M .

Let $M \subset \mathbb{R}^m$ be a compact subset, $Y \subset \mathbb{R}^m$ a finite set of points and $\mathcal{C}(Y)$ the corresponding Voronoi complex. Let $\mathcal{C}_m(Y, M)$ be the collection of those m -cells in the Voronoi complex $\mathcal{C}(Y)$ whose intersection with M is not empty and $\mathcal{C}(Y, M)$ the polyhedral complex consisting of the cells in $\mathcal{C}_m(Y, M)$ and their faces. Then $M \subset |\mathcal{C}(Y, M)|$.

We conclude by using a theorem of [Cai61] that says that if M is a C^2 -submanifold with C^2 -boundary then for a suitably fine mesh of points $Y \subset \mathbb{R}^m$ the inclusion $M \subset |\mathcal{C}(Y, M)|$ is a homotopy equivalence. \square

4. Simple normal crossing varieties

Let \mathcal{C} be a purely m -dimensional, compact subcomplex of a simple Voronoi complex in \mathbb{R}^m . Our aim is to construct a projective simple normal crossing variety $V(\mathcal{C})$ whose dual complex naturally identifies with the Delaunay triangulation of \mathcal{C} .

26 (First attempt). For each m -polytope $P_i \in \mathcal{C}$ we associate a copy $\mathbb{P}_{(i)}^m = \mathbb{CP}^m$. For a subvariety $W \subset \mathbb{CP}^m$ we let $W_{(i)}$ or $W^{(i)}$ denote the corresponding subvariety of $\mathbb{P}_{(i)}^m$.

If P_i and P_j have a common face F_{ij} of dimension $m - 1$ then the complexification of the affine span of F_{ij} gives hyperplanes $H_{ij}^{(i)} \subset \mathbb{P}_{(i)}^m$ and $H_{ij}^{(j)} \subset \mathbb{P}_{(j)}^m$. Moreover, $H_{ij}^{(i)}$ and $H_{ij}^{(j)}$ come with a natural identification $\sigma_{ij} : H_{ij}^{(i)} \cong H_{ij}^{(j)}$.

We use σ_{ij} to glue $\mathbb{P}_{(i)}^m$ and $\mathbb{P}_{(j)}^m$ together. The resulting variety is isomorphic to the union of 2 hyperplanes in \mathbb{CP}^{m+1} .

It is harder to see what happens if we try to perform all these gluings σ_{ij} simultaneously.

Let $\amalg_i \mathbb{P}_{(i)}^m$ denote the disjoint union of all the $\mathbb{P}_{(i)}^m$. Each σ_{ij} defines a relation that identifies a point $p_{(i)} \in H_{ij}^{(i)} \subset \mathbb{P}_{(i)}^m$ with its image $p_{(j)} = \sigma_{ij}(p_{(i)}) \in H_{ij}^{(j)} \subset \mathbb{P}_{(j)}^m$. Let Σ denote the equivalence relation generated by all the σ_{ij} .

It is easy to see (cf. [Kol12, Lem.17]) that there is a projective algebraic variety

$$\amalg_i \mathbb{P}_{(i)}^m \longrightarrow (\amalg_i \mathbb{P}_{(i)}^m)/\Sigma \longrightarrow \mathbb{CP}^m$$

whose points are exactly the equivalence classes of Σ .

This gives the correct simple normal crossing variety if $m = 1$ but already for $m = 2$ we have problems. For instance, consider three 2-cells P_i, P_j, P_k such that P_i and P_j have a common face F_{ij} , P_j and P_k have a common face F_{jk} but $P_i \cap P_k = \emptyset$. The problem is that while F_{ij} and F_{jk} are disjoint, their complexified spans are lines in \mathbb{CP}^2 hence they intersect at a point q . Thus σ_{ij} identifies $q_{(i)} \in \mathbb{P}_{(i)}^2$ with $q_{(j)} \in \mathbb{P}_{(j)}^2$ and σ_{jk} identifies $q_{(j)} \in \mathbb{P}_{(j)}^2$ with $q_{(k)} \in \mathbb{P}_{(k)}^2$ thus the equivalence relation Σ identifies $q_{(i)} \in \mathbb{P}_{(i)}^2$ with $q_{(k)} \in \mathbb{P}_{(k)}^2$. Thus in $(\amalg_i \mathbb{P}_{(i)}^m)/\Sigma$ the images of $\mathbb{P}_{(i)}^2$ and of $\mathbb{P}_{(k)}^2$ are not disjoint.

In order to get the correct simple normal crossing variety, we need to remove these extra intersection points. In higher dimensions we need to remove various linear subspaces as well.

DEFINITION 27 (Essential and parasitic intersections). Let \mathcal{C} be a Voronoi complex on \mathbb{R}^m defined by the points $\{y_i : i \in I\}$. We have the linear subspaces H_J defined in (24). Assume for simplicity that $J_1 \neq J_2$ implies that $H_{J_1} \neq H_{J_2}$.

Let $P \subset \mathbb{R}^m$ be a Voronoi cell. We say that H_J is *essential* for P if it is the affine span of a face of P . Otherwise it is called *parasitic* for P .

LEMMA 28. *Let $P \subset \mathbb{R}^m$ be a simple Voronoi cell.*

- (1) *Every essential subspace L of dimension $\leq m - 2$ is contained in a unique smallest parasitic subspace which has dimension $\dim L + 1$.*
- (2) *The intersection of two parasitic subspaces is again parasitic.*

Proof. There is a point $y_p \in P$ and a subset $J \subset I$ such that H_{ip} are spans of faces of P for $i \in J$ and $L = \cap_{i \in J} H_{ip}$. Thus the unique $\dim L + 1$ -dimensional parasitic subspace containing L is H_J .

Assume that L_1, L_2 are parasitic. If $L_1 \cap L_2$ is essential then there is a unique smallest parasitic subspace $L' \supset L_1 \cap L_2$. Then $L' \subset L_i$ a contradiction. \square

29 (Removing parasitic intersections). Let $\{H_s : s \in S\}$ be a finite set of hyperplanes of \mathbb{CP}^m . For $Q \subset S$ set $H_Q := \cap_{s \in Q} H_s$. Let $\mathcal{P} \subset 2^S$ be a subset closed under unions.

Set $\pi_0 : P^0 \cong \mathbb{CP}^m$. If $\pi_r : P^r \rightarrow \mathbb{CP}^m$ is already defined then let $P^{r+1} \rightarrow P^r$ denote the blow-up of the union of birational transforms of all the H_Q such that $Q \in \mathcal{P}$ and $\dim H_Q = r$. Then π_{r+1} is the composite $P^{r+1} \rightarrow P^r \rightarrow \mathbb{CP}^m$.

Note that we blow up a disjoint union of smooth subvarieties since any intersection of the r -dimensional H_Q is lower dimensional, hence it was removed by an earlier blow up. Finally set $\Pi : \tilde{P} := P^{m-2} \rightarrow \mathbb{CP}^m$.

Let \mathcal{C} be a pure dimensional subcomplex of a Voronoi complex as in (25). For each cell $P_i \in \mathcal{C}$ we use (29) with

$$\mathcal{P}_i := \{\text{parasitic intersections for } P_i\}$$

to obtain $\tilde{P}_{(i)}$. Note that if P_i and P_j have a common codimension 1 face F_{ij} then we perform the same blow-ups on the complexifications $H_{ij}^{(i)} \subset \mathbb{P}_{(i)}^m$ and $H_{ij}^{(j)} \subset \mathbb{P}_{(j)}^m$. Thus $\sigma_{ij} : H_{ij}^{(i)} \cong H_{ij}^{(j)}$ lifts to the birational transforms

$$\tilde{\sigma}_{ij} : \tilde{H}_{ij}^{(i)} \cong \tilde{H}_{ij}^{(j)}.$$

As before, the $\tilde{\sigma}_{ij}$ define an equivalence relation $\tilde{\Sigma}$ on $\Pi_i \tilde{P}_{(i)}$. With these changes, the approach outlined in Paragraph 26 does work and we get the following.

THEOREM 30. [KK11, Prop.28] *With the above notation there is a projective, simple normal crossing variety*

$$V(\mathcal{C}) := (\Pi_i \tilde{P}_{(i)}) / \tilde{\Sigma}$$

with the following properties.

- (1) *There is a finite morphism $\Pi_i \tilde{P}_{(i)} \longrightarrow V(\mathcal{C})$ whose fibers are exactly the equivalence classes of $\tilde{\Sigma}$.*
- (2) *The dual complex $\mathcal{D}(V(\mathcal{C}))$ is naturally identified with the Delaunay triangulation of \mathcal{C} .*

Comments on the proof. The existence of $V(\mathcal{C})$ is relatively easy either directly as in [KK11, Prop.31] or using the general theory of quotients by finite equivalence relations as in [Kol12].

As we noted in Paragraph 18 the projectivity of such quotients is a rather delicate question since the maps $\tilde{P}_{(i)} \rightarrow \mathbb{CP}^m$ are not finite any more.

The main advantage we have here is that each $\tilde{P}_{(i)}$ comes with a specific sequence of blow-ups $\Pi_i : \tilde{P}_{(i)} \rightarrow \mathbb{CP}^m$ and this enables us to write down explicit, invertible, ample subsheaves $A_i \subset \Pi_i^* \mathcal{O}_{\mathbb{CP}^m}(N)$ for some $N \gg 1$ that glue together to give an ample invertible sheaf on $V(T)$. For details see [KK11, Par.32]. \square

The culmination of the results of the last 2 sections is the following.

THEOREM 31. [KK11, Thm.29] *Let T be a finite cell complex. Then there is a projective simple normal crossing variety Z_T such that*

- (1) $\mathcal{D}(Z_T)$ is homotopy equivalent to T ,
- (2) $\pi_1(Z_T) \cong \pi_1(T)$ and
- (3) $H^i(Z_T, \mathcal{O}_{Z_T}) \cong H^i(T, \mathbb{C})$ for every $i \geq 0$.

Proof. We have already established (1) in (30), moreover the construction yields a simple normal crossing variety Z_T whose strata are all rational varieties. In particular every stratum $W \subset Z_T$ is simply connected and $H^r(W, \mathcal{O}_W) = 0$ for every $r > 0$. Thus (2–3) follow from Lemmas 32–33. \square

The proof of the following lemma is essentially in [GS75, pp.68–72]. More explicit versions can be found in [FM83, pp.26–27] and [Ish85, ABW09].

LEMMA 32. *Let X be a simple normal crossing variety over \mathbb{C} with irreducible components $\{X_i : i \in I\}$. Let $T = D(X)$ be the dual complex of X .*

- (1) *There are natural injections $H^r(T, \mathbb{C}) \hookrightarrow H^r(X, \mathcal{O}_X)$ for every r .*
- (2) *Assume that $H^r(W, \mathcal{O}_W) = 0$ for every $r > 0$ and for every stratum $W \subset X$. Then $H^r(X, \mathcal{O}_X) = H^r(T, \mathbb{C})$ for every r .* \square

The following comparison result is rather straightforward.

LEMMA 33. [Cor92, Prop.3.1] *Using the notation of (32) assume that every stratum $W \subset X$ is 1-connected. Then $\pi_1(X) \cong \pi_1(\mathcal{D}(X))$.* \square

5. Generic embeddings of simple normal crossing varieties

The following is a summary of the construction of [Kol11]; see also [Kol13b, Sec.3.4] for an improved version.

34. Let Z be a projective, local complete intersection variety of dimension n and choose any embedding $Z \subset P$ into a smooth projective variety of dimension N . (We can take $P = \mathbb{P}^N$ for $N \gg 1$.) Let L be a sufficiently ample line bundle on P . Let $Z \subset Y_1 \subset P$ be the complete intersection of $(N - n - 1)$ general sections of $L(-Z)$. Set

$$Y := B_{(-Z)}Y_1 := \text{Proj}_{Y_1} \sum_{m=0}^{\infty} \mathcal{O}_{Y_1}(mZ).$$

(Note that this is not the blow-up of Z but the blow-up of its inverse in the class group.)

It is proved in [Kol11] that the birational transform of Z in Y is a Cartier divisor isomorphic to Z and there is a contraction morphism

$$\begin{array}{ccc} Z & \subset & Y \\ \downarrow & & \downarrow \pi \\ 0 & \in & X \end{array} \tag{34.1}$$

such that $Y \setminus Z \cong X \setminus \{0\}$. If Y is smooth then $\mathcal{DR}(0 \in X) = \mathcal{D}(Z)$ and we are done with Theorem 7. However, the construction of [Kol11] yields a

smooth variety Y only if $\dim Z = 1$ or Z is smooth. (By (19) this limitation is not unexpected.)

In order to resolve singularities of Y we need a detailed description of them. This is a local question, so we may assume that $Z \subset \mathbb{C}_{\mathbf{x}}^N$ is a complete intersection defined by $f_1 = \dots = f_{N-n} = 0$. Let $Z \subset Y_1 \subset \mathbb{C}^N$ be a general complete intersection defined by equations

$$h_{i,1}f_1 + \dots + h_{i,N-n}f_{N-n} = 0 \quad \text{for } i = 1, \dots, N-n-1.$$

Let $H = (h_{ij})$ be the $(N-n-1) \times (N-n)$ matrix of the system and H_i the submatrix obtained by removing the i th column. By [Kol11] or [Kol13b, Sec.3.2], an open neighborhood of $Z \subset Y$ is defined by the equations

$$(f_i = (-1)^i \cdot t \cdot \det H_i : i = 1, \dots, N-n) \subset \mathbb{C}_{\mathbf{x}}^N \times \mathbb{C}_t. \quad (34.2)$$

Assume now that Z has hypersurface singularities. Up-to permuting the f_i and passing to a smaller open set, we may assume that df_2, \dots, df_{N-n} are linearly independent everywhere along Z . Then the singularities of Y all come from the equation

$$f_1 = -t \cdot \det H_1. \quad (34.3)$$

Our aim is to write down local normal forms for Y along Z in the normal crossing case.

On \mathbb{C}^N there is a stratification $\mathbb{C}^N = R_0 \supset R_1 \supset \dots$ where R_i is the set of points where $\operatorname{rank} H_1 \leq (N-n-1) - i$. Since the h_{ij} are general, $\operatorname{codim}_W R_i = i^2$ and we may assume that every stratum of Z is transversal to each $R_i \setminus R_{i+1}$ (cf. Paragraph 37).

Let $S \subset Z$ be any stratum and $p \in S$ a point such that $p \in R_m \setminus R_{m+1}$. We can choose local coordinates $\{x_1, \dots, x_d\}$ and $\{y_{rs} : 1 \leq r, s \leq m\}$ such that, in a neighborhood of p ,

$$f_1 = x_1 \cdots x_d \quad \text{and} \quad \det H_1 = \det(y_{rs} : 1 \leq r, s \leq m).$$

Note that $m^2 \leq \dim S = n-d$, thus we can add $n-d-m^2$ further coordinates y_{ij} to get a complete local coordinate system on S .

Then the n coordinates $\{x_k, y_{ij}\}$ determine a map

$$\sigma : \mathbb{C}^N \times \mathbb{C}_t \rightarrow \mathbb{C}^n \times \mathbb{C}_t$$

such that $\sigma(Y)$ is defined by the equation

$$x_1 \cdots x_d = t \cdot \det(y_{rs} : 1 \leq r, s \leq m).$$

Since df_2, \dots, df_{N-n} are linearly independent along Z , we see that $\sigma|_Y$ is étale along $Z \subset Y$.

We can summarize these considerations as follows.

PROPOSITION 35. *Let Z be a normal crossing variety of dimension n . Then there is a normal singularity $(0 \in X)$ of dimension $n+1$ and a proper, birational morphism $\pi : Y \rightarrow X$ such that $\operatorname{red} \pi^{-1}(0) \cong Z$ and for every point $p \in \pi^{-1}(0)$ we can choose local (étale or analytic) coordinates called*

$\{x_i : i \in I_p\}$ and $\{y_{rs} : 1 \leq r, s \leq m_p\}$ (plus possibly other unnamed coordinates) such that one can write the local equations of $Z \subset Y$ as

$$\left(\prod_{i \in I_p} x_i = t = 0\right) \subset \left(\prod_{i \in I_p} x_i = t \cdot \det(y_{rs} : 1 \leq r, s \leq m_p)\right) \subset \mathbb{C}^{n+2}. \quad \square$$

36 (Proof of Theorem 7). Let T be a finite cell complex. By (31) there is a projective simple normal crossing variety Z such that $\mathcal{D}(Z)$ is homotopy equivalent to T , $\pi_1(Z) \cong \pi_1(T)$ and $H^i(Z, \mathcal{O}_Z) \cong H^i(T, \mathbb{C})$ for every $i \geq 0$.

Then Proposition 35 constructs a singularity $(0 \in X)$ with a partial resolution

$$\begin{array}{ccc} Z & \subset & Y \\ \downarrow & & \downarrow \pi \\ 0 & \in & X \end{array} \quad (36.1)$$

The hardest is to check that we can resolve the singularities of Y without changing the homotopy type of the dual complex of the exceptional divisor. This is done in Section 6.

In order to show (7.2–3) we need further information about the varieties and maps in (36.1).

First, Y has rational singularities. This is easy to read off from their equations. (For the purposes of Theorem 3, we only need the case $\dim Y = 3$ when the only singularities we have are ordinary double points with local equation $x_1 x_2 = t y_{11}$.)

Second, we can arrange that Z has very negative normal bundle in Y . By a general argument this implies that $R^i \pi_* \mathcal{O}_Y \cong H^i(Z, \mathcal{O}_Z)$, proving (7.3); see [Kol11, Prop.9] for details.

Finally we need to compare $\pi_1(Z)$ with $\pi_1(L(0 \in X))$. There is always a surjection

$$\pi_1(L(0 \in X)) \twoheadrightarrow \pi_1(Z) \quad (36.2)$$

but it can have a large kernel. We claim however, that with suitable choices we can arrange that (36.2) is an isomorphism. It is easiest to work not on $Z \subset Y$ but on a resolution $Z' \subset Y'$.

More generally, let W be a smooth variety, $D = \cup_i D_i \subset W$ a simple normal crossing divisor and $T \supset D$ a regular neighborhood with boundary $M = \partial T$. There is a natural (up to homotopy) retraction map $T \rightarrow D$ which induces $M \rightarrow D$ hence a surjection $\pi_1(M) \twoheadrightarrow \pi_1(D)$ whose kernel is generated (as a normal subgroup) by the simple loops γ_i around the D_i .

In order to understand this kernel, assume first that D is smooth. Then $M \rightarrow D$ is a circle bundle hence there is an exact sequence

$$\pi_2(D) \xrightarrow{c_1 \cap} \mathbb{Z} \cong \pi_1(\mathbb{S}^1) \rightarrow \pi_1(M) \rightarrow \pi_1(D) \rightarrow 1$$

where c_1 is the Chern class of the normal bundle of D in X . Thus if $c_1 \cap \alpha = 1$ for some $\alpha \in \pi_2(D)$ then $\pi_1(M) \cong \pi_1(D)$. In the general case, arguing as above we see that $\pi_1(M) \cong \pi_1(D)$ if the following holds:

- (3) For every i there is a class $\alpha_i \in \pi_2(D_i^0)$ such that $c_1(N_{D_i, X}) \cap \alpha_i = 1$ where $D_i^0 := D_i \setminus \{\text{other components of } D\}$.

Condition (3) is typically very easy to achieve in our constructions. Indeed, we obtain the D_i^0 by starting with \mathbb{CP}^m , blowing it up many times and then removing a few divisors. Thus we end up with very large $H_2(D_i^0, \mathbb{Z})$ and typically the D_i^0 are even simply connected, hence $\pi_2(D_i^0) = H_2(D_i^0, \mathbb{Z})$. \square

37 (Determinantal varieties). We have used the following basic properties of determinantal varieties. These are quite easy to prove directly; see [Har95, 12.2 and 14.16] for a more general case.

Let V be a smooth, affine variety, and $\mathcal{L} \subset \mathcal{O}_V$ a finite dimensional sub vector space without common zeros. Let $H = (h_{ij})$ be an $n \times n$ matrix whose entries are general elements in \mathcal{L} . For a point $p \in V$ set $m_p = \text{corank } H(p)$. Then there are local analytic coordinates $\{y_{rs} : 1 \leq r, s \leq m_p\}$ (plus possibly other unnamed coordinates) such that, in a neighborhood of p ,

$$\det H = \det(y_{rs} : 1 \leq r, s \leq m_p).$$

In particular, $\text{mult}_p(\det H) = \text{corank } H(p)$, for every m the set of points $R_m \subset V$ where $\text{corank } H(p) \geq m$ is a subvariety of pure codimension m^2 and $\text{Sing } R_m = R_{m+1}$.

6. Resolution of generic embeddings

In this section we start with the varieties constructed in Proposition 35 and resolve their singularities. Surprisingly, the resolution process described in Paragraphs 39–44 leaves the dual complex unchanged and we get the following.

THEOREM 38. *Let Z be a projective simple normal crossing variety of dimension n . Then there is a normal singularity $(0 \in X)$ of dimension $(n+1)$ and a resolution $\pi : Y \rightarrow X$ such that $E := \pi^{-1}(0) \subset Y$ is a simple normal crossing divisor and its dual complex $\mathcal{D}(E)$ is naturally identified with $\mathcal{D}(Z)$. (More precisely, there is a morphism $E \rightarrow Z$ that induces a birational map on every stratum.)*

39 (Inductive set-up for resolution). The object we try to resolve is a triple

$$(Y, E, F) := (Y, \sum_{i \in I} E_i, \sum_{j \in J} a_j F_j) \quad (39.1)$$

where Y is a variety over \mathbb{C} , E_i, F_j are codimension 1 subvarieties and $a_j \in \mathbb{N}$. (The construction (34) produces a triple $(Y, E := Z, F := \emptyset)$. The role of the F_j is to keep track of the exceptional divisors as we resolve the singularities of Y .)

We assume that E is a simple normal crossing variety and for every point $p \in E$ there is a (Euclidean) open neighborhood $p \in Y_p \subset Y$, an embedding $\sigma_p : Y_p \hookrightarrow \mathbb{C}^{\dim Y + 1}$ whose image can be described as follows.

There are subsets $I_p \subset I$ and $J_p \subset J$, a natural number $m_p \in \mathbb{N}$ and coordinates in $\mathbb{C}^{\dim Y + 1}$ called

$$\{x_i : i \in I_p\}, \{y_{rs} : 1 \leq r, s \leq m_p\}, \{z_j : j \in J_p\} \quad \text{and} \quad t$$

(plus possibly other unnamed coordinates) such that $\sigma_p(Y_p) \subset \mathbb{C}^{\dim Y+1}$ is an open subset of the hypersurface

$$\prod_{i \in I_p} x_i = t \cdot \det(y_{rs} : 1 \leq r, s \leq m_p) \cdot \prod_{j \in J_p} z_j^{a_j}. \quad (39.2)$$

Furthermore,

$$\begin{aligned} \sigma_p(E_i) &= (t = x_i = 0) \cap \sigma_p(Y_p) \quad \text{for } i \in I_p \quad \text{and} \\ \sigma_p(F_j) &= (z_j = 0) \cap \sigma_p(Y_p) \quad \text{for } j \in J_p. \end{aligned}$$

We do not impose any compatibility condition between the local equations on overlapping charts.

We say that (Y, E, F) is *resolved* at p if Y is smooth at p .

The key technical result of this section is the following.

PROPOSITION 40. *Let (Y, E, F) be a triple as above. Then there is a resolution of singularities $\pi : (Y', E', F') \rightarrow (Y, E, F)$ such that*

- (1) *Y' is smooth and E' is a simple normal crossing divisor,*
- (2) *$E' = \pi^{-1}(E)$,*
- (3) *every stratum of E' is mapped birationally to a stratum of E and*
- (4) *π induces an identification $\mathcal{D}(E') = \mathcal{D}(E)$.*

Proof. The resolution will be a composite of explicit blow-ups of smooth subvarieties (except at the last step). We use the local equations to describe the blow-up centers locally. Thus we need to know which locally defined subvarieties make sense globally. For example, choosing a divisor F_{j_1} specifies the local divisor $(z_{j_1} = 0)$ at every point $p \in F_{j_1}$. Similarly, choosing two divisors E_{i_1}, E_{i_2} gives the local subvarieties $(t = x_{i_1} = x_{i_2} = 0)$ at every point $p \in E_{i_1} \cap E_{i_2}$. (Here it is quite important that the divisors E_i are themselves smooth. The algorithm does not seem to work if the E_i have self-intersections.) Note that by contrast $(x_{i_1} = x_{i_2} = 0) \subset Y$ defines a local divisor which has no global meaning. Similarly, the vanishing of any of the coordinate functions y_{rs} has no global meaning.

To a point $p \in \text{Sing } E$ we associate the local invariant

$$\text{Deg}(p) := (\deg_x(p), \deg_y(p), \deg_z(p)) = (|I_p|, m_p, \sum_{j \in J_p} a_j).$$

It is clear that $\deg_x(p)$ and $\deg_z(p)$ do not depend on the local coordinates chosen. We see in (42) that $\deg_y(p)$ is also well defined if $p \in \text{Sing } E$. The degrees $\deg_x(p), \deg_y(p), \deg_z(p)$ are constructible and upper semi continuous functions on $\text{Sing } E$.

Note that Y is smooth at p iff either $\text{Deg}(p) = (1, *, *)$ or $\text{Deg}(p) = (*, 0, 0)$. If $\deg_x(p) = 1$ then we can rewrite the equation (39.2) as

$$x' = t \cdot \prod_j z_j^{a_j} \quad \text{where} \quad x' := x_1 + t \cdot (1 - \det(y_{rs})) \cdot \prod_j z_j^{a_j},$$

so if Y is smooth then $(Y, E + F)$ has only simple normal crossings along E . Thus the resolution constructed in Theorem 38 is a log resolution.

The usual method of Hironaka would start by blowing up the *highest* multiplicity points. This introduces new and rather complicated exceptional

divisors and I have not been able to understand how the dual complex changes.

In our case, it turns out to be better to look at a locus where $\deg_y(p)$ is maximal but instead of maximizing $\deg_x(p)$ or $\deg_z(p)$ we maximize the dimension. Thus we blow up subvarieties along which Y is not equimultiple. Usually this leads to a morass, but our equations separate the variables into distinct groups which makes these blow-ups easy to compute.

One can think of this as mixing the main step of the Hironaka method with the order reduction for monomial ideals (see, for instance, [Kol07b, Step 3 of 3.111]).

After some preliminary remarks about blow-ups of simple normal crossing varieties the proof of (40) is carried out in a series of steps (42–44).

We start with the locus where $\deg_y(p)$ is maximal and by a sequence of blow-ups we eventually achieve that $\deg_y(p) \leq 1$ for every singular point p . This, however, increases \deg_z . Then in 3 similar steps we lower the maximum of \deg_z until we achieve that $\deg_z(p) \leq 1$ for every singular point p . Finally we take care of the singular points where $\deg_y(p) + \deg_z(p) \geq 1$. \square

41 (Blowing up simple normal crossing varieties). Let Z be a simple normal crossing variety and $W \subset Z$ a subvariety. We say that W has *simple normal crossing* with Z if for each point $p \in Z$ there is an open neighborhood Z_p , an embedding $Z_p \hookrightarrow \mathbb{C}^{n+1}$ and subsets $I_p, J_p \subset \{0, \dots, n\}$ such that

$$Z_p = (\prod_{i \in I_p} x_i = 0) \quad \text{and} \quad W \cap Z_p = (x_j = 0 : j \in J_p).$$

This implies that for every stratum $Z_J \subset Z$ the intersection $W \cap Z_J$ is smooth (even scheme theoretically).

If W has simple normal crossing with Z then the blow-up $B_W Z$ is again a simple normal crossing variety. If W is one of the strata of Z , then $\mathcal{D}(B_W Z)$ is obtained from $\mathcal{D}(Z)$ by removing the cell corresponding to W and every other cell whose closure contains it. Otherwise $\mathcal{D}(B_W Z) = \mathcal{D}(Z)$. (In the terminology of [Kol13b, Sec.2.4], $B_W Z \rightarrow Z$ is a thrifty modification.)

As an example, let $Z = (x_1 x_2 x_3 = 0) \subset \mathbb{C}^3$. There are 7 strata and $\mathcal{D}(Z)$ is the 2-simplex whose vertices correspond to the planes $(x_i = 0)$.

Let us blow up a point $W = \{p\} \subset Z$ to get $B_p Z \subset B_p \mathbb{C}^3$. Note that the exceptional divisor $E \subset B_p \mathbb{C}^3$ is *not* a part of $B_p Z$ and $B_p Z$ still has 3 irreducible components.

If p is the origin, then the triple intersection is removed and $\mathcal{D}(B_p Z)$ is the boundary of the 2-simplex.

If p is not the origin, then $B_p Z$ still has 7 strata naturally corresponding to the strata of Z and $\mathcal{D}(B_p Z)$ is the 2-simplex.

We will be interested in situations where Y is a hypersurface in \mathbb{C}^{n+2} and $Z \subset Y$ is a Cartier divisor that is a simple normal crossing variety. Let $W \subset Y$ be a smooth, irreducible subvariety, not contained in Z such that

- (1) the scheme theoretic intersection $W \cap Z$ has simple normal crossing with Z

- (2) $\text{mult}_{Z \cap W} Z = \text{mult}_W Y$. (Note that this holds if $W \subset \text{Sing } Y$ and $\text{mult}_{Z \cap W} Z = 2$.)

Choose local coordinates (x_0, \dots, x_n, t) such that $W = (x_0 = \dots = x_i = 0)$ and $Z = (t = 0) \subset Y$. Let $f(x_0, \dots, x_n, t) = 0$ be the local equation of Y .

Blow up W to get $\pi : B_W Y \rightarrow Y$. Up to permuting the indices $0, \dots, i$, the blow-up $B_W Y$ is covered by coordinate charts described by the coordinate change

$$(x_0, x_1, \dots, x_i, x_{i+1}, \dots, x_n, t) = (x'_0, x'_1 x'_0, \dots, x'_i x'_0, x_{i+1}, \dots, x_n, t).$$

If $\text{mult}_W Y = d$ then the local equation of $B_W Y$ in the above chart becomes

$$(x'_0)^{-d} f(x'_0, x'_1 x'_0, \dots, x'_i x'_0, x_{i+1}, \dots, x_n, t) = 0.$$

By assumption (2), $(x'_0)^d$ is also the largest power that divides

$$f(x'_0, x'_1 x'_0, \dots, x'_i x'_0, x_{i+1}, \dots, x_n, 0),$$

hence $\pi^{-1}(Z) = B_{W \cap Z} Z$.

Observe finally that the conditions (1–2) can not be fulfilled in any interesting way if Y is smooth. Since we want $Z \cap W$ to be scheme theoretically smooth, if Y is smooth then condition (1) implies that $Z \cap W$ is disjoint from $\text{Sing } Z$.

(As an example, let $Y = \mathbb{C}^3$ and $Z = (xyz = 0)$. Take $W := (x = y = z)$. Note that W is transversal to every irreducible component of Z but $W \cap Z$ is a non-reduced point. The preimage of Z in $B_W Y$ does not have simple normal crossings.)

There are, however, plenty of examples where Y is singular along $Z \cap W$ and these are exactly the singular points that we want to resolve.

42 (Resolving the determinantal part). Let m be the largest size of a determinant occurring at a non-resolved point. Assume that $m \geq 2$ and let $p \in Y$ be a non-resolved point with $m_p = m$.

Away from $E \cup F$ the local equation of Y is

$$\prod_{i \in I_p} x_i = \det(y_{rs} : 1 \leq r, s \leq m).$$

Thus, the singular set of $Y_p \setminus (E \cup F)$ is

$$\bigcup_{(i, i')} (\text{rank}(y_{rs}) \leq m - 2) \cap (x_i = x_{i'} = 0)$$

where the union runs through all 2-element subsets $\{i, i'\} \subset I_p$. Thus the irreducible components of $\text{Sing } Y \setminus (E \cup F)$ are in natural one-to-one correspondence with the irreducible components of $\text{Sing } E$ and the value of $m = \deg_y(p)$ is determined by the multiplicity of any of these irreducible components at p .

Pick $i_1, i_2 \in I$ and we work locally with a subvariety

$$W'_p(i_1, i_2) := (\text{rank}(y_{rs}) \leq m - 2) \cap (x_{i_1} = x_{i_2} = 0).$$

Note that $W'_p(i_1, i_2)$ is singular if $m > 2$ and the subset of its highest multiplicity points is given by $\text{rank}(y_{rs}) = 0$. Therefore the locally defined subvarieties

$$W_p(i_1, i_2) := (y_{rs} = 0 : 1 \leq r, s \leq m) \cap (x_{i_1} = x_{i_2} = 0).$$

glue together to a well defined global smooth subvariety $W := W(i_1, i_2)$.

E is defined by ($t = 0$) thus $E \cap W$ has the same local equations as $W_p(i_1, i_2)$. In particular, $E \cap W$ has simple normal crossings with E and $E \cap W$ is not a stratum of E ; its codimension in the stratum $(x_{i_1} = x_{i_2} = 0)$ is m^2 .

Furthermore, E has multiplicity 2 along $E \cap W$, hence (41.2) also holds and so

$$\mathcal{D}(B_{E \cap W}) = \mathcal{D}(E).$$

We blow up $W \subset Y$. We will check that the new triple is again of the form (39). The local degree $\text{Deg}(p)$ is unchanged over $Y \setminus W$. The key assertion is that, over W , the maximum value of $\text{Deg}(p)$ (with respect to the lexicographic ordering) decreases. By repeating this procedure for every irreducible components of $\text{Sing } E$, we decrease the maximum value of $\text{Deg}(p)$. We can repeat this until we reach $\deg_y(p) \leq 1$ for every non-resolved point $p \in Y$.

(Note that this procedure requires an actual ordering of the irreducible components of $\text{Sing } E$, which is a non-canonical choice. If a finite group acts on Y , our resolution usually can not be chosen equivariant.)

Now to the local computation of the blow-up. Fix a point $p \in W$ and set $I_p^* := I_p \setminus \{i_1, i_2\}$. We write the local equation of Y as

$$x_{i_1} x_{i_2} \cdot L = t \cdot \det(y_{rs}) \cdot R \quad \text{where} \quad L := \prod_{i \in I_p^*} x_i \quad \text{and} \quad R := \prod_{j \in J_p} z_j^{a_j}.$$

Since $W = (x_{i_1} = x_{i_2} = y_{rs} = 0 : 1 \leq r, s \leq m)$ there are two types of local charts on the blow-up.

(1) There are two charts of the first type. Up to interchanging the subscripts 1, 2, these are given by the coordinate change

$$(x_{i_1}, x_{i_2}, y_{rs} : 1 \leq r, s \leq m) = (x'_{i_1}, x'_{i_2} x'_{i_1}, y'_{rs} x'_{i_1} : 1 \leq r, s \leq m).$$

After setting $z_w := x'_{i_1}$ the new local equation is

$$x'_{i_2} \cdot L = t \cdot \det(y'_{rs}) \cdot (z_w^{m^2-2} \cdot R).$$

The exceptional divisor is added to the F -divisors with coefficient $m^2 - 2$ and the new degree is $(\deg_x(p) - 1, \deg_y(p), \deg_z(p) + m^2 - 2)$.

(2) There are m^2 charts of the second type. Up to re-indexing the m^2 pairs (r, s) these are given by the coordinate change

$$(x_{i_1}, x_{i_2}, y_{rs} : 1 \leq r, s \leq m) = (x'_{i_1} y''_{mm}, x'_{i_2} y''_{mm}, y'_{rs} y''_{mm} : 1 \leq r, s \leq m)$$

except when $r = s = m$ where we set $y_{mm} = y''_{mm}$. It is convenient to set $y'_{mm} = 1$ and $z_w := y''_{mm}$. Then the new local equation is

$$x'_{i_1} x'_{i_2} \cdot L = t \cdot \det(y'_{rs} : 1 \leq r, s \leq m) \cdot (z_w^{m^2-2} \cdot R).$$

Note that the (m, m) entry of (y'_{rs}) is 1. By row and column operations we see that

$$\det(y'_{rs} : 1 \leq r, s \leq m) = \det(y'_{rs} - y'_{rm} y'_{ms} : 1 \leq r, s \leq m-1).$$

By setting $y''_{rs} := y'_{rs} - y'_{rm} y'_{ms}$ we have new local equations

$$x'_{i_1} x'_{i_2} L = t \cdot \det(y''_{rs} : 1 \leq r, s \leq m-1) \cdot (z_w^{m^2-2} \cdot R)$$

and the new degree is $(\deg_x(p), \deg_y(p) - 1, \deg_z(p) + m^2 - 2)$.

Outcome. After these blow ups we have a triple (Y, E, F) such that at non-resolved points the local equations are

$$\prod_{i \in I_p} x_i = t \cdot y \cdot \prod_{j \in J_p} z_j^{a_j} \quad \text{or} \quad \prod_{i \in I_p} x_i = t \cdot \prod_{j \in J_p} z_j^{a_j}. \quad (42.3)$$

(Note that we can not just declare that y is also a z -variable. The z_j are local equations of the divisors F_j while $(y=0)$ has no global meaning.)

43 (Resolving the monomial part). Following (42.3), the local equations are

$$\prod_{i \in I_p} x_i = t \cdot y^c \cdot \prod_{j \in J_p} z_j^{a_j} \quad \text{where } c \in \{0, 1\}.$$

We lower the degree of the z -monomial in 3 steps.

Step 1. Assume that there is a non-resolved point with $a_{j_1} \geq 2$.

The singular set of F_{j_1} is then

$$\bigcup_{(i, i')} (z_{j_1} = x_i = x_{i'} = 0)$$

where the union runs through all 2-element subsets $\{i, i'\} \subset I$. Pick an irreducible component of it, call it $W(i_1, i_2, j_1) := (z_{j_1} = x_{i_1} = x_{i_2} = 0)$.

Set $I_p^* := I_p \setminus \{i_1, i_2\}$, $J_p^* := J_p \setminus \{j_1\}$ and write the local equations as

$$x_{i_1} x_{i_2} \cdot L = t z_j^{a_j} \cdot R \quad \text{where } L := \prod_{i \in I_p^*} x_i \quad \text{and} \quad R := y^c \cdot \prod_{j \in J_p^*} z_j^{a_j}.$$

There are 3 local charts on the blow-up:

- (1) $(x_{i_1}, x_{i_2}, z_j) = (x'_{i_1}, x'_{i_2} x'_{i_1}, z'_j x'_{i_1})$ and, after setting $z_w := x'_{i_1}$ the new local equation is

$$x'_{i_2} \cdot L = t \cdot z_w^{a_j-2} z'_j{}^{a_j} \cdot R.$$

The new degree is $(\deg_x(p) - 1, \deg_y(p), \deg_z(p) + a_j - 2)$.

- (2) Same as above with the subscripts 1, 2 interchanged.

- (3) $(x_{i_1}, x_{i_2}, z_j) = (x'_{i_1} z'_j, x'_{i_2} z'_j, z'_j)$ with new local equation

$$x'_{i_1} x'_{i_2} \cdot L = t \cdot z'_j{}^{a_j-2} \cdot R.$$

The new degree is $(\deg_x(p), \deg_y(p), \deg_z(p) - 2)$.

Step 2. Assume that there is a non-resolved point with $a_{j_1} = a_{j_2} = 1$. The singular set of $F_{j_1} \cap F_{j_2}$ is then

$$\bigcup_{(i,i')} (z_{j_1} = z_{j_2} = x_i = x_{i'} = 0).$$

where the union runs through all 2-element subsets $\{i, i'\} \subset I$. Pick an irreducible component of it, call it $W(i_1, i_2, j_1, j_2) := (z_{j_1} = z_{j_2} = x_{i_1} = x_{i_2} = 0)$.

Set $I_p^* := I_p \setminus \{i_1, i_2\}$, $J_p^* := J_p \setminus \{j_1, j_2\}$ and we write the local equations as

$$x_{i_1} x_{i_2} \cdot L = t z_{j_1} z_{j_2} \cdot R \quad \text{where} \quad L := \prod_{i \in I_p^*} x_i \quad \text{and} \quad R := y^c \cdot \prod_{j \in J_p^*} z_j^{a_j}.$$

There are two types of local charts on the blow-up.

- (1) In the chart $(x_{i_1}, x_{i_2}, z_{j_1}, z_{j_2}) = (x'_{i_1}, x'_{i_2} x'_{i_1}, z'_{j_1} x'_{i_1}, z'_{j_2} x'_{i_1})$ the new local equation is

$$x'_{i_2} \cdot L = t \cdot z'_{j_1} z'_{j_2} \cdot R.$$

and the new degree is $(\deg_x(p) - 1, \deg_y(p), \deg_z(p))$. A similar chart is obtained by interchanging the subscripts i_1, i_2 .

- (2) In the chart $(x_{i_1}, x_{i_2}, z_{j_1}, z_{j_2}) = (x'_{i_1} z'_{j_1}, x'_{i_2} z'_{j_1}, z'_{j_1}, z'_{j_2} z'_{j_1})$. the new local equation is

$$x'_{i_1} x'_{i_2} \cdot L = t \cdot z'_{j_2} \cdot R.$$

The new degree is $(\deg_x(p), \deg_y(p), \deg_z(p) - 1)$.

A similar chart is obtained by interchanging the subscripts j_1, j_2 .

By repeated application of these two steps we are reduced to the case where $\deg_z(p) \leq 1$ at all non-resolved points.

Step 3. Assume that there is a non-resolved point with $\deg_y(p) = \deg_z(p) = 1$.

The singular set of Y is

$$\bigcup_{(i,i')} (y = z = x_i = x_{i'} = 0).$$

Pick an irreducible component of it, call it $W(i_1, i_2) := (y = z = x_{i_1} = x_{i_2} = 0)$. The blow up computation is the same as in Step 2.

As before we see that at each step the conditions (41.1–2) hold, hence $\mathcal{D}(E)$ is unchanged.

Outcome. After these blow-ups we have a triple (Y, E, F) such that at non-resolved points the local equations are

$$\prod_{i \in I_p} x_i = t \cdot y, \quad \prod_{i \in I_p} x_i = t \cdot z_1 \quad \text{or} \quad \prod_{i \in I_p} x_i = t. \quad (43.4)$$

As before, the y and z variables have different meaning, but we can rename z_1 as y . Thus we have only one non-resolved local form left: $\prod x_i = ty$.

44 (Resolving the multiplicity 2 part). Here we have a local equation $x_{i_1} \cdots x_{i_d} = ty$ where $d \geq 2$. We would like to blow up $(x_{i_1} = y = 0)$, but, as we noted, this subvariety is not globally defined. However, a rare occurrence helps us out. Usually the blow-up of a smooth subvariety determines its center uniquely. However, this is not the case for codimension 1 centers. Thus we could get a globally well defined blow-up even from centers that are not globally well defined.

Note that the inverse of $(x_{i_1} = y = 0)$ in the local Picard group of Y is $E_{i_1} = (x_{i_1} = t = 0)$, which is globally defined. Thus

$$\text{Proj}_Y \sum_{m \geq 0} \mathcal{O}_Y(mE_{i_1})$$

is well defined, and locally it is isomorphic to the blow-up $B_{(x_{i_1}=y=0)}Y$. (A priori, we would need to take the normalization of $B_{(x_{i_1}=y=0)}Y$, but it is actually normal.) Thus we have 2 local charts.

- (1) $(x_{i_1}, y) = (x'_{i_1}, y'x'_{i_1})$ and the new local equation is $(x_{i_2} \cdots x_{i_d} = ty')$. The new local degree is $(d-1, 1, 0)$.
- (2) $(x_{i_1}, y) = (x'_{i_1}y', y')$ and the new local equation is $(x'_{i_1} \cdot x_{i_2} \cdots x_{i_d} = t)$. The new local degree is $(d, 0, 0)$.

Outcome. After all these blow-ups we have a triple $(Y, \sum_{i \in I} E_i, \sum_{j \in J} a_j F_j)$ where $\sum_{i \in I} E_i$ is a simple normal crossing divisor and Y is smooth along $\sum_{i \in I} E_i$.

This completes the proof of Proposition 40. □

45 (Proof of Theorem 8). Assume that T is \mathbb{Q} -acyclic. Then, by (31) there is a simple normal crossing variety Z_T such that $H^i(Z_T, \mathcal{O}_{Z_T}) = 0$ for $i > 0$. Then [Kol11, Prop.9] shows that, for L sufficiently ample, the singularity $(0 \in X_T)$ constructed in (34) and (35) is rational. By (40) we conclude that $\mathcal{DR}(0 \in X_T) \cong \mathcal{D}(Z_T)$ is homotopy equivalent to T .

7. Cohen–Macaulay singularities

DEFINITION 46. *Cohen–Macaulay singularities* form the largest class where Serre duality holds. That is, if X is a projective variety of pure dimension n then X has Cohen–Macaulay singularities iff $H^i(X, L)$ is dual to $H^{n-i}(X, \omega_X \otimes L^{-1})$ for every line bundle L . A pleasant property is that if $D \subset X$ is Cartier divisor in a scheme then D is Cohen–Macaulay iff X is Cohen–Macaulay in a neighborhood of D . See [Har77, pp.184–186] or [KM98b, Sec.5.5] for details.

For local questions it is more convenient to use a characterization using local cohomology due to [Gro67, Sec.3.3]: X is Cohen–Macaulay iff $H_x^i(X, \mathcal{O}_X) = 0$ for every $x \in X$ and $i < \dim X$.

Every normal surface is Cohen–Macaulay, so the topology of the links of Cohen–Macaulay singularities starts to become interesting when $\dim X \geq 3$.

DEFINITION 47. Recall that a group G is called *perfect* if it has no nontrivial abelian quotients. Equivalently, if $G = [G, G]$ or if $H_1(G, \mathbb{Z}) = 0$.

We say that G is \mathbb{Q} -*perfect* if every abelian quotient is torsion. Equivalently, if $H_1(G, \mathbb{Q}) = 0$.

The following theorem describes the fundamental group of the link of Cohen–Macaulay singularities. Note, however, that the most natural part is the equivalence (48.1) \Leftrightarrow (48.5), relating the fundamental group of the link to the vanishing of $R^1 f_* \mathcal{O}_Y$ for a resolution $f : Y \rightarrow X$.

THEOREM 48. *For a finitely presented group G the following are equivalent.*

- (1) G is \mathbb{Q} -perfect (47).
- (2) G is the fundamental group of the link of an isolated Cohen–Macaulay singularity of dimension = 3.
- (3) G is the fundamental group of the link of an isolated Cohen–Macaulay singularity of dimension ≥ 3 .
- (4) G is the fundamental group of the link of a Cohen–Macaulay singularity whose singular set has codimension ≥ 3 .
- (5) G is the fundamental group of the link of a 1-rational singularity (52).

Proof. It is clear that (2) \Rightarrow (3) \Rightarrow (4) and (49) shows that (4) \Rightarrow (5). The implication (5) \Rightarrow (1) is proved in (51).

Let us prove (1) \Rightarrow (2). By (31) there is a simple normal crossing variety Z such that $\pi_1(Z) \cong G$. By a singular version of the Lefschetz hyperplane theorem (see, for instance, [GM88, Sec.II.1.2]), by taking general hyperplane sections we obtain a simple normal crossing surface S such that $\pi_1(S) \cong G$. Thus $H^1(S, \mathbb{Q}) = 0$ and by Hodge theory this implies that $H^1(S, \mathcal{O}_S) = 0$.

By (35) there is a 3-dimensional isolated singularity ($x \in X$) with a partial resolution $f : Y \rightarrow X$ whose exceptional divisor is $E \cong S$ and $R^1 f_* \mathcal{O}_Y \cong H^1(E, \mathcal{O}_E) = 0$. In this case the singularities of Y are the simplest possible: we have only ordinary nodes with equation $(x_1 x_2 = t y_{11})$. These are resolved in 1 step by blowing up ($x_1 = t = 0$) and they have no effect on our computations.

Thus X is Cohen–Macaulay by (50). \square

LEMMA 49. *Let X be a normal variety with Cohen–Macaulay singularities (S_3 would be sufficient) and $f : Y \rightarrow X$ a resolution of singularities. Then $\text{Supp } R^1 f_* \mathcal{O}_Y$ has pure codimension 2. Thus if $\text{Sing } X$ has codimension ≥ 3 then $R^1 f_* \mathcal{O}_Y = 0$.*

Proof. By localizing at a generic point of $\text{Supp } R^1 f_* \mathcal{O}_Y$ (or by taking a generic hyperplane section) we may assume that $\text{Supp } R^1 f_* \mathcal{O}_Y = \{x\}$ is a closed point. Set $E := f^{-1}(x)$. There is a Leray spectral sequence

$$H_x^i(X, R^j f_* \mathcal{O}_X) \Rightarrow H_E^{i+j}(Y, \mathcal{O}_Y). \quad (49.1)$$

By a straightforward duality (see, e.g. [Kol13b, 10.44]) $H_E^r(Y, \mathcal{O}_Y)$ is dual to the stalk of $R^{n-r}f_*\omega_Y$ which is zero for $r < n$ by [GR70]. Thus (49.1) gives an exact sequence

$$H_x^1(X, \mathcal{O}_X) \rightarrow H_E^1(Y, \mathcal{O}_Y) \rightarrow H_x^0(X, R^1f_*\mathcal{O}_X) \rightarrow H_x^2(X, \mathcal{O}_X).$$

If X is Cohen–Macaulay and $\dim X \geq 3$ then $H_x^1(X, \mathcal{O}_X) = H_x^2(X, \mathcal{O}_X) = 0$, thus

$$(R^1f_*\mathcal{O}_X)_x \cong H_x^0(X, R^1f_*\mathcal{O}_X) \cong H_E^1(Y, \mathcal{O}_Y) = 0. \quad \square$$

For isolated singularities, one has the following converse

LEMMA 50. *Let $(x \in X)$ be a normal, isolated singularity with a resolution $f : Y \rightarrow X$. Then X is Cohen–Macaulay iff $R^i f_* \mathcal{O}_Y = 0$ for $0 < i < n - 1$.*

Proof. The spectral sequence (49.1) implies that we have isomorphisms

$$R^i f_* \mathcal{O}_Y \cong H_x^i(X, \mathcal{O}_X) \quad \text{for } 0 < i < n - 1$$

and $H_x^1(X, \mathcal{O}_X) = 0$ since X is normal. \square

LEMMA 51. *Let X be a normal variety with 1-rational singularities (52) and $x \in X$ a point with link $L := L(x \in X)$. Then $H^1(L, \mathbb{Q}) = 0$.*

Proof. Let $f : Y \rightarrow X$ be a resolution such that $E := f^{-1}(x)$ is a simple normal crossing divisor. By [Ste83, 2.14] the natural maps $R^i f_* \mathcal{O}_Y \rightarrow H^i(E, \mathcal{O}_E)$ are surjective, thus $H^1(E, \mathcal{O}_E) = 0$ hence $H^1(E, \mathbb{Q}) = 0$ by Hodge theory.

Next we prove that $H^1(E, \mathbb{Q}) = H^1(L, \mathbb{Q})$. Let $x \in N_X \subset X$ be a neighborhood of x such that $\partial N_X = L$ and $N_Y := f^{-1}(N_X)$ the corresponding neighborhood of E with boundary $\partial N_Y := L_Y$. Since $L_Y \rightarrow L$ has connected fibers, $H^1(L, \mathbb{Q}) \hookrightarrow H^1(L_Y, \mathbb{Q})$ thus it is enough to prove that $H^1(L_Y, \mathbb{Q}) = 0$. The exact cohomology sequence of the pair (N_Y, L_Y) gives

$$0 = H^1(E, \mathbb{Q}) = H^1(N_Y, \mathbb{Q}) \rightarrow H^1(L_Y, \mathbb{Q}) \rightarrow H^2(N_Y, L_Y, \mathbb{Q}) \xrightarrow{\alpha} H^2(N_Y, \mathbb{Q})$$

By Poincaré duality $H^2(N_Y, L_Y, \mathbb{Q}) \cong H_{2n-2}(N_Y, \mathbb{Q})$. Since N_Y retracts to E we see that $H_{2n-2}(N_Y, \mathbb{Q})$ is freely generated by the classes of exceptional divisors $E = \cup_i E_i$. The map α sends $\sum m_i [E_i]$ to $c_1(\mathcal{O}_{N_Y}(\sum m_i E_i))$ and we need to show that the latter are nonzero. This follows from the Hodge index theorem. \square

8. Rational singularities

DEFINITION 52. A quasi projective variety X has *rational singularities* if for one (equivalently every) resolution of singularities $p : Y \rightarrow X$ and for every algebraic (or holomorphic) vector bundle F on X , the natural maps $H^i(X, F) \rightarrow H^i(Y, p^*F)$ are isomorphisms. Thus, for purposes of computing cohomology of vector bundles, X behaves like a smooth variety. Rational implies Cohen–Macaulay. See [KM98b, Sec.5.1] for details.

A more frequently used equivalent definition is the following. X has rational singularities iff the higher direct images $R^i f_* \mathcal{O}_Y$ are zero for $i > 0$ for one (equivalently every) resolution of singularities $p : Y \rightarrow X$.

We say that X has *1-rational singularities* if $R^1 f_* \mathcal{O}_Y = 0$ for one (equivalently every) resolution of singularities $p : Y \rightarrow X$.

53 (Proof of Theorem 8). Let $p : Y \rightarrow X$ be a resolution of singularities such that $E_x := p^{-1}(x)$ is a simple normal crossing divisor. As we noted in the proof of (51), $R^i f_* \mathcal{O}_Y \rightarrow H^i(E, \mathcal{O}_E)$ is surjective, thus $H^i(E, \mathcal{O}_E) = 0$ hence $H^i(\mathcal{DR}(x \in X), \mathbb{Q}) = 0$ by (32). Thus $\mathcal{DR}(x \in X)$ is \mathbb{Q} -acyclic.

Conversely, if T is \mathbb{Q} -acyclic then Theorem 7 constructs a singularity which is rational by (7.3). \square

Let L be the link of a rational singularity ($x \in X$). Since X is Cohen–Macaulay, we know that $\pi_1(L)$ is \mathbb{Q} -perfect (48). It is not known what else can one say about fundamental groups of links of rational singularities, but the fundamental group of the dual complex can be completely described.

DEFINITION 54. A group G is called *superperfect* if $H_1(G, \mathbb{Z}) = H_2(G, \mathbb{Z}) = 0$; see [Ber02]. We say that G is \mathbb{Q} -superperfect if $H_1(G, \mathbb{Q}) = H_2(G, \mathbb{Q}) = 0$. Note that every finite group is \mathbb{Q} -superperfect. Other examples are the infinite dihedral group or $\mathrm{SL}(2, \mathbb{Z})$.

COROLLARY 55. [KK11, Thm.42] *Let $(x \in X)$ be a rational singularity. Then $\pi_1(\mathcal{DR}(X))$ is \mathbb{Q} -superperfect. Conversely, for every finitely presented, \mathbb{Q} -superperfect group G there is a 6-dimensional rational singularity $(x \in X)$ such that*

$$\pi_1(\mathcal{DR}(X)) = \pi_1(\mathcal{R}(X)) = \pi_1(L(x \in X)) \cong G.$$

Proof. By a slight variant of the results of [Ker69, KM63], for every finitely presented, \mathbb{Q} -superperfect group G there is a \mathbb{Q} -acyclic, 5-dimensional manifold (with boundary) M whose fundamental group is isomorphic to G . Using this M in (8) we get a rational singularity $(x \in X)$ as desired.

Note that just applying the general construction would give 11 dimensional examples. See [KK11, Sec.7] on how to lower the dimension to 6.³ \square

9. Questions and problems

Questions about fundamental groups.

In principle, for any finitely presented group G one can follow the proof of [KK11] and construct links L such that $\pi_1(L) \cong G$. However, in almost all cases, the general methods lead to very complicated examples. It would be useful to start with some interesting groups and obtain examples that are understandable. For example, Higman’s group

$$H = \langle x_i : x_i[x_i, x_{i+1}], i \in \mathbb{Z}/4\mathbb{Z} \rangle$$

is perfect, infinite and contains no proper finite index subgroups [Hig51].

³A different construction giving 4 and 5 dimensional examples is in [Kol13a].

PROBLEM 56. Find an explicit link whose fundamental group is Higman's group. (It would be especially interesting to find examples that occur “naturally” in algebraic geometry.)

Note that our results give links with a given fundamental group but, as far as we can tell, these groups get killed in the larger quasi-projective varieties. (In particular, we do not answer the question [Ser77, p.19] whether Higman's group can be the fundamental group of a smooth variety.) This leads to the following.

QUESTION 57. Let G be a finitely presented group. Is there a quasi-projective variety X with an isolated singularity ($x \in X$) such that $\pi_1(L(x \in X)) \cong G$ and the natural map $\pi_1(L(x \in X)) \rightarrow \pi_1(X \setminus \{x\})$ is an injection?

As Kapovich pointed out, it is not known if every finitely presented group occurs as a subgroup of the fundamental group of a smooth projective or quasi-projective variety.

We saw in (55) that \mathbb{Q} -superperfect groups are exactly those that occur as $\pi_1(\mathcal{DR}(X))$ for rational singularities. Moreover, every \mathbb{Q} -superperfect group can be the fundamental group of a link of a rational singularity. However, there are rational singularities such that the fundamental group of their link is not \mathbb{Q} -superperfect. As an example, let S be a fake projective quadric whose universal cover is the 2-disc $\mathbb{D} \times \mathbb{D}$ (cf. [Bea96, Ex.X.13.4]). Let $C(S)$ be a cone over S with link $L(S)$. Then

$$H^2(L(S), \mathbb{Q}) \cong H^2(S, \mathbb{Q})/\mathbb{Q} \cong \mathbb{Q}$$

and the universal cover of L is an \mathbb{R} -bundle over $\mathbb{D} \times \mathbb{D}$ hence contractible. Thus

$$H^2(\pi_1(L(S)), \mathbb{Q}) \cong H^2(L(S), \mathbb{Q}) \cong \mathbb{Q},$$

so $\pi_1(L(S))$ is not \mathbb{Q} -superperfect. This leads us to the following, possibly very hard, question.

PROBLEM 58. Characterize the fundamental groups of links of rational singularities.

In this context it is worthwhile to mention the following.

CONJECTURE 59 (Carlson–Toledo). *The fundamental group of a smooth projective variety is not \mathbb{Q} -superperfect (unless it is finite).*

More generally, the original conjecture of Carlson and Toledo asserts that the image

$$\text{im}[H^2(\pi_1(X), \mathbb{Q}) \rightarrow H^2(X, \mathbb{Q})]$$

is nonzero and contains a (possibly degenerate) Kähler class, see [Kol95, 18.16]. For a partial solution see [Rez02].

Our examples show that for every finitely presented group G there is a *reducible* simple normal crossing surface S such that $\pi_1(S) \cong G$. By [Sim10], for every finitely presented group G there is a (very singular) *irreducible*

variety Z such that $\pi_1(Z) \cong G$. It is natural to hope to combine these results. [Kap12] proves that for every finitely presented group G there is an irreducible surface S with normal crossing and Whitney umbrella singularities (also called pinch points, given locally as $x^2 = y^2z$) such that $\pi_1(S) \cong G$.

PROBLEM 60. [Sim10] What can one say about the fundamental groups of irreducible surfaces with normal crossing singularities?

Although closely related, the next question should have a quite different answer.

PROBLEM 61. What can one say about the fundamental groups of normal, projective varieties or surfaces? Are these two classes of groups the same?

Many of the known restrictions on fundamental groups of smooth varieties also apply to normal varieties. For instance, the theory of Albanese varieties implies that the rank of $H_2(X, \mathbb{Q})$ is even for normal, projective varieties X . Another example is the following. By [Siu87] any surjection $\pi_1(X) \twoheadrightarrow \pi_1(C)$ to the fundamental group of a curve C of genus ≥ 2 factors as

$$\pi_1(X) \xrightarrow{g_*} \pi_1(C') \twoheadrightarrow \pi_1(C)$$

where $g : X \rightarrow C'$ is a morphism. (In general there is no morphism $C' \rightarrow C$.)

We claim that this also holds for normal varieties Y . Indeed, let $\pi : Y' \rightarrow Y$ be a resolution of singularities. Any surjection $\pi_1(Y) \twoheadrightarrow \pi_1(C)$ induces $\pi_1(Y') \twoheadrightarrow \pi_1(C)$, hence we get a morphism $g' : Y' \rightarrow C'$. Let $B \subset Y'$ be an irreducible curve that is contacted by π . Then $\pi_1(B) \rightarrow \pi_1(Y)$ is trivial and so is $\pi_1(B) \rightarrow \pi_1(C)$. If $g'|_B : B \rightarrow C'$ is not constant then the induced map $\pi_1(B) \rightarrow \pi_1(C')$ has finite index image. This is impossible since the composite $\pi_1(B) \rightarrow \pi_1(C') \rightarrow \pi_1(C)$ is trivial. Thus g' descends to $g : Y \rightarrow C'$.

For further such results see [Gro89, GL91, Cat91, Cat96].

Algebraically one can think of the link as the punctured spectrum of the Henselisation (or completion) of the local ring of $x \in X$. Although one can not choose a base point, it should be possible to define an algebraic fundamental group. All the examples in Theorem 3 can be realized on varieties defined over \mathbb{Q} . Thus they should have an algebraic fundamental group $\pi_1^{\text{alg}}(L(0 \in X_{\mathbb{Q}}))$ which is an extension of the profinite completion of $\pi_1(L(0 \in X))$ and of the Galois group $\text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$.

PROBLEM 62. Define and describe the possible groups $\pi_1^{\text{alg}}(L(0 \in X_{\mathbb{Q}}))$.

Questions about the topology of links.

We saw that the fundamental groups of links can be quite different from fundamental groups of quasi-projective varieties. However, our results say very little about the cohomology or other topological properties of links. It

turns out that links have numerous restrictive topological properties. I thank J. Shaneson and L. Maxim for bringing many of these to my attention.

63 (Which manifolds can be links?). Let M be a differentiable manifold that is diffeomorphic to the link L of an isolated complex singularity of dimension n . Then M satisfies the following.

63.1. $\dim_{\mathbb{R}} M = 2n - 1$ is odd and M is orientable. Resolution of singularities shows that M is cobordant to 0.

63.2. The decomposition $T_X|_L \cong T_L + N_{L,X}$ shows that T_M is stably complex. In particular, its odd integral Stiefel–Whitney classes are zero [Mas61]. (More generally, this holds for orientable real hypersurfaces in complex manifolds.)

63.3. The cohomology groups $H^i(L, \mathbb{Q})$ carry a natural mixed Hodge structure; see [PS08, Sec.6.3] for a detailed treatment and references. Using these, [DH88] proves that the cup product $H^i(L, \mathbb{Q}) \times H^j(L, \mathbb{Q}) \rightarrow H^{i+j}(L, \mathbb{Q})$ is zero if $i, j < n$ and $i + j \geq n$. In particular, the torus \mathbb{T}^{2n-1} can not be a link. If X is a smooth projective variety then $X \times \mathbb{S}^1$ can not be a link. Further results along this direction are in [PP08].

63.4. By [CS91, p.548], the components of the Todd–Hirzebruch L-genus of M vanish above the middle dimension. More generally, the purity of the Chern classes and weight considerations as in (63.3) show that the $c_i(T_X|_L)$ are torsion above the middle dimension. Thus all Pontryagin classes of L are torsion above the middle dimension. See also [CMS08a, CMS08b] for further results on the topology of singular algebraic varieties which give restrictions on links as special cases.

There is no reason to believe that this list is complete and it would be useful to construct many different links to get some idea of what other restrictions may hold.

Let $(0 \in X) \subset (0 \in \mathbb{C}^N)$ be an isolated singularity of dimension n and $L = X \cap \mathbb{S}^{2N-1}(\epsilon)$ its link. If $X_0 := X$ is smoothable in a family $\{X_t \subset \mathbb{C}^N\}$ then L bounds a Stein manifold $U_t := X_t \cap \mathbb{B}^{2N}(\epsilon)$ and U_t is homotopic to an n -dimensional compact simplicial complex. This imposes strong restrictions on the topology of smoothable links; some of these were used in [PP08]. Interestingly, these restrictions use the integral structure of the cohomology groups. This leads to the following intriguing possibility.

QUESTION 64. Let L be a link of dimension $2n - 1$. Does L bound a \mathbb{Q} -homology manifold U (of dimension $2n$) that is \mathbb{Q} -homotopic to an n -dimensional, finite simplicial complex?

There is very little evidence to support the above speculation but it is consistent with known restrictions on the topology of links and it would explain many of them. On the other hand, I was unable to find such U even in some simple cases. For instance, if $(0 \in X)$ is a cone over an Abelian

variety (or a product of curves of genus ≥ 2) of dimension ≥ 2 then algebraic deformations of X do not produce such a U .

Restricting to the cohomology rings, here are two simple questions.

QUESTION 65 (Cohomology of links). Is the sequence of Betti numbers of a complex link arbitrary? Can one describe the possible algebras $H^*(L, \mathbb{Q})$?

QUESTION 66 (Cohomology of links of weighted cones). We saw in (11) that the first Betti number of the link of a weighted cone (of dimension > 1) is even. One can ask if this is the only restriction on the Betti numbers of a complex link of a weighted cone.

Philosophically, one of the main results on the topology of smooth projective varieties, proved in [DGMS75, Sul77], says that for simply connected varieties the integral cohomology ring and the Pontryagin classes determine the differentiable structure up to finite ambiguity. It is natural to ask what happens for links.

QUESTION 67. To what extent is the diffeomorphism type of a simply connected link L determined by the cohomology ring $H^*(L, \mathbb{Z})$ plus some characteristic classes?

A positive answer to (67) would imply that general links are indeed very similar to weighted homogeneous links and to projective varieties.

Questions about $\mathcal{DR}(0 \in X)$.

The preprint version contained several questions about dual complexes of dlt pairs; these are corrected and solved in [dFKX12].

Embeddings of simple normal crossing varieties.

In many contexts it has been a difficulty that not every variety with simple normal crossing singularities can be realized as a hypersurface in a smooth variety. See for instance [Fuj09, BM11, BP11, Kol13b] for such examples and for various partial solutions.

As we discussed in (19), recent examples of [Fuj12a, Fuj12b] show that the answer to the following may be quite complicated.

QUESTION 68. Which proper, complex, simple normal crossing spaces can be realized as hypersurfaces in a complex manifold?

QUESTION 69. Which projective simple normal crossing varieties can be realized as hypersurfaces in a smooth projective variety?

Note that, in principle it could happen that there is a projective simple normal crossing variety that can be realized as a hypersurface in a complex manifold but not in a smooth projective variety.

Let Y be a smooth variety and $D \subset Y$ a compact divisor. Let $D \subset N \subset Y$ be a regular neighborhood with smooth boundary ∂N . If D is the exceptional divisor of a resolution of an isolated singularity $x \in X$

then ∂N is homeomorphic to the link $L(x \in X)$. It is clear that D and $c_1(N_{D,X}) \in H^2(D, \mathbb{Z})$ determine the boundary ∂N , but I found it very hard to compute concrete examples.

PROBLEM 70. Find an effective method to compute the cohomology or the fundamental group of ∂N , at least when D is a simple normal crossing divisor.

References

- [ABC⁺96] J. Amorós, M. Burger, K. Corlette, D. Kotschick, and D. Toledo, *Fundamental groups of compact Kähler manifolds*, Mathematical Surveys and Monographs, vol. 44, American Mathematical Society, Providence, RI, 1996. MR 1379330 (97d:32037)
- [ABW09] D. Arapura, P. Bakhtary, and J. Włodarczyk, *The combinatorial part of the cohomology of a singular variety*, ArXiv:0902.4234, 2009.
- [ABW11] ———, *Weights on cohomology, invariants of singularities, and dual complexes*, ArXiv e-prints (2011).
- [Art70] Michael Artin, *Algebraization of formal moduli. II. Existence of modifications*, Ann. of Math. (2) **91** (1970), 88–135. MR 0260747 (41 #5370)
- [Bar65] D. Barden, *Simply connected five-manifolds*, Ann. of Math. (2) **82** (1965), 365–385. MR MR0184241 (32 #1714)
- [Bea96] Arnaud Beauville, *Complex algebraic surfaces*, second ed., London Mathematical Society Student Texts, vol. 34, Cambridge University Press, Cambridge, 1996, Translated from the 1978 French original by R. Barlow, with assistance from N. I. Shepherd-Barron and M. Reid. MR MR97e:14045
- [Ber02] A. J. Berrick, *A topologist’s view of perfect and acyclic groups*, Invitations to geometry and topology, Oxf. Grad. Texts Math., vol. 7, Oxford Univ. Press, Oxford, 2002, pp. 1–28. MR 1967745 (2004c:20001)
- [BG00] Charles P. Boyer and Krzysztof Galicki, *On Sasakian-Einstein geometry*, Internat. J. Math. **11** (2000), no. 7, 873–909. MR 2001k:53081
- [BG08] ———, *Sasakian geometry*, Oxford Mathematical Monographs, Oxford University Press, Oxford, 2008. MR 2382957 (2009c:53058)
- [BGK05] Charles P. Boyer, Krzysztof Galicki, and János Kollár, *Einstein metrics on spheres*, Ann. of Math. (2) **162** (2005), no. 1, 557–580. MR MR2178969 (2006j:53058)
- [BGKT05] Charles P. Boyer, Krzysztof Galicki, János Kollár, and Evan Thomas, *Einstein metrics on exotic spheres in dimensions 7, 11, and 15*, Experiment. Math. **14** (2005), no. 1, 59–64. MR 2146519 (2006a:53042)
- [BM11] Edward Bierstone and Pierre D. Milman, *Resolution except for minimal singularities I*, arXiv.org:1107.5595, 2011.
- [BP11] Edward Bierstone and Franklin V. Pacheco, *Resolution of singularities of pairs preserving semi-simple normal crossings*, arXiv.org:1109.3205, 2011.
- [Bri66] Egbert Brieskorn, *Beispiele zur Differentialtopologie von Singularitäten*, Invent. Math. **2** (1966), 1–14. MR 34 #6788
- [Cai61] Stewart S. Cairns, *A simple triangulation method for smooth manifolds*, Bull. Amer. Math. Soc. **67** (1961), 389–390. MR 0149491 (26 #6978)
- [Cat91] Fabrizio Catanese, *Moduli and classification of irregular Kaehler manifolds (and algebraic varieties) with Albanese general type fibrations*, Invent. Math. **104** (1991), no. 2, 263–289. MR 1098610 (92f:32049)
- [Cat96] ———, *Fundamental groups with few relations*, Higher-dimensional complex varieties (Trento, 1994), de Gruyter, Berlin, 1996, pp. 163–165. MR 1463177 (98i:32047)

- [CMS08a] Sylvain E. Cappell, Laurentiu G. Maxim, and Julius L. Shaneson, *Euler characteristics of algebraic varieties*, Comm. Pure Appl. Math. **61** (2008), no. 3, 409–421. MR 2376847 (2009f:14038)
- [CMS08b] ———, *Hodge genera of algebraic varieties. I*, Comm. Pure Appl. Math. **61** (2008), no. 3, 422–449. MR 2376848 (2009f:14039)
- [Coh73] Marshall M. Cohen, *A course in simple-homotopy theory*, Springer-Verlag, New York, 1973, Graduate Texts in Mathematics, Vol. 10. MR 0362320 (50 #14762)
- [Cor92] Jon Michael Corson, *Complexes of groups*, Proc. London Math. Soc. (3) **65** (1992), no. 1, 199–224. MR 1162493 (93h:57003)
- [CS91] Sylvain E. Cappell and Julius L. Shaneson, *Stratifiable maps and topological invariants*, J. Amer. Math. Soc. **4** (1991), no. 3, 521–551. MR 1102578 (92d:57024)
- [CS08] Kevin Corlette and Carlos Simpson, *On the classification of rank-two representations of quasiprojective fundamental groups*, Compos. Math. **144** (2008), no. 5, 1271–1331. MR 2457528 (2010i:14006)
- [Dem88] Michel Demazure, *Anneaux gradués normaux*, Introduction à la théorie des singularités, II, Travaux en Cours, vol. 37, Hermann, Paris, 1988, pp. 35–68. MR 91k:14004
- [dFKX12] Tommaso de Fernex, János Kollár, and Chenyang Xu, *The dual complex of singularities*, ArXiv e-prints (2012).
- [DGMS75] Pierre Deligne, Phillip Griffiths, John Morgan, and Dennis Sullivan, *Real homotopy theory of Kähler manifolds*, Invent. Math. **29** (1975), no. 3, 245–274. MR MR0382702 (52 #3584)
- [DH88] Alan H. Durfee and Richard M. Hain, *Mixed Hodge structures on the homotopy of links*, Math. Ann. **280** (1988), no. 1, 69–83. MR 928298 (89c:14012)
- [Dol83] Igor V. Dolgachev, *On the link space of a Gorenstein quasihomogeneous surface singularity*, Math. Ann. **265** (1983), no. 4, 529–540. MR 721886 (85k:32024)
- [DPS09] Alexandru Dimca, Stefan Papadima, and Alexander I. Suciu, *Topology and geometry of cohomology jump loci*, Duke Math. J. **148** (2009), no. 3, 405–457. MR 2527322 (2011b:14047)
- [FM83] Robert Friedman and David R. Morrison (eds.), *The birational geometry of degenerations*, Progr. Math., vol. 29, Birkhäuser Boston, Mass., 1983. MR 690262 (84g:14032)
- [Fuj09] Osamu Fujino, *Introduction to the log minimal model program for log canonical pairs*, arXiv.org:0907.1506, 2009.
- [Fuj12a] Kento Fujita, *Simple normal crossing Fano varieties and log Fano manifolds*, ArXiv e-prints (2012).
- [Fuj12b] ———, *The Mukai conjecture for log Fano manifolds*, ArXiv e-prints (2012).
- [FZ03] Hubert Flenner and Mikhail Zaidenberg, *Normal affine surfaces with \mathbb{C}^* -actions*, Osaka J. Math. **40** (2003), no. 4, 981–1009. MR 2 020 670
- [GL91] Mark Green and Robert Lazarsfeld, *Higher obstructions to deforming cohomology groups of line bundles*, J. Amer. Math. Soc. **4** (1991), no. 1, 87–103. MR MR1076513 (92i:32021)
- [GM88] Mark Goresky and Robert MacPherson, *Stratified Morse theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), vol. 14, Springer-Verlag, Berlin, 1988. MR 932724 (90d:57039)
- [Gor80] Gerald Leonard Gordon, *On a simplicial complex associated to the monodromy*, Trans. Amer. Math. Soc. **261** (1980), no. 1, 93–101. MR 576865 (81j:32017)
- [GR70] Hans Grauert and Oswald Riemenschneider, *Verschwindungssätze für analytische Kohomologiegruppen auf komplexen Räumen*, Invent. Math. **11** (1970), 263–292. MR MR0302938 (46 #2081)
- [Gro67] Alexander Grothendieck, *Local cohomology*, Lecture Notes in Mathematics, Vol. 41, Springer-Verlag, Berlin, 1967. MR 0224620 (37 #219)

- [Gro68] ———, *Cohomologie locale des faisceaux cohérents et théorèmes de Lefschetz locaux et globaux (SGA 2)*, North-Holland Publishing Co., Amsterdam, 1968, Augmenté d'un exposé par Michèle Raynaud, Séminaire de Géométrie Algébrique du Bois-Marie, 1962, Advanced Studies in Pure Mathematics, Vol. 2. MR 0476737 (57 #16294)
- [Gro89] Michel Gromov, *Sur le groupe fondamental d'une variété kähleriennne*, C. R. Acad. Sci. Paris Sér. I Math. **308** (1989), no. 3, 67–70. MR 983460 (90i:53090)
- [GS75] Phillip Griffiths and Wilfried Schmid, *Recent developments in Hodge theory: a discussion of techniques and results*, Discrete subgroups of Lie groups and applicatons to moduli (Internat. Colloq., Bombay, 1973), Oxford Univ. Press, Bombay, 1975, pp. 31–127. MR 0419850 (54 #7868)
- [Har77] Robin Hartshorne, *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157 (57 #3116)
- [Har95] Joe Harris, *Algebraic geometry*, Graduate Texts in Mathematics, vol. 133, Springer-Verlag, New York, 1995, A first course, Corrected reprint of the 1992 original. MR MR1416564 (97e:14001)
- [Hat02] Allen Hatcher, *Algebraic topology*, Cambridge University Press, Cambridge, 2002. MR 1867354 (2002k:55001)
- [Hig51] Graham Higman, *A finitely generated infinite simple group*, J. London Math. Soc. **26** (1951), 61–64. MR 0038348 (12,390c)
- [Hir62] Morris W. Hirsch, *Smooth regular neighborhoods*, Ann. of Math. (2) **76** (1962), 524–530. MR 0149492 (26 #6979)
- [Ish85] Shihoko Ishii, *On isolated Gorenstein singularities*, Math. Ann. **270** (1985), no. 4, 541–554. MR MR776171 (86j:32024)
- [Kap12] M. Kapovich, *Dirichlet fundamental domains and complex-projective varieties*, ArXiv e-prints (2012).
- [Ker69] Michel A. Kervaire, *Smooth homology spheres and their fundamental groups*, Trans. Amer. Math. Soc. **144** (1969), 67–72. MR 0253347 (40 #6562)
- [KK11] Michael Kapovich and János Kollár, *Fundamental groups of links of isolated singularities*, Journal AMS (to appear) ArXiv e-prints (2011).
- [KM63] Michel A. Kervaire and John W. Milnor, *Groups of homotopy spheres. I*, Ann. of Math. (2) **77** (1963), 504–537. MR 0148075 (26 #5584)
- [KM98a] Michael Kapovich and John J. Millson, *On representation varieties of Artin groups, projective arrangements and the fundamental groups of smooth complex algebraic varieties*, Inst. Hautes Études Sci. Publ. Math. (1998), no. 88, 5–95 (1999). MR 1733326 (2001d:14024)
- [KM98b] János Kollár and Shigefumi Mori, *Birational geometry of algebraic varieties*, Cambridge Tracts in Mathematics, vol. 134, Cambridge University Press, Cambridge, 1998, With the collaboration of C. H. Clemens and A. Corti, Translated from the 1998 Japanese original. MR 1658959 (2000b:14018)
- [Kob63] Shoshichi Kobayashi, *Topology of positively pinched Kähler manifolds*, Tôhoku Math. J. (2) **15** (1963), 121–139. MR 0154235 (27 #4185)
- [Kol95] János Kollár, *Shafarevich maps and automorphic forms*, M. B. Porter Lectures, Princeton University Press, Princeton, NJ, 1995. MR 1341589 (96i:14016)
- [Kol05] ———, *Einstein metrics on five-dimensional Seifert bundles*, J. Geom. Anal. **15** (2005), no. 3, 445–476. MR MR2190241 (2007c:53056)
- [Kol06] ———, *Circle actions on simply connected 5-manifolds*, Topology **45** (2006), no. 3, 643–671. MR 2218760 (2006m:57044)
- [Kol07a] ———, *Einstein metrics on connected sums of $S^2 \times S^3$* , J. Differential Geom. **75** (2007), no. 2, 259–272. MR MR2286822 (2007k:53061)
- [Kol07b] ———, *Lectures on resolution of singularities*, Annals of Mathematics Studies, vol. 166, Princeton University Press, Princeton, NJ, 2007. MR 2289519

- [Kol09] ———, *Positive Sasakian structures on 5-manifolds*, Riemannian topology and geometric structures on manifolds, Progr. Math., vol. 271, Birkhäuser Boston, Boston, MA, 2009, pp. 93–117. MR 2494170 (2010i:53077)
- [Kol11] ———, *New examples of terminal and log canonical singularities*, arXiv:1107.2864, 2011.
- [Kol12] ———, *Quotients by finite equivalence relations*, Current developments in algebraic geometry, Math. Sci. Res. Inst. Publ., vol. 59, Cambridge Univ. Press, Cambridge, 2012, With an appendix by Claudiu Raicu, pp. 227–256. MR 2931872
- [Kol13a] ———, *Simple normal crossing varieties with prescribed dual complex*, ArXiv e-prints (2013).
- [Kol13b] ———, *Singularities of the minimal model program*, Cambridge University Press, Cambridge, 2013, With the collaboration of S. Kovács.
- [Kul77] Vik. S. Kulikov, *Degenerations of K3 surfaces and Enriques surfaces*, Izv. Akad. Nauk SSSR Ser. Mat. **41** (1977), no. 5, 1008–1042, 1199. MR 0506296 (58 #22087b)
- [Mas61] W. S. Massey, *Obstructions to the existence of almost complex structures*, Bull. Amer. Math. Soc. **67** (1961), 559–564. MR 0133137 (24 #A2971)
- [Mor78] John W. Morgan, *The algebraic topology of smooth algebraic varieties*, Inst. Hautes Études Sci. Publ. Math. (1978), no. 48, 137–204. MR 516917 (80e:55020)
- [Mum61] David Mumford, *The topology of normal singularities of an algebraic surface and a criterion for simplicity*, Inst. Hautes Études Sci. Publ. Math. (1961), no. 9, 5–22. MR 0153682 (27 #3643)
- [Neu81] Walter D. Neumann, *A calculus for plumbing applied to the topology of complex surface singularities and degenerating complex curves*, Trans. Amer. Math. Soc. **268** (1981), no. 2, 299–344. MR 632532 (84a:32015)
- [OW75] Peter Orlik and Philip Wagreich, *Seifert n-manifolds*, Invent. Math. **28** (1975), 137–159. MR 50 #13596
- [Pay09] Sam Payne, *Lecture at MSRI*, <http://www.msri.org/web/msri/online-videos/-/video/showVideo/3674>, 2009.
- [Pay11] Sam Payne, *Boundary complexes and weight filtrations*, ArXiv e-prints (2011).
- [Per77] Ulf Persson, *On degenerations of algebraic surfaces*, Mem. Amer. Math. Soc. **11** (1977), no. 189, xv+144. MR 0466149 (57 #6030)
- [Pin77] H. Pinkham, *Normal surface singularities with C^* action*, Math. Ann. **227** (1977), no. 2, 183–193. MR 55 #5623
- [PP08] Patrick Popescu-Pampu, *On the cohomology rings of holomorphically fillable manifolds*, Singularities II, Contemp. Math., vol. 475, Amer. Math. Soc., Providence, RI, 2008, pp. 169–188. MR 2454366 (2010h:32039)
- [PS08] Chris A. M. Peters and Joseph H. M. Steenbrink, *Mixed Hodge structures*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge., vol. 52, Springer-Verlag, Berlin, 2008. MR MR2393625
- [Rez02] Alexander Reznikov, *The structure of Kähler groups. I. Second cohomology*, Motives, polylogarithms and Hodge theory, Part II (Irvine, CA, 1998), Int. Press Lect. Ser., vol. 3, Int. Press, Somerville, MA, 2002, pp. 717–730. MR 1978716 (2004c:32042)
- [Sco83] Peter Scott, *The geometries of 3-manifolds*, Bull. London Math. Soc. **15** (1983), no. 5, 401–487. MR 84m:57009
- [Sei32] Herbert Seifert, *Topologie dreidimensionaler gefaserte Räume*, Acta Math. **60** (1932), 148–238.
- [Ser77] Jean-Pierre Serre, *Arbres, amalgames, SL_2* , Société Mathématique de France, Paris, 1977, Avec un sommaire anglais, Rédigé avec la collaboration de Hyman Bass, Astérisque, No. 46. MR 0476875 (57 #16426)

- [Sim10] Carlos Simpson, *Local systems on proper algebraic V-manifolds*, arXiv1010.3363, 2010.
- [Siu87] Yum Tong Siu, *Strong rigidity for Kähler manifolds and the construction of bounded holomorphic functions*, Discrete groups in geometry and analysis (New Haven, Conn., 1984), Progr. Math., vol. 67, Birkhäuser Boston, Boston, MA, 1987, pp. 124–151. MR 900825 (89i:32044)
- [Sma62] Stephen Smale, *On the structure of 5-manifolds*, Ann. of Math. (2) **75** (1962), 38–46. MR 25 #4544
- [Ste83] J. H. M. Steenbrink, *Mixed Hodge structures associated with isolated singularities*, Singularities, Part 2 (Arcata, Calif., 1981), Proc. Sympos. Pure Math., vol. 40, Amer. Math. Soc., Providence, RI, 1983, pp. 513–536. MR 713277 (85d:32044)
- [Ste08] D. A. Stepanov, *A note on resolution of rational and hypersurface singularities*, Proc. Amer. Math. Soc. **136** (2008), no. 8, 2647–2654. MR 2399025 (2009g:32060)
- [Sul77] Dennis Sullivan, *Infinitesimal computations in topology*, Inst. Hautes Études Sci. Publ. Math. (1977), no. 47, 269–331 (1978). MR 0646078 (58 #31119)
- [Thu07] Amaury Thuillier, *Géométrie toroïdale et géométrie analytique non archimédienne. Application au type d’homotopie de certains schémas formels*, Manuscripta Math. **123** (2007), no. 4, 381–451. MR 2320738 (2008g:14038)

PRINCETON UNIVERSITY, PRINCETON, NJ 08544-1000, USA

E-mail address: kollar@math.princeton.edu

This page intentionally left blank

Calabi energies of extremal toric surfaces

Claude LeBrun

ABSTRACT. We derive a formula for the L^2 norm of the scalar curvature of any extremal Kähler metric on a compact toric manifold, stated purely in terms of the geometry of the corresponding moment polytope. The main interest of this formula pertains to the case of complex dimension 2, where it plays a key role in construction of Bach-flat metrics on appropriate 4-manifolds.

1. Introduction

In an audacious attempt to endow complex algebraic varieties with canonical Riemannian metrics, Eugenio Calabi [11] initiated a systematic study of the squared L^2 -norm

$$(1.1) \quad \mathcal{C}(g) = \int_M s^2 \, d\mu$$

of the scalar curvature, considered as a functional on the space of Kähler metrics g on a given compact complex manifold (M, J) ; here s and $d\mu$ of course denote the scalar curvature and Riemannian volume form of the given metric g . Given a Kähler class $\Omega \in H^{1,1}(M, \mathbb{R}) \subset H^2(M, \mathbb{R})$, his aim was to minimize the functional $\mathcal{C}(g)$ among all Kähler metrics $g = \omega(\cdot, J\cdot)$ with Kähler class $[\omega] = \Omega$. Calabi showed that the Euler-Lagrange equation for this variational problem is equivalent to requiring that $\nabla^{1,0}s$ be a holomorphic vector field, and he introduced the terminology *extremal Kähler metrics* for the solutions of this equation. It was later shown [13] that any extremal Kähler metric on a compact complex manifold actually minimizes the Calabi energy (1.1) in its Kähler class. Moreover, when such a minimizer exists, it is actually unique in its Kähler class, modulo automorphisms of the complex manifold [14, 20, 43]. Our knowledge of existence remains imperfect, but considerable progress [2, 16, 21] has recently been made in the toric case that is focus of the present paper. However, a relatively elementary argument [39] shows that the set of Kähler classes represented

Supported in part by NSF grants DMS-0905159 and DMS-1205953.

by extremal Kähler metrics on a compact complex manifold (M, J) is necessarily open in $H^{1,1}(M, \mathbb{R})$.

Rather than minimizing the squared L^2 -norm of the scalar curvature, as in (1.1), one might be tempted to instead minimize the squared L^2 -norm of, say, the Riemann curvature tensor or the Ricci tensor. However, Calabi also observed [11] that, after appropriate normalization, such functionals only differ from (1.1) by a constant depending on the Kähler class. In this respect, real dimension four occupies a privileged position; not only does (1.1) become scale invariant in this dimension, but the relevant constants only depend on the topology of M^4 , and so are independent of the Kähler class in question. For example, the Riemann curvature \mathcal{R} and the Ricci tensor r satisfy

$$\begin{aligned} \int_M |\mathcal{R}|^2 d\mu &= -8\pi^2(\chi + 3\tau)(M) + \frac{1}{4}\mathcal{C}(g) \\ \int_M |r|^2 d\mu &= -8\pi^2(2\chi + 3\tau)(M) + \frac{1}{2}\mathcal{C}(g) \end{aligned}$$

for any compact Kähler manifold (M, g, J) of complex dimension 2, where $\chi(M)$ and $\tau(M)$ are respectively the Euler characteristic and signature of the compact oriented 4-manifold M . Similarly, the Weyl curvature W , which is the conformally invariant part of the Riemann tensor \mathcal{R} , satisfies

$$(1.2) \quad \int_M |W|^2 d\mu = -12\pi^2\tau(M) + \frac{1}{12}\mathcal{C}(g)$$

for any compact Kähler surface (M, g, J) . Thus, if a Kähler metric g on M^4 is a critical point of any of these Riemannian functionals, considered as a function on the bigger space of all Riemannian metrics on M , it must, in particular, be an *extremal* Kähler metric. In connection with (1.2), this observation has interesting consequences, some of which will be touched on in this article.

The primary goal of this article is to calculate the Calabi energy of any extremal Kähler metric on any *toric surface* — that is, on any simply connected compact complex manifold of complex dimension two which carries a compatible effective action of the 2-torus $T^2 = S^1 \times S^1$. Any Kähler class on a toric surface is represented by a T^2 -invariant Kähler metric, and, relative to such a metric, the action is generated by two periodic Hamiltonian vector fields. This pair of Hamiltonians gives us an \mathbb{R}^2 -valued moment map, under which the image of our complex surface is a convex polygon $P \subset \mathbb{R}^2$. Moreover, modulo translations and $SL(2, \mathbb{Z})$ transformations, the moment polygon P only depends on the given the Kähler class. Euclidean area measure on the interior of P then allows us to define a barycenter for P and a moment-of-inertia matrix Π of P relative to this barycenter. The edges of P have rational slope, and are therefore endowed with preferred rescalings $d\lambda$ of 1-dimensional Lebesgue measure, chosen so that intervals of unit length correspond to separation vectors which are indivisible elements

of the integer lattice \mathbb{Z}^2 . This allows us to also define a barycenter of the perimeter of P , and hence also a vector $\vec{\mathfrak{D}} \in \mathbb{R}^2$ connecting the barycenter of the interior to the barycenter of the perimeter. Combining these ingredients, we then obtain a convenient formula for the Calabi energy of any extremal toric surface:

THEOREM A. *Let (M, J, Ω) be a toric surface with fixed Kähler class, and let P be the associated moment polygon. Then any Kähler metric g with Kähler form $\omega \in \Omega$ has scalar curvature s satisfying*

$$\frac{1}{32\pi^2} \int_M s^2 d\mu_g \geq \frac{|\partial P|^2}{2} \left(\frac{1}{|P|} + \vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}} \right)$$

with equality iff g is an extremal Kähler metric. Here $|P|$ denotes the area of the interior of P , $|\partial P|$ is the λ -length of its boundary, Π is the moment-of-inertia matrix of P , and $\vec{\mathfrak{D}}$ is the vector joining the barycenter P to the barycenter of ∂P .

We give two proofs of this result. Our first proof, which is specifically adapted to complex dimension 2, can be found in §5 below. Then, in §6, we prove a generalization, Theorem B, which holds for toric manifolds of arbitrary complex dimension. However, both proofs crucially depend on a detailed understanding of both the Futaki invariant and toric manifolds. We have therefore found it useful to preface our main calculations with a careful exploration of the underpinnings of these ideas. The article then concludes with a discussion of examples that illustrate our current knowledge of Bach-flat Kähler metrics.

2. The Futaki Invariant

If (M^{2m}, J) is a compact complex m -manifold of Kähler type, and if

$$\mathfrak{h} = H^0(M, \mathcal{O}(T^{1,0}M))$$

is the associated Lie algebra of holomorphic fields on M , the *Futaki invariant* assigns an element $\mathfrak{F}(\Omega)$ of the Lie coalgebra \mathfrak{h}^* to every Kähler class Ω on (M, J) . To construct this element, let g be a Kähler metric, with Kähler class $[\omega] = \Omega$, scalar curvature s , Green's operator \mathcal{G} , and volume form $d\mu$. We then define the Futaki invariant

$$\mathfrak{F}(\Omega) : H^0(M, \mathcal{O}(T^{1,0}M)) \longrightarrow \mathbb{C}$$

to be the linear functional

$$\Xi \longmapsto -2 \int_M \Xi(\mathcal{G}s) d\mu.$$

It is a remarkable fact, due to Futaki [24], Bando [6], and Calabi [12], that $\mathfrak{F}(\Omega)$ only depends on the Kähler class Ω , and not on the particular metric g chosen to represent it.

We will now assume henceforth that $b_1(M) = 0$. Since (M, J) is of Kähler type, the Hodge decomposition then tells us that $H^{0,1}(M) = 0$, and

it therefore follows [12, 40] that every holomorphic vector field Ξ on M can be written as $\nabla^{1,0}f$ for some smooth function $f = f_\Xi$, called a holomorphy potential. This allows us to re-express the Futaki invariant as

$$(2.1) \quad \mathfrak{F}(\Xi, \Omega) := [\mathfrak{F}(\Omega)](\Xi) = - \int_M (s - \bar{s}) f_\Xi d\mu$$

where \bar{s} denotes the average value of the scalar curvature, which can be computed by the topological formula

$$\bar{s} = 4\pi m \frac{c_1 \cdot \Omega^{m-1}}{\Omega^m}.$$

Of course, the negative sign appearing in (2.1) is strictly a matter of convention, and is used here primarily to ensure consistency with [40]. Also note that the \bar{s} term in (2.1) could be dropped if one required that the holonomy potential f_Ξ be normalized to have integral zero; however, we will find it useful to avoid systematically imposing such a normalization.

Let \mathbf{H} now denote the identity component of the automorphism group of (M, J) , so that \mathfrak{h} is its Lie algebra. Because the assumption that $b_1(M) = 0$ implies that \mathbf{H} is a linear algebraic group [22], we can define its unipotent radical \mathbf{R}_u to consist of the unipotent elements of its maximal solvable normal subgroup. If $\mathbf{G} \subset \mathbf{H}$ is a maximal compact subgroup, and if $\mathbf{G}_{\mathbb{C}} \subset \mathbf{H}$ is its complexification, then $\mathbf{G}_{\mathbb{C}}$ projects isomorphically onto the quotient group \mathbf{H}/\mathbf{R}_u . The Chevalley decomposition [17] moreover expresses \mathbf{H} as a semi-direct product

$$\mathbf{H} = \mathbf{G}_{\mathbb{C}} \ltimes \mathbf{R}_u$$

and we have a corresponding split short exact sequence

$$0 \rightarrow \mathfrak{r}_u \rightarrow \mathfrak{h} \rightarrow \mathfrak{g}_{\mathbb{C}} \rightarrow 0$$

of Lie algebras.

In their pioneering work on extremal Kähler vector fields [25], Futaki and Mabuchi next restricted the Futaki invariant \mathfrak{F} to $\mathfrak{g}_{\mathbb{C}} \subset \mathfrak{h}$. However, under mild hypotheses, this is not actually necessary:

PROPOSITION 2.1. *Let (M^{2m}, J) be a compact complex m -manifold of Kähler type for which $h^{1,0} = h^{2,0} = 0$. Then the Futaki invariant $\mathfrak{F}(\Omega) \in \mathfrak{h}^*$ automatically annihilates the Lie algebra \mathfrak{r}_u of the unipotent radical, and so belongs to $\mathfrak{g}_{\mathbb{C}}^*$. Moreover, this element is automatically real, and so belongs to \mathfrak{g}^* .*

As we show in Appendix A, this is actually a straightforward consequence of a theorem of Nakagawa [45].

Because the Futaki invariant is invariant under biholomorphisms, it is unchanged by the action of \mathbf{H} on \mathfrak{h} . It follows that $\mathfrak{F}(\Omega)$ must vanish when restricted to the derived subalgebra $[\mathfrak{h}, \mathfrak{h}]$. Thus, $\mathfrak{F}(\Omega) : \mathfrak{h} \rightarrow \mathbb{C}$ is actually a Lie-algebra character. In particular, $\mathfrak{F}(\Omega)$ annihilates the derived subalgebra $[\mathfrak{g}, \mathfrak{g}]$ of the maximal compact. Since the compactness of \mathbf{G} implies that it is a reductive Lie group, $\mathfrak{g} = [\mathfrak{g}, \mathfrak{g}] \oplus \mathfrak{z}$, where \mathfrak{z} is the center of \mathfrak{g} . We thus conclude

that $\mathfrak{F}(\Omega) \in \mathfrak{z}^*$ for any Kähler class Ω whenever M is as in Proposition A.3. Since \mathfrak{z} is contained in the Lie algebra of any maximal torus $\mathbf{T} \subset \mathbf{G}$, we thus deduce the following important fact:

PROPOSITION 2.2. *Let (M^{2m}, J) be a compact complex m -manifold of Kähler type for which $h^{1,0} = h^{2,0} = 0$. Let \mathbf{T} be a maximal torus in $\text{Aut}(M, J)$, and let \mathfrak{t} be the Lie algebra of \mathbf{T} . Then, for any Kähler class Ω on M , the Futaki invariant $\mathfrak{F}(\Omega)$ naturally belongs to \mathfrak{t}^* . In particular, $\mathfrak{F}(\Omega)$ is completely determined by its restriction to \mathfrak{t} .*

Now, for a fixed \mathbf{G} -invariant metric g , we have already noticed that every Killing field ξ on (M, g) is represented by a unique Hamiltonian f_ξ with $\int_M f_\xi d\mu = 0$, and that the Lie bracket on \mathfrak{g} is thereby transformed into the Poisson bracket on (M, ω) :

$$f_{[\xi, \eta]} = \{f_\xi, f_\eta\} = -\omega^{-1}(df_\xi, df_\eta).$$

Following Futaki and Mabuchi [25], we may therefore introduce a bilinear form \mathbb{B} on the real Lie algebra \mathfrak{g} by restricting the L^2 norm of (M, g) to the space of these Hamiltonians:

$$\mathbb{B}(\xi, \eta) = \int_M f_\xi f_\eta d\mu_g = \frac{1}{m!} \int_M f_\xi f_\eta \omega^m.$$

Since a straightforward version of Moser stability shows that the Kähler forms of any two \mathbf{G} -invariant metrics in a fixed Kähler class are \mathbf{G} -equivariantly symplectomorphic, this inner product only depends on Ω and the maximal compact $\mathbf{G} < \mathbf{H}$, not on the representative metric g . Moreover, since any two maximal compacts are conjugate in \mathbf{H} , one can show [25] that the corresponding complex-bilinear form on $\mathfrak{g}_\mathbb{C} = \mathfrak{h}/\mathfrak{r}_u$ is actually independent of the choice of maximal compact \mathbf{G} .

Since \mathbb{B} is positive-definite, and so defines an isomorphism $\mathfrak{g} \rightarrow \mathfrak{g}^*$, it also has a well-defined inverse which gives a positive-definite bilinear form

$$\mathbb{B}^{-1} : \mathfrak{g}^* \times \mathfrak{g}^* \rightarrow \mathbb{R}$$

on the Lie coalgebra of our maximal compact. On the other hand, assuming that (M, J) is as in Proposition A.3, we have already seen that $\mathfrak{F}(\Omega) \in \mathfrak{g}^*$ for any Kähler class Ω on M . Thus, the number

$$(2.2) \quad \|\mathfrak{F}(\Omega)\|^2 := \mathbb{B}^{-1}(\mathfrak{F}(\Omega), \mathfrak{F}(\Omega))$$

is independent of choices, and so is an invariant of (M, J, Ω) .

To see why this number has an important differential-geometric significance, let us first suppose that g is a \mathbf{G} -invariant Kähler metric with Kähler class Ω , and let \mathbb{P} be orthogonal projection in the real Hilbert space $L^2(M, g)$ to the subspace of normalized Hamiltonians representing the Lie algebra \mathfrak{g} of Killing fields on (M, g) . Restricting equation (2.1) to $\mathfrak{g} \subset \mathfrak{h}$, one observes that $\mathfrak{F}(\Omega) : \mathfrak{g} \rightarrow \mathbb{R}$ is exactly given by the \mathbb{B} -inner-product with the Killing

field whose Hamiltonian is $-\mathbf{p}(s - \bar{s})$. We thus immediately have

$$\int_M [\mathbf{p}(s - \bar{s})]^2 d\mu_g = \|\mathfrak{F}(\Omega)\|^2$$

and, since the projection \mathbf{p} is norm-decreasing, it follows that

$$(2.3) \quad \int_M (s - \bar{s})^2 d\mu_g \geq \|\mathfrak{F}(\Omega)\|^2$$

for any \mathbf{G} -invariant Kähler metric with Kähler class Ω . It is a remarkable fact, proved by Xiuxiong Chen [13], that inequality (2.3) actually holds even if g is *not* assumed to be \mathbf{G} -invariant. Moreover, equality holds in (2.3) if and only if $\nabla^{1,0}s$ is a holomorphic vector field, which is precisely the condition [11, 12] for g to be an extremal Kähler metric.

The bilinear form \mathbb{B} on \mathfrak{g} is bi-invariant. In particular, the center \mathfrak{z} of \mathfrak{g} is \mathbb{B} -orthogonal to the semi-simple factor $[\mathfrak{g}, \mathfrak{g}]$ of \mathfrak{g} . Thus, a computation of $\|\mathfrak{F}(\Omega)\|^2$ does not require a complete knowledge of the bilinear form \mathbb{B} ; only a knowledge of its restriction to \mathfrak{z} is required. This observation allows us to prove the following:

COROLLARY 2.3. *Let (M, J) be as in Proposition 2.2, let \mathbf{T} be a maximal torus in the complex automorphism group of (M, J) , and let \mathfrak{t} denote the Lie algebra of \mathbf{T} . If g is any \mathbf{T} -invariant Kähler metric with Kähler class Ω , and if*

$$\mathbb{B}_{\mathbf{T}} : \mathfrak{t} \times \mathfrak{t} \rightarrow \mathbb{R}$$

is the g -induced L^2 -norm restricted to normalized Hamiltonians, then

$$\|\mathfrak{F}(\Omega)\|^2 = \mathbb{B}_{\mathbf{T}}^{-1}(\mathfrak{F}(\Omega), \mathfrak{F}(\Omega))$$

where $\mathbb{B}_{\mathbf{T}}^{-1}$ denotes the inner product on \mathfrak{t}^ induced by $\mathbb{B}_{\mathbf{T}}$.*

PROOF. Let \mathbf{G} be a maximal compact subgroup of \mathbf{H} containing \mathbf{T} . Then, by Proposition 2.2, the assertion certainly holds for any \mathbf{G} -invariant Kähler metric \tilde{g} in Ω . However, by averaging, any \mathbf{T} -invariant Kähler metric with Kähler class Ω can be joined to \tilde{g} by a path of such metrics, and is therefore \mathbf{T} -equivariantly symplectomorphic to \tilde{g} by Moser stability. The claim therefore follows, since $\mathfrak{F}(\Omega) \in \mathfrak{t}^*$ is completely determined by (M, J, Ω) , while $\mathbb{B}_{\mathbf{T}}$ is completely determined by the symplectic form and normalized Hamiltonians representing elements of \mathfrak{t} . \square

3. Toric Manifolds

We now specialize our discussion to the toric case. For clarity, our presentation will be self-contained, and will include idiosyncratic proofs of various standard facts about toric geometry. For more orthodox expositions of some of these fundamentals, the reader might do well to consult [23] and [27].

We define a *toric manifold* to be a (connected) compact complex m -manifold (M^{2m}, J) of Kähler type which has non-zero Euler characteristic

and which is equipped a group of automorphisms generated by m commuting, periodic, J -preserving real vector fields which are linearly independent in the space of vector fields on M . Thus, the relevant group of automorphisms \mathbf{T} is required to be the image of the m -torus under some Lie group homomorphism $T^m \rightarrow \text{Aut}(M, J)$ which induces an injection of Lie algebras. Notice that our definition implies that there must be a fixed point $p \in M$ of this T^m -action. Indeed, the fixed point set of any circle action on a smooth compact manifold is [32] a disjoint union of smooth compact manifolds with total Euler characteristic equal to the Euler characteristic of the ambient space; by induction on the number of circle factors, it follows that the fixed-point set of any torus action on M therefore has total Euler characteristic $\chi(M) \neq 0$, and so, in particular, cannot be empty.

In light of this, let $p \in M$ be a fixed point of the given T^m -action on a toric manifold (M^{2m}, J) , and, by averaging, also choose a Kähler metric g on M which is T^m -invariant. Then T^m acts on $T_p M \cong \mathbb{C}^m$ in a manner preserving both g and J , giving us a unitary representation $T^m \rightarrow \mathbf{U}(m)$. Since the action of T^m on $T_p M$ completely determines the action on M via the exponential map $T_p M \rightarrow M$ of g , and since, by hypothesis, the Lie algebra of T^m injects into the vector fields on M , it follows that the above unitary representation gives rise to a faithful representation of $\mathbf{T} < \text{Aut}(M, J)$. However, $\mathbf{U}(m)$ has rank m , so the image of $T^m \rightarrow \mathbf{U}(m)$ must be a maximal torus in $\mathbf{U}(m)$; thus, after a change of basis of \mathbb{C}^m , \mathbf{T} may be identified with the standard maximal torus $\mathbf{U}(1) \times \cdots \times \mathbf{U}(1) \subset \mathbf{U}(m)$ consisting of diagonal matrices. In particular, $\mathbf{T} < \text{Aut}(M, J)$ is intrinsically an m -torus, and has many free orbits. Since the origin in \mathbb{C}^m is the only fixed point of the diagonal torus in $\mathbf{U}(m)$, it also follows that p must be an isolated fixed point of \mathbf{T} . But since the same argument applies equally well to any other fixed point, this shows that the fixed-point set $M_{\mathbf{T}}$ of \mathbf{T} is discrete, and therefore finite. In particular, $\chi(M)$ must equal the cardinality of $M_{\mathbf{T}}$, so the Euler characteristic of M is necessarily positive.

The above arguments in particular show that the toric condition can be reformulated as follows: a toric m -manifold is a compact complex m -manifold (M, J) of Kähler type, together with an m -torus $\mathbf{T} \subset \text{Aut}(M, J)$ that has both a free orbit \mathcal{Q} and a fixed point p . To check the equivalence, note that this reformulation implies that the Euler characteristic $\chi(M)$ is positive, because the fixed-point set $M_{\mathbf{T}}$ is necessarily finite, and by hypothesis is also non-empty.

Now let (M, J, \mathbf{T}) be a toric m -manifold, and let $j : \mathcal{Q} \hookrightarrow M$ be the inclusion of a free \mathbf{T} -orbit. Since \mathbf{T} also has a fixed-point p , and since any two \mathbf{T} -orbits are homotopic, it follows that j is homotopic to a constant map. Consequently, the induced homomorphism $j^* : H^k(M) \rightarrow H^k(\mathcal{Q})$ must be the zero map in all dimensions $k > 0$. However, the restriction of the Kähler form $\omega = g(J \cdot, \cdot)$ to $\mathcal{Q} \approx \mathbf{T}$ is an invariant 2-form on $\mathbf{T} \approx T^m$. Since every deRham class on T^m contains a unique invariant form, and since $j^*[\omega] = 0 \in H^2(T^m, \mathbb{R})$, it follows that $j^*\omega$ must vanish identically. Thus \mathcal{Q} is

a Lagrangian submanifold, which is to say that $T\mathcal{Q}$ is everywhere orthogonal to $J(T\mathcal{Q})$. In particular, if ξ_1, \dots, ξ_m are the generators of the \mathbf{T} -action, the corresponding holomorphic vector fields $\Xi_j = -(J\xi_j + i\xi_j)/2$ span $T^{1,0}M$ in a neighborhood of \mathcal{Q} . Integrating the flows of the commuting vector fields ξ_j and $J\xi_j$, we thus obtain a holomorphic action of the complexified torus $(\mathbb{C}^\times)^m$ which has both a fixed point and an open orbit \mathcal{U} .

In particular, (M, J) carries m holomorphic vector fields Ξ_1, \dots, Ξ_m which vanish at p , but which nonetheless span $T^{1,0}(M)$ at a generic point. It follows that M cannot carry a non-trivial holomorphic k -form $\alpha \in H^{k,0}(M)$ for any $k > 0$, since, for any choice of j_1, \dots, j_k , the “component” functions $\alpha(\Xi_{j_1}, \dots, \Xi_{j_k})$ would be holomorphic, and hence constant, and yet would have to vanish at the fixed point p . In particular, we may invoke Kodaira’s observation [33] that any Kähler manifold with $H^{2,0} = 0$ admits Hodge metrics, and so is projective. This gives us the following result:

LEMMA 3.1. *Any toric manifold M is projective algebraic, and satisfies $H^{k,0}(M) = 0$ for all $k > 0$.*

In particular, the identity component $\mathbf{H} = \text{Aut}^0(M, J)$ of the automorphism group of our toric m -manifold is linear algebraic. Let $\mathbf{T} < \mathbf{H}$ be the m -torus associated with the toric structure of (M, J) . Using the Chevalley decomposition, we can then choose a maximal compact subgroup $\mathbf{G} < \mathbf{H}$ containing \mathbf{T} . Also choose a \mathbf{G} -invariant Kähler metric g on M and a fixed point p of \mathbf{T} . We will now study the centralizer $Z(\mathbf{T}) < \mathbf{G}$, consisting of elements of \mathbf{G} that commute with all elements of \mathbf{T} . Observe that

$$a \in Z(\mathbf{T}), b \in \mathbf{T} \implies b(a(p)) = a(b(p)) = a(p),$$

so that $Z(\mathbf{T})$ acts by permutation on the finite set $M_{\mathbf{T}}$ of fixed points. In particular, the identity component $Z^0(\mathbf{T})$ of $Z(\mathbf{T})$ must send p to itself. Once more invoking the exponential map of g , we thus obtain a faithful unitary representation of $Z^0(\mathbf{T})$ by considering its induced action on $T_p M \cong \mathbb{C}^m$. However, the image of $Z^0(\mathbf{T})$ in $\mathbf{U}(m)$ must then be a subgroup of the centralizer of the diagonal torus $\mathbf{U}(1) \times \dots \times \mathbf{U}(1)$ in $\mathbf{U}(m)$. But since the latter centralizer is just the diagonal torus itself, we conclude that $Z^0(\mathbf{T}) = \mathbf{T}$. It follows that \mathbf{T} is a maximal torus in \mathbf{G} , and hence also in $\mathbf{H} = \mathbf{G}_{\mathbb{C}} \ltimes \mathbf{R}_{\mathbf{u}}$:

LEMMA 3.2. *Let (M^{2m}, J) be a toric manifold, and let $\mathbf{T} < \text{Aut}(M, J)$ be the associated m -torus. Then \mathbf{T} is a maximal torus in $\text{Aut}(M, J)$.*

Combining this result with Lemma 3.1 and Proposition 2.2, we can thus generalize [46, Theorem 1.9] to irrational Kähler classes:

PROPOSITION 3.3. *Let (M^{2m}, J) be a toric manifold, let \mathbf{T} be the given m -torus in its automorphism group, and let \mathfrak{t} be the Lie algebra of \mathbf{T} . Then, for any Kähler class Ω on M , the Futaki invariant $\mathfrak{F}(\Omega)$ naturally belongs to \mathfrak{t}^* . In particular, $\mathfrak{F}(\Omega)$ is completely determined by its restriction to \mathfrak{t} .*

However, we will not simply need to know where $\mathfrak{F}(\Omega)$ lives; our goal will require us to calculate its norm with respect to the relevant bilinear form. Fortunately, Lemma 3.2 and Corollary 2.3 together imply the following result:

PROPOSITION 3.4. *Let (M^{2m}, J) be a toric manifold, let \mathbf{T} be the given m -torus in its automorphism group, and let \mathfrak{t} be the Lie algebra of \mathbf{T} . If g is any \mathbf{T} -invariant Kähler metric with Kähler class Ω , and if*

$$\mathbb{B}_{\mathbf{T}} : \mathfrak{t} \times \mathfrak{t} \rightarrow \mathbb{R}$$

is the g -induced L^2 -norm restricted to normalized Hamiltonians, then

$$\|\mathfrak{F}(\Omega)\|^2 = \mathbb{B}_{\mathbf{T}}^{-1}(\mathfrak{F}(\Omega), \mathfrak{F}(\Omega))$$

where $\mathbb{B}_{\mathbf{T}}^{-1}$ denotes the inner product on \mathfrak{t}^ induced by $\mathbb{B}_{\mathbf{T}}$.*

Of course, Lemma 3.1 has many other interesting applications. For example, by Hodge symmetry, it implies the Todd genus is given by

$$\chi(M, \mathcal{O}) = \sum_k (-1)^k h^{0,k}(M) = 1$$

for any toric manifold M . Since the same argument could also be applied to any finite covering of M , whereas $\chi(M, \mathcal{O})$ is multiplicative under coverings, one immediately sees that M cannot have non-trivial finite covering spaces. In particular, this implies that $H_1(M, \mathbb{Z}) = 0$.

However, one can easily do much better. Choose a \mathbf{T} -invariant Kähler metric g with Kähler form ω . Because $b_1(M) = 0$ by Lemma 3.1, the symplectic vector fields ξ_1, \dots, ξ_m must then have Hamiltonians, so that $\xi_j = J\nabla f_j$ for suitable functions f_1, \dots, f_m . Let a_1, \dots, a_m be real numbers which are linearly independent over \mathbb{Q} , and let $f = \sum_j a_j f_j$. The corresponding symplectic vector field $\xi = \sum_j a_j \xi_j$ is thus a Killing field for g , and its flow is dense in the torus $\mathbf{T} < \text{Aut}(M, J)$. Consequently, ξ vanishes only at the fixed points of \mathbf{T} . Since ξ is Killing, with only isolated zeroes, it then follows that $\nabla \xi$ is non-degenerate at each fixed point p of \mathbf{T} , in the sense that it defines an isomorphism $T_p \rightarrow T_p$. Since $\nabla_a \nabla_b f = \omega_{bc} \nabla_a \xi^c$, this implies that the Hessian of f is non-degenerate at each zero of df ; that is, f is a Morse function on M . However, since ξ is the real part of a holomorphic vector field, $\bar{\partial} \partial^\# f = 0$, and this is equivalent to saying that the Riemannian Hessian $\nabla \nabla f$ is everywhere J -invariant. Since the Riemannian Hessian coincides with the naïve Hessian at a critical point, this shows that every critical point of f must have even index. It follows [44] that M is homotopy equivalent to a CW complex consisting entirely of even-dimensional cells. In particular, we obtain the following:

LEMMA 3.5. *Any toric manifold is simply connected, and has trivial homology in all odd dimensions.*

Finally, notice that Lemma 3.1 implies that the canonical line bundle $K = \Lambda^{m,0}$ of a toric m -manifold has no non-trivial holomorphic sections. However, essentially the same argument also shows that positive powers K^ℓ cannot have non-trivial holomorphic sections either, since the pairing of such a section with $(\Xi_1 \wedge \cdots \wedge \Xi_m)^{\otimes \ell}$ would again result in a constant function which would have to vanish at p . Thus, all the plurigenera $p_\ell = h^0(\mathcal{O}(K^\ell))$ of any toric manifold must vanish. In other words:

LEMMA 3.6. *Any toric manifold has Kodaira dimension $-\infty$.*

4. The Virtual Action

As previously discussed in connection with (2.3), a theorem of Chen [13] says that any Kähler metric g on a compact complex manifold M satisfies

$$(4.1) \quad \int_M (s - \bar{s})^2 d\mu_g \geq \|\mathfrak{F}(\Omega)\|^2 ,$$

where $\Omega = [\omega]$ is the Kähler class of g ; moreover, equality holds iff g is an extremal Kähler metric. On the other hand,

$$(4.2) \quad \int_M s^2 d\mu_g = \int_M (s - \bar{s})^2 d\mu_g + \int_M \bar{s}^2 d\mu_g$$

as may be seen by applying the Pythagorean theorem to L^2 -norms. Since s is the trace of the Ricci tensor with respect to the metric, and because the Ricci form is essentially the curvature of the canonical line bundle, we also know that

$$(4.3) \quad \int_M s d\mu = \frac{4\pi c_1 \cdot \Omega^{m-1}}{(m-1)!}$$

in complex dimension m ; meanwhile, the volume of an m -dimensional Kähler m -manifold is just given by

$$\int_M d\mu = \frac{\Omega^m}{m!} .$$

Hence

$$\int_M \bar{s}^2 d\mu = \frac{(\int_M s d\mu)^2}{\int_M d\mu} = \frac{16\pi^2 m}{(m-1)!} \frac{(c_1 \cdot \Omega^{m-1})^2}{\Omega^m}$$

and (4.1) thus implies that

$$(4.4) \quad \int_M s^2 d\mu_g \geq \frac{16\pi^2 m}{(m-1)!} \frac{(c_1 \cdot \Omega^{m-1})^2}{\Omega^m} + \|\mathfrak{F}(\Omega)\|^2$$

with equality iff g is an extremal Kähler metric on (M^{2m}, J) .

Now specializing to the case of complex dimension $m = 2$, we have

$$\int_M s^2 d\mu_g \geq 32\pi^2 \frac{(c_1 \cdot \Omega)^2}{\Omega^2} + \|\mathfrak{F}(\Omega)\|^2$$

for any Kähler metric g with Kähler class $[\omega] = \Omega$ on a compact complex surface (M^4, J) . In other words, if we define a function on the Kähler cone by

$$\mathcal{A}(\Omega) := \frac{(c_1 \cdot \Omega)^2}{\Omega^2} + \frac{1}{32\pi^2} \|\mathfrak{F}(\Omega)\|^2 ,$$

then

$$(4.5) \quad \frac{1}{32\pi^2} \int_M s_g^2 d\mu_g \geq \mathcal{A}(\Omega)$$

for any Kähler metric g with Kähler class Ω , with equality iff g is an extremal Kähler metric. The function $\mathcal{A}(\Omega)$ will be called the *virtual action*. Our normalization has been chosen so that $\mathcal{A}(\Omega) \geq c_1^2(M)$, with equality iff the Futaki invariant vanishes and Ω is a multiple of c_1 . (Incidentally, the latter occurs iff Ω is the Kähler class of a Kähler-Einstein metric on (M^4, J) [4, 47, 52, 54].) The fact that the virtual action $\mathcal{A}(\Omega)$ is homogeneous of degree 0 in Ω corresponds to the fact that the Calabi energy $\mathcal{C}(g)$ is scale-invariant in real dimension four.

In complex dimension $m = 2$, one important reason for studying the Calabi energy \mathcal{C} is the manner in which (1.2) relates it to the *Weyl functional*

$$\mathcal{W}(g) = \int_M |W|_g^2 d\mu_g$$

where the Weyl curvature W is the conformally invariant piece of the curvature tensor. It is easy to check that \mathcal{W} is also conformally invariant, and may therefore be considered as a functional on the space of conformal classes of Riemannian metrics. Critical points of the Weyl functional are characterized [5, 9] by the vanishing of the *Bach tensor*

$$B_{ab} := (\nabla^c \nabla^d + \frac{1}{2} r^{cd}) W_{acbd}$$

and so are said to be *Bach-flat*; obviously, this is a conformally invariant condition. The Bianchi identities immediately imply that any Einstein metric on a 4-manifold is Bach-flat, and it therefore follows that any conformally Einstein metric is Bach-flat, too. The converse, however, is false; for example, self-dual and anti-self-dual metrics are also Bach-flat, and such metrics exist on many compact 4-manifolds [34, 35, 51] that do not admit Einstein metrics.

When the Weyl functional \mathcal{W} is restricted to the space of Kähler metrics, equation (1.2) shows that it becomes equivalent to the Calabi energy \mathcal{C} . Nonetheless, the following result [15] may come as something of a surprise:

PROPOSITION 4.1. *Let g be a Kähler metric on a compact complex surface. Then g is Bach-flat if and only if*

- g is an extremal Kähler metric, and
- its Kähler class Ω is a critical point of the virtual action \mathcal{A} .

This gives rise to a remarkable method of constructing Einstein metrics, courtesy of a beautiful result of Derdziński [19, Proposition 4]:

PROPOSITION 4.2. *If the scalar curvature s of a Bach-flat Kähler metric g on a complex surface (M^4, J) is not identically zero, then the conformally related metric $h = s^{-2}g$ is Einstein on the open set $s \neq 0$ where it is defined.*

5. Toric Surfaces

We will now prove Theorem A by computing the virtual action $\mathcal{A}(\Omega)$ for any Kähler class on a toric surface. An important intermediate step in this process involves an explicit computation of the Futaki invariant $\mathfrak{F}(\Omega)$. Up to a universal constant, our answer agrees with that of various other authors [21, 25, 41, 49], but determining the correct constant is crucial for our purposes. For this reason, our first proof will be based on the author's formula [34] for the scalar curvature of a Kähler surface with isometric S^1 action.

By a *toric surface*, we mean a toric manifold (M, J, \mathbf{T}) of complex dimension two. This is equivalent¹ to saying that (M^4, J) is a simply connected compact complex surface equipped with a 2-torus $\mathbf{T} < \text{Aut}(M, J)$. By Castelnuovo's criterion [7, 26], Lemma 3.1 and Lemma 3.6, any toric surface (M, J) can be obtained from either \mathbb{CP}_2 or a Hirzebruch surface by blowing up points. Indeed, since the holomorphic vector fields generating the torus action on M automatically descend to the minimal model, the toric structure of (M, J) can be obtained from a toric structure on \mathbb{CP}_2 or a Hirzebruch surface by iteratively blowing up fixed points of the torus action. For more direct proofs, using the toric machinery of fans or moment polytopes, see [23, 27].

Let $(M^4, J, \mathbf{T}, \Omega)$ now be a toric surface with fixed Kähler class. By averaging, we can then find a \mathbf{T} -invariant Kähler metric g on (M, J) with Kähler form $\omega \in \Omega$. Choose an isomorphism $\mathbf{T} \cong \mathbb{R}^2/\mathbb{Z}^2$, and denote the corresponding generating vector fields of period 1 by ξ_1 and ξ_2 . Since $b_1(M) = 0$, there are Hamiltonian functions x_1 and x_2 on M with $\xi_j = J \text{grad } x_j$, $j = 1, 2$. This makes (M, ω) into a Hamiltonian T^2 -space in the sense of [27]. In particular, the image of M under $\vec{x} = (x_1, x_2)$ is [3, 28] a convex polygon $P \subset \mathbb{R}^2$ whose area is exactly the volume of (M, g) . The map $\vec{x} : M \rightarrow \mathbb{R}^2$ is called the *moment map*, and its image $P = \vec{x}(M)$ will be called the *moment polygon*. Of course, since we have not insisted that the Hamiltonians x_k have integral zero, our moment map is only determined up to translations of \mathbb{R}^2 . Modulo this ambiguity, however, the moment polygon is uniquely determined by (M, ω, \mathbf{T}) , together with the chosen basis (ξ_1, ξ_2) for the Lie algebra \mathbf{t} of \mathbf{T} . Moreover, since a straightforward Moser-stability argument shows that any two \mathbf{T} -invariant Kähler forms in Ω are \mathbf{T} -equivariantly symplectomorphic, the moment polygon really only depends on $(M, J, \Omega, (\xi_1, \xi_2))$. However, outer automorphisms of \mathbf{T} can be used to

¹In one direction, this equivalence follows because any simply connected compact complex surface is of Kähler type [10, 50] and has positive Euler characteristic. On the other hand, the converse follows from Lemma 3.5.

alter (ξ_1, ξ_2) by an $\mathbf{SL}(2, \mathbb{Z})$ transformation, and this in turn changes the moment polygon by an $\mathbf{SL}(2, \mathbb{Z})$ transformation of \mathbb{R}^2 . Moreover, since the vertices of P correspond to the fixed points of \mathbf{T} , and because the action of \mathbf{T} on the tangent space of any fixed point can be identified with that of the diagonal torus $\mathbf{U}(1) \times \mathbf{U}(1) \subset \mathbf{U}(2)$, a neighborhood of any corner of P can be transformed into a neighborhood of the origin in the positive quadrant of \mathbb{R}^2 by an element of $\mathbf{SL}(2, \mathbb{Z})$ and a translation [18]. Polygons with the latter property are said to be *Delzant*, and any Delzant polygon arises from a uniquely determined toric surface, equipped with a uniquely determined Kähler class [27].

We now introduce a measure $d\lambda$ on the boundary ∂P of our moment polygon. To do this, first notice that each edge of P is the image of a rational curve $C_i \cong \mathbb{CP}_1$ in (M, J) which is fixed by an S^1 subgroup of T^2 , and hence by a \mathbb{C}^\times subgroup of the complexified torus $\mathbb{C}^\times \times \mathbb{C}^\times$. We then define the measure $d\lambda$ along the edge $\ell_i = \vec{x}(C_i)$ to be the push-forward, via \vec{x} , of the smooth area measure on C_i given by the restriction of the Kähler form ω . Since a rational linear combination of the x_k is a Hamiltonian for rotation of C_i about two fixed points, $d\lambda$ is a constant times 1-dimensional Lebesgue measure on the line segment ℓ_i , with total length

$$\int_{\ell_i} d\lambda = \int_{C_i} \omega =: \mathcal{A}_i$$

equal to the area of corresponding holomorphic curve in M . Here the index i is understood to run over the edges of ∂P .

When an edge is parallel to either axis, $d\lambda$ just becomes standard Euclidean length measure. More generally, on an arbitrary edge, it must coincide with the pull-back of Euclidean length via any $\mathbf{SL}(2, \mathbb{Z})$ transformation which sends the edge to a segment parallel to an axis. Because P is a Delzant polygon, this contains enough information to completely determine $d\lambda$, and leads to a consistent definition of the measure because the stabilizer

$$\left\{ \pm \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} \mid k \in \mathbb{Z} \right\}$$

of the x_1 -axis in $\mathbf{SL}(2, \mathbb{Z})$ preserves Euclidean length on this axis. However, the Euclidean algorithm of elementary number theory implies that every pair (p, q) of relatively prime non-zero integers belongs to the $\mathbf{SL}(2, \mathbb{Z})$ -orbit of $(1, 0)$. One can therefore compute edge-lengths with respect to $d\lambda$ by means of the following recipe: Given an edge of P which is not parallel to either axis, its slope m is a non-zero rational number, and so can be expressed in lowest terms as $m = q/p$, where p and q are relatively prime non-zero integers. The displacement vector \vec{v} representing the difference between the two endpoints of the edge can thus be written as $\vec{v} = (up, uq)$ for some $u \in \mathbb{R} - \{0\}$. The length of the edge with respect to $d\lambda$ then equals $|u|$.

We can now associate two different barycenters with our moment polygon. First, there is the barycenter $\vec{x} = (\bar{x}_1, \bar{x}_2)$ of the interior of P , as defined

by

$$\bar{x}_k = \int_P x_k \, da = \frac{\int_P x_k \, da}{\int_P da}$$

where da is standard 2-dimensional Lebesgue measure in \mathbb{R}^2 . Second, there is the barycenter $\langle \vec{x} \rangle = (\langle x_1 \rangle, \langle x_2 \rangle)$ of the perimeter ∂P , defined by

$$\langle x_k \rangle = \int_{\partial P} x_k \, d\lambda = \frac{\int_{\partial P} x_k \, d\lambda}{\int_{\partial P} d\lambda}$$

These two barycenters certainly need not coincide in general. It is therefore natural to consider the displacement vector

$$\vec{\mathfrak{D}} = \langle \vec{x} \rangle - \bar{x}$$

that measures their separation. Notice that $\vec{\mathfrak{D}}$ is translation invariant — it is unchanged if we alter the Hamiltonians (x_1, x_2) by adding constants.

Next, we introduce the moment-of-inertia matrix Π of P , which encodes the moment of inertia of the polygon about an arbitrary axis in \mathbb{R}^2 passing through its barycenter \bar{x} . Thus Π is the positive-definite symmetric 2×2 matrix with entries given by

$$\Pi_{jk} = \int_P (x_j - \bar{x}_j)(x_k - \bar{x}_k) da$$

where da once again denotes the usual Euclidean area form on the interior of P , and exactly equals the push-forward of the metric volume measure on M . For our purposes, it is important to notice that Π is always an invertible matrix.

Finally, let $|\partial P| = \int_{\partial P} d\lambda = \sum_i \mathcal{A}_i$ denote the perimeter of the moment polygon with respect to the measure $d\lambda$ introduced above, and let $|P| = \int_P da$ denote the area of its interior in the usual sense. With these notational conventions, we are now ready to state the main result of this section:

THEOREM 5.1. *If (M, J, Ω) is any toric surface with fixed Kähler class, then*

$$(5.1) \quad \mathcal{A}(\Omega) = \frac{|\partial P|^2}{2} \left(\frac{1}{|P|} + \vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}} \right)$$

where P is the moment polygon determined by the given T^2 -action.

The proof of Theorem 5.1 crucially depends on a computation of the Futaki invariant, which, we recall, is a character on the Lie algebra of holomorphic vector fields. Let us therefore consider the holomorphic vector fields $\Xi_k = \nabla^{1,0} x_k$ whose holomorphy potentials are the Hamiltonians of the periodic Killing fields ξ_k . These are explicitly given by

$$\Xi_k = -\frac{1}{2} (J\xi_k + i\xi_k).$$

PROPOSITION 5.2. *Suppose that (M, J, Ω) is a toric surface with fixed Kähler class, and let Ξ_k be the generators of the associated complex torus action, normalized as above. Let*

$$\mathfrak{F}_k := \mathfrak{F}(\Xi_k, \Omega)$$

be the corresponding components of the Futaki invariant of (M, J, Ω) . Then the vector $\vec{\mathfrak{F}} = (\mathfrak{F}_1, \mathfrak{F}_2)$ is explicitly given by

$$\vec{\mathfrak{F}} = -4\pi |\partial P| \vec{\mathfrak{D}}$$

where $|\partial P|$ again denotes the weighted perimeter of the moment polygon P , and $\vec{\mathfrak{D}}$ is again the vector joining the barycenters of the interior and weighted boundary of P .

PROOF. More explicitly, the assertion is that

$$(5.2) \quad \mathfrak{F}_k = -4\pi \sum_i \left(\langle x_k \rangle_i - \bar{x}_k \right) \mathcal{A}_i$$

where \bar{x}_k is once again the k^{th} coordinate of the barycenter of the interior of the moment polygon P , $\langle x_k \rangle_i$ is the k^{th} coordinate of the center of the i^{th} edge of P , and \mathcal{A}_i is the weighted length of i^{th} edge.

We will now prove (5.2) using a method [31, 40] which is broadly applicable to \mathbb{C}^\times -actions, but which nicely simplifies in the toric case. We thus make a choice of $k = 1$ or 2 , and set $\Xi = \Xi_k$, $\xi = \xi_k$, and $x = x_k$ for this choice of k . In order to facilitate comparison with [31, 34, 40], set $\eta = \xi/2\pi$, so that η is a symplectic vector field of period 2π , with Hamiltonian $t = x/2\pi$. Let $\Sigma = M // \mathbb{C}^\times$ be the stable quotient of (M, J) by the action generated by Ξ , and observe that the following interesting special properties hold in our toric setting:

- the stable quotient Σ has genus 0; and
- all the isolated \mathbb{C}^\times fixed points project to just two points $q_1, q_2 \in \Sigma$.

Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}$, respectively, denote the minimum and maximum of the Hamiltonian t , so that $t(M) = [\mathbf{a}, \mathbf{b}]$. If $t^{-1}(\{\mathbf{a}\})$ or $t^{-1}(\{\mathbf{b}\})$ is an isolated fixed point, blow up M there to obtain \hat{M} , and pull the metric g back to \hat{M} as a degenerate metric; otherwise, let $\hat{M} = M$. We then have a holomorphic quotient map $\varpi : \hat{M} \rightarrow \Sigma$. Let C_+ and C_- be the holomorphic curves in \hat{M} given by $t^{-1}(\mathbf{b})$ and $t^{-1}(\mathbf{a})$, respectively. Except when they are just artifacts produced by blowing up, the curves C_\pm number among the rational curves C_i which project to the sides of the moment polygon P ; the others, after proper transform if necessary, form a sub-collection $\{E_j\} \subset \{C_i\}$ characterized by $\varpi^{-1}(\{q_1, q_2\}) = \cup_j E_j$ for a preferred pair of distinct points $q_1, q_2 \in \Sigma$. Each E_j is the closure of a \mathbb{C}^\times -orbit, and we will let $m_j \in \mathbb{Z}^+$ denote the order of the isotropy of \mathbb{C}^\times acting on the relevant orbit. Also let t_j^- and t_j^+ denote the minimum and maximum of t on E_j , so that $t(E_j) = [t_j^-, t_j^+]$, and observe that

$$\langle t \rangle_j := (t_j^- + t_j^+)/2$$

coincides with the average value of t on E_j with respect to g -area measure.

Let us now define $\varphi : \hat{M} \rightarrow \Sigma \times [\mathbf{a}, \mathbf{b}]$ to be the map $\varpi \times t$. If p_1, \dots, p_m are the images in $\Sigma \times (\mathbf{a}, \mathbf{b})$ of the isolated fixed points, and if

$$X = [\Sigma \times (\mathbf{a}, \mathbf{b})] - \{p_1, \dots, p_m\},$$

then the open dense set $Y = \varphi^{-1}(X) \subset \hat{M}$ map be viewed as an orbifold S^1 -principal bundle over X , and comes equipped with a unique connection 1-form θ whose kernel is g -orthogonal to η and which satisfies $\theta(\eta) = 1$. We may now express the given Kähler metric g as

$$g = w \check{g}(t) + w dt^{\otimes 2} + w^{-1} \theta^{\otimes 2},$$

for a positive functions $w > 0$ on X and a family orbifold metrics $\check{g}(t)$ on Σ .

Because g , w and dt are geometrically defined, $\check{g}(t)$ is an invariantly defined, t -dependent orbifold Kähler metric on Σ for all regular values of t ; moreover, it is a smooth well-defined tensor field on all of $(\Sigma - \{q_1, q_2\}) \times (\mathbf{a}, \mathbf{b})$. Now notice that the Kähler quotient of M associated with a regular value of the Hamiltonian is manifestly $(\Sigma, w \check{g}(t))$, and must therefore tend to the restriction of g to C_{\pm} as $t \rightarrow \mathbf{a}$ or \mathbf{b} . On the other hand, $w^{-1} = g(\eta, \eta)$ by construction, and since η is a Killing field of period 2π and Hamiltonian t , we have $g(\eta, \eta) = 2|t - \mathbf{a}| + O(|t - \mathbf{a}|^2)$ near $t = \mathbf{a}$, and similarly near $t = \mathbf{b}$. Thus [31, 40], letting $\check{\omega}(t)$ be the Kähler form of $\check{g}(t)$, we have

$$\begin{aligned} \check{\omega}|_{t=\mathbf{a}} &= \check{\omega}|_{t=\mathbf{b}} = 0 \\ \frac{d}{dt} \check{\omega} \Big|_{t=\mathbf{a}} &= 2\omega|_{C^-} \\ \frac{d}{dt} \check{\omega} \Big|_{t=\mathbf{b}} &= -2\omega|_{C^+}. \end{aligned}$$

More surprisingly, the calculations underlying the hyperbolic ansatz of [34] show [40, equation (3.16)] that the scalar curvature density of g may be globally expressed on $Y \subset M$ as

$$s d\mu = \left[2\check{\rho} - \frac{d^2}{dt^2} \check{\omega} \right] \wedge dt \wedge \theta$$

where $\check{\rho}(t)$ is the Ricci form of $\check{g}(t)$. However, for regular values of $t \in (\mathbf{a}, \mathbf{b})$, the Gauss-Bonnet formula for orbifolds tells us that

$$\begin{aligned} \frac{1}{2\pi} \int_{\Sigma} \check{\rho}(t) &= \chi(\Sigma) - \sum_j \delta_j(t) \left(1 - \frac{1}{m_j}\right) \\ &= \chi(S^2) - 2 + \sum_j \frac{1}{m_j} \delta_j(t) \\ &= \sum_j \frac{1}{m_j} \delta_j(t) \end{aligned}$$

where we have introduced the characteristic function

$$\delta_j(t) = \begin{cases} 1 & t_j^- < t < t_j^+ \\ 0 & \text{otherwise} \end{cases}$$

of (t_j^-, t_j^+) in order to keep track of which two curves E_j meet a given regular level-set of the Hamiltonian function t .

Now the Futaki invariant is defined in terms of the L^2 inner product of the scalar curvature s of g with normalized holomorphy potentials. It is therefore pertinent to observe that

$$\begin{aligned} \int_M ts \, d\mu &= \int_Y ts \, d\mu \\ &= \int_Y t \left[2\check{\rho} - \frac{d^2}{dt^2} \check{\omega} \right] \wedge dt \wedge \theta \\ &= 4\pi \int_{\mathbf{a}}^{\mathbf{b}} t \left[\int_{\Sigma} \check{\rho} \right] dt - 2\pi \int_{\Sigma} \left[\int_{\mathbf{a}}^{\mathbf{b}} t \frac{d^2}{dt^2} \check{\omega} \right] dt \\ &= 4\pi \int_{\mathbf{a}}^{\mathbf{b}} 2\pi \left[\sum_j \frac{1}{m_j} \delta_j(t) \right] t \, dt - 2\pi \int_{\Sigma} \left(\left[t \frac{d}{dt} \check{\omega} \right]_{\mathbf{a}}^{\mathbf{b}} - \int_{\mathbf{a}}^{\mathbf{b}} \frac{d\check{\omega}}{dt} dt \right) \\ &= 4\pi \sum_j \frac{2\pi}{m_j} \int_{t_j^-}^{t_j^+} t \, dt - 2\pi \int_{\Sigma} \left(-2\mathbf{b}\omega \Big|_{t=\mathbf{b}} - 2\mathbf{a}\omega \Big|_{t=\mathbf{a}} - [\check{\omega}]_{\mathbf{a}}^{\mathbf{b}} \right) \\ &= 4\pi \sum_j \frac{2\pi(t_j^+ - t_j^-)}{m_j} \frac{t_j^+ + t_j^-}{2} + 4\pi \left(\mathbf{a} [\omega] \cdot C^- + \mathbf{b} [\omega] \cdot C^+ \right) \\ &= 4\pi \sum_j ([\omega] \cdot E_j) \langle t \rangle_j + 4\pi \left(\mathbf{a} [\omega] \cdot C^- + \mathbf{b} [\omega] \cdot C^+ \right) \\ &= 4\pi \sum_i \langle t \rangle_i \mathcal{A}_i = 2 \sum_i \langle x \rangle_i \mathcal{A}_i \end{aligned}$$

where $\mathcal{A}_i = [\omega] \cdot C_i$ is once again the area of C_i . Since the holomorphy potential of the holomorphic vector field Ξ is $x = 2\pi t$, we therefore have

$$\begin{aligned} -\mathfrak{F}(\Xi, [\omega]) &= \int_M s(x - \bar{x}) \, d\mu \\ &= 2\pi \int_M st \, d\mu - \bar{x} \int_M s \, d\mu \\ (5.3) \quad &= \left(4\pi \sum_i \langle x \rangle_i \mathcal{A}_i \right) - \bar{x} \left(4\pi c_1 \cdot [\omega] \right) \end{aligned}$$

where \bar{x} again denotes the average value of x on M .

Next, notice that $\cup_i C_i$ is the zero locus of the holomorphic section $\Xi_1 \wedge \Xi_2$ of the anti-conical line-bundle $K^{-1} = \wedge^2 T^{1,0}$, and that, since the imaginary parts of Ξ_1 and Ξ_2 are Killing fields, this section is transverse to the zero section away from the intersection points $C_i \cap C_j$. It follows that the homology

class of $\cup_i C_i$ is Poincaré dual to $c_1(M, J) = c_1(K^{-1})$. Hence

$$c_1 \cdot [\omega] = \sum_i C_i \cdot [\omega] = \sum_i \mathcal{A}_i$$

so that (5.3) simplifies to become

$$\mathfrak{F}(\Xi, [\omega]) = -4\pi \sum_i (\langle x \rangle_i - \bar{x}) \mathcal{A}_i$$

and (5.2) therefore follows by setting $\Xi = \Xi_k$ and $x = x_k$. \square

With this preparation, we can now calculate $\mathcal{A}(\Omega)$ for any toric surface.

PROOF OF THEOREM 5.1. Relative to the basis given by the normalized holomorphy potentials $\{x_k - \bar{x}_k \mid k = 1, 2\}$, Proposition 5.2 tells us that the restriction of the Futaki invariant to \mathfrak{t} is given by

$$\vec{\mathfrak{F}} = (\mathfrak{F}_1, \mathfrak{F}_2) = -4\pi |\partial P| \vec{\mathfrak{D}}.$$

Since the L^2 inner product $\mathbb{B}_{\mathbf{T}}$ on \mathfrak{t} is given in this basis by the moment-of-inertia matrix

$$\Pi = \left[\int_P (x_j - \bar{x}_j)(x_k - \bar{x}_k) d\alpha \right] = \left[\int_M (x_j - \bar{x}_j)(x_k - \bar{x}_k) d\mu \right],$$

the dual inner product $\mathbb{B}_{\mathbf{T}}^{-1}$ on \mathfrak{t}^* is represented by the inverse matrix Π^{-1} , and Proposition 3.4 therefore tells us that

$$\|\mathfrak{F}\|^2 = \vec{\mathfrak{F}} \cdot \Pi^{-1} \vec{\mathfrak{F}} = 16\pi^2 |\partial P|^2 \vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}}.$$

Since the first Chern class is Poincaré dual to the homology class of $\cup C_i$,

$$c_1 \cdot [\omega] = \sum_i C_i \cdot [\omega] = \sum_i \mathcal{A}_i = |\partial P|,$$

while M has volume $|P| = [\omega]^2/2$. Thus

$$\mathcal{A}(\Omega) = \frac{(c_1 \cdot [\omega])^2}{[\omega]^2} + \frac{1}{32\pi^2} \|\mathfrak{F}\|^2 = \frac{|\partial P|^2}{2} \left(\frac{1}{|P|} + \vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}} \right)$$

exactly as claimed. \square

By (4.5), Theorem A is now an immediately immediate corollary.

6. The Abreu Formalism

The proof of Theorem A given in §5 was based on the author's formula [34] for the scalar curvature of Kähler surfaces with isometric S^1 actions. This section will present a different proof, which is based on Abreu's beautiful formula [1] for the scalar curvature of a toric manifold, and makes crucial use of an integration-by-parts trick due to Donaldson [21]. While this second proof is certainly more elegant and natural, there are unfortunately many numerical factors involved in this formalism that are typically misreported in the literature, and we will need to correct these imprecisions in order to obtain our result. This will be well worth the effort,

however, insofar as this second proof works equally well in all complex dimensions. The reader should note, however, that the higher-dimensional version of Theorem A is of much less differential-geometric interest than the corresponding statement in complex dimension 2; it is only in real dimension 4 that the Calabi energy is intimately tied to the Weyl functional and conformally Einstein metrics.

We thus begin by considering a toric manifold (M^{2m}, J, \mathbf{T}) of complex dimension m , equipped with a Kähler metric g which is invariant under the action of the m -torus $\mathbf{T} \cong T^m$. Choosing an isomorphism $\mathbf{T} \cong \mathbb{R}^m / \mathbb{Z}^m$, we then let (ξ_1, \dots, ξ_m) be the m unit-period vector fields generating \mathbf{T} associated with this choice, and let (Ξ_1, \dots, Ξ_m) be the holomorphic vector fields defined by $\Xi_j = \xi_j^{1,0}$. Let (x_1, \dots, x_m) be Hamiltonians for (ξ_1, \dots, ξ_m) , and note that these are consequently also holomorphy potentials for (Ξ_1, \dots, Ξ_m) . The function $\vec{x} : M \rightarrow \mathbb{R}^m$ given by (x_1, \dots, x_m) is then a *moment map* for this T^m -action, and its image $\vec{x}(M)$ is called the associated *moment polytope*. Once again, the moment polytope has the *Delzant property*: a neighborhood of any vertex $\in P$ can be transformed into a neighborhood of $\vec{0} \in [0, \infty)^m$ by an element of $\mathbf{SL}(m, \mathbb{Z}) \ltimes \mathbb{R}^m$. The $2m$ -dimensional volume measure on M now pushes forward, by integration on the fibers, to the standard m -dimensional Euclidean measure on \mathbb{R}^m , which we will again denote by $d\alpha$ to emphasize our special interest in the case of $m = 2$. The boundary ∂P is the image of a union of toric complex hypersurfaces in M , and the push-forward of $(2m - 2)$ -dimensional Riemannian measure induces an $(m - 1)$ -dimensional measure $d\lambda$ on ∂P which, on each face, is $\mathbf{SL}(n, \mathbb{Z})$ -equivalent to the standard $(m - 1)$ -dimensional Euclidean measure on the hyperplane $x_1 = 0$.

For consistency with [1, 21], it will be convenient to also consider the vector fields $\eta_j = \xi_j / 2\pi$ of period 2π , and their Hamiltonians $t^j = x_j / 2\pi$; the corresponding moment map is then $\vec{t} = (t^1, \dots, t^m)$, and its image $\tilde{P} = \vec{t}(M)$ can then be transformed into P by dilating by a factor of 2π . Following Donaldson, we will use $d\mu$ to denote m -dimensional Euclidean measure on \tilde{P} , and $d\sigma$ to denote the $(m - 1)$ -dimensional measure on $\partial \tilde{P}$ which, on each face, is $\mathbf{SL}(n, \mathbb{Z})$ -equivalent to $(m - 1)$ -dimensional Euclidean measure on the hyperplane $t^1 = 0$. Identifying \tilde{P} with P via the obvious homothety, we thus have $d\alpha = (2\pi)^m d\mu$ and $d\lambda = (2\pi)^{m-1} d\sigma$.

On the open dense set $\vec{t}^{-1}(\text{Int } \tilde{P}) \subset M$, Abreu observed that our T^m -invariant Kähler metric can be expressed as

$$g = V_{jk} dt^j \otimes dt^k + V^{jk} d\vartheta_j \otimes d\vartheta_k$$

where $V : \tilde{P} \rightarrow \mathbb{R}$ is a convex potential function, $[V_{jk}]$ is the Hessian matrix of V , $[V^{jk}]$ is its inverse matrix, and the ϑ_j are standard angle coordinates on $T^m = S^1 \times \cdots \times S^1$. The potential V is Legendre dual to a Kähler potential for g ; it is continuous on \tilde{P} and smooth in its interior. Moreover, it satisfies the so-called Guillemin-Abreu boundary condition: near a face

given by $L = 0$, where the affine linear function $L : \mathbb{R}^m \rightarrow \mathbb{R}$ is non-negative on \tilde{P} and where dL is an indivisible element of the integer lattice $(\mathbb{Z}^m)^*$, V differs from $\frac{1}{2}L \log L$ by a smooth function. (Note that the factor of $1/2$ is missing from [21, p. 303], and will lead to a compensating correction below.) The scalar curvature s of g is then expressible in terms of V via Abreu's beautiful formula [1, 21]

$$(6.1) \quad s = -(V^{jk})_{,jk} := - \sum_{j,k=1}^m \frac{\partial^2 V^{jk}}{\partial t_j \partial t_k},$$

where we have followed Donaldson's conventions in order to give s its standard Riemannian value.

In this setting, Donaldson [21, Lemma 3.3.5] derives the integration-by-parts formula

$$(6.2) \quad \int_{\tilde{P}} V^{jk} f_{,jk} d\mu = \int_{\tilde{P}} (V^{jk})_{,jk} f d\mu + 2 \int_{\partial \tilde{P}} f d\sigma$$

for any convex function f . Note, however, that the factor of 2 in front of the boundary term does not actually appear in [21], but is needed to compensate for the factor of $1/2$ in the corrected Abreu-Guillemain boundary conditions. We also give the boundary term a different sign, because we are treating $d\sigma$ as a measure rather than as an exterior differential form.

Example Let (M, g) be the unit 2-sphere, with sectional curvature $K = 1$, and hence with scalar curvature $s = 2K = 2$. Equip (M, g) with the S^1 action given by period- 2π rotation around the z -axis, with Hamiltonian $t = z$ and moment polytope $\tilde{P} = [-1, 1]$. In cylindrical coordinates, our metric becomes

$$g = \frac{dt^2}{1-t^2} + (1-t^2)d\vartheta^2$$

so that the potential V must satisfy $V_{,11} = 1/(1-t^2)$ and $V^{,11} = 1-t^2$. A suitable choice of V is therefore

$$V = \frac{1}{2}(1+t)\log(1+t) + \frac{1}{2}(1-t)\log(1-t)$$

and we note that this satisfies the Guillemain-Abreu boundary conditions discussed above. The Abreu formula (6.1) now correctly calculates the scalar curvature

$$s = -(V^{,11})_{,11} = -\frac{d^2}{dt^2}(1-t^2) = 2$$

of g . Also notice that integration by parts gives

$$\int_{-1}^1 (1-t^2)f'' dt = \int_{-1}^1 (1-t^2)''f dt + 2[f(-1) + f(1)]$$

as predicted by (6.2). ◊

Example Let (M^{2m}, g) be the Riemannian product $S^2 \times \cdots \times S^2$ of m copies of the unit 2-sphere, with equipped with the product T^m -action. The moment polytope is now the m -cube $\tilde{P} = [-1, 1]^m$, and the metric is again represented by a symplectic potential

$$V = \frac{1}{2} \sum_j [(1 + t^j) \log(1 + t^j) + (1 - t^j) \log(1 - t^j)]$$

which satisfies our corrected Guillemin-Abreu boundary conditions. The Abreu formula (6.1) now predicts that the scalar curvature of g is

$$s = -(V^{ij})_{ij} = -\sum_j \frac{\partial^2}{\partial(t^j)^2} (1 - t_j^2) = 2m ,$$

in agreement with the additivity of the scalar curvature under Riemannian products. Integrating the j^{th} term by parts twice in the j^{th} variable, we have

$$\int_{\tilde{P}} \left[\sum_j [1 - (t^j)^2] \frac{\partial^2 f}{\partial(t^j)^2} \right] d\mu = \int_{\tilde{P}} \left[\sum_j \frac{\partial^2 [1 - (t^j)^2]}{\partial(t^j)^2} \right] f d\mu + 2 \int_{\partial\tilde{P}} f d\sigma ,$$

for any smooth f , thereby double-checking (6.2) in complex dimension m . \diamondsuit

By linearity, (6.2) also holds [21, Corollary 3.3.10] if f is any difference of convex functions. In particular, (6.2) applies to any affine linear function f on \mathbb{R}^m ; and since any such f satisfies $f_{jk} = 0$, (6.1) and (6.2) tell us that

$$0 = \int_{\tilde{P}} (-s)f d\mu + 2 \int_{\partial\tilde{P}} f d\sigma$$

for any affine-linear function. Applying the dilation that relates \tilde{P} and P , we therefore obtain

$$(6.3) \quad \int_P s f da = 4\pi \int_{\partial P} f d\lambda$$

for any affine-linear f . In particular, if we take $f = x_k - \bar{x}_k$, we obtain

$$\int_P x_k(s - \bar{s}) da = \int_P (x_k - \bar{x}_k)s da = 4\pi \int_{\partial P} (x_k - \bar{x}_k) d\lambda$$

which in turn implies that

$$\int_M x_k(s - \bar{s}) d\mu = 4\pi \int_{\partial P} (x_k - \bar{x}_k) d\lambda$$

because da is the push-forward of the volume measure of (M, g) . However, x_k is a holomorphy potential for the holomorphic vector field Ξ_k , so (2.1) tells us that the component

$$\mathfrak{F}_k := \mathfrak{F}(\Xi_k, \Omega)$$

of the Futaki invariant is given by

$$\mathfrak{F}_k = -4\pi \int_{\partial P} (x_k - \bar{x}_k) d\lambda .$$

On the other hand,

$$\frac{1}{|\partial P|} \int_{\partial P} (x_k - \bar{x}_k) d\lambda = \langle x_k - \bar{x}_k \rangle = \langle x_k \rangle - \bar{x}_k = \mathfrak{D}_k$$

where $|\partial P|$ denotes the λ -measure of the boundary, $\langle \cdot \rangle$ is the average with respect to $d\lambda$, and where \mathfrak{D}_k is the k^{th} component of the vector \mathfrak{D} which points from the barycenter of P to the barycenter of ∂P . Thus the Futaki invariant $\mathfrak{F}(\Omega) = \vec{\mathfrak{F}} = (\mathfrak{F}_1, \dots, \mathfrak{F}_m)$ is given by

$$(6.4) \quad \vec{\mathfrak{F}} = -4\pi |\partial P| \vec{\mathfrak{D}}$$

and we have thus reproved Proposition 5.2 in arbitrary complex dimension m .

Now notice that, by taking normalized Hamiltonians, the Lie algebra \mathfrak{t} of our maximal torus \mathbf{T} is naturally identified with those affine-linear functions $\mathbb{R}^m \rightarrow \mathbb{R}$ which send the barycenter \bar{x} of our moment polytope to 0. From this view-point, it is now apparent that $\mathfrak{F}(\Omega) = -4\pi |\partial P| \vec{\mathfrak{D}}$ actually belongs to \mathfrak{t}^* , as it should. In these same terms, though, the “moment-of inertia” matrix Π defined by

$$(6.5) \quad \Pi_{jk} = \int_P (x_j - \bar{x}_j)(x_k - \bar{x}_k) da$$

represents the L^2 inner product

$$\mathbb{B}_{\mathbf{T}} : \mathfrak{t} \times \mathfrak{t} \rightarrow \mathbb{R} ,$$

while its inverse matrix Π^{-1} represents the dual inner product

$$\mathbb{B}_{\mathbf{T}}^{-1} : \mathfrak{t}^* \times \mathfrak{t}^* \rightarrow \mathbb{R} .$$

By Corollary 2.3 and (6.4), we thus have

$$\|\mathfrak{F}(\Omega)\|^2 = \mathbb{B}_{\mathbf{T}}^{-1}(\mathfrak{F}(\Omega), \mathfrak{F}(\Omega)) = 16\pi^2 |\partial P|^2 \vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}} .$$

Chen’s inequality (4.1) therefore tells us² that any Kähler metric on a toric manifold satisfies

$$\int_M (s - \bar{s})^2 d\mu \geq 16\pi^2 |\partial P|^2 \vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}}$$

where the moment polytope P is determined solely by the toric manifold M and the Kähler class Ω ; moreover, equality holds iff g is extremal.

On the other hand, setting $f = 1$ in (6.3) yields

$$\int_P s da = 4\pi \int_{\partial P} d\lambda ,$$

²Here it is worth reiterating that, while the inequality (4.1) is essentially elementary when g is \mathbf{T} -invariant, it is a deep and remarkable result that this same inequality in fact holds for completely arbitrary Kähler metrics.

so that

$$\int_M s \, d\mu = 4\pi |\partial P| ,$$

a fact which the reader may enjoy comparing with (4.3). Since (M, g) has volume $|P|$, we therefore see that

$$\int_M \bar{s}^2 \, d\mu = \frac{(\int_M s \, d\mu)^2}{\int_M d\mu} = 16\pi^2 \frac{|\partial P|^2}{|P|}$$

and the Pythagorean theorem (4.2) therefore implies the following result:

THEOREM B. *Let $(M^{2m}, J, \Omega, \mathbf{T})$ be a toric complex m -manifold with fixed Kähler class, and let $P \subset \mathbb{R}^m$ be the associated moment polytope. Then the scalar curvature s of any Kähler metric g with Kähler form $\omega \in \Omega$ satisfies*

$$(6.6) \quad \frac{1}{16\pi^2} \int_M s^2 d\mu_g \geq |\partial P|^2 \left(\frac{1}{|P|} + \vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}} \right) ,$$

with equality iff g is an extremal Kähler metric. Here $|P|$ denotes the m -volume of the interior of P , $|\partial P|$ is the λ -volume of its boundary, the moment-of-inertia matrix Π of P is defined by (6.5), and $\vec{\mathfrak{D}}$ is the vector joining the barycenter P to the barycenter of ∂P .

Specializing to the case of $m = 2$ gives a second proof of Theorem A.

Notice that the sharp lower bound (6.6) is in fact independent of dimension. However, this feature of the result actually depends on our conventions regarding the moment polytope and the generators of the action. For example, if we had instead chosen the periodicity of our generators to be 2π instead of 1, we would have been led to instead use the polytope \tilde{P} , and we would have then been forced to introduce an inconvenient scaling factor, since

$$\frac{|\partial P|^2}{|P|} = (2\pi)^{m-2} \frac{|\partial \tilde{P}|^2}{|\tilde{P}|}$$

But it is also worth noticing that this awkward scaling factor magically disappears when $m = 2$. This reflects the fact that the Calabi energy is invariant under rescaling in real dimension four, and that rescaling a Kähler class exactly results in a rescaling of the associated moment polytope.

In particular, for the purpose of calculating the virtual action \mathcal{A} for toric surfaces, we would have obtained exactly the same formula if we had used the rescaled polygon \tilde{P} instead of the polygon P emphasized by this article. Nonetheless, the use of P has other practical advantages, even when $m = 2$. For example, the λ -length of sides of P directly represents the areas of holomorphic curves in M , unmediated by factors of 2π . In practice, this avoids repeatedly having to cancel powers of 2π when calculating $\mathcal{A}(\Omega)$ in explicit examples. This will now become apparent, as we next illustrate Theorem A by applying it to specific toric surfaces.

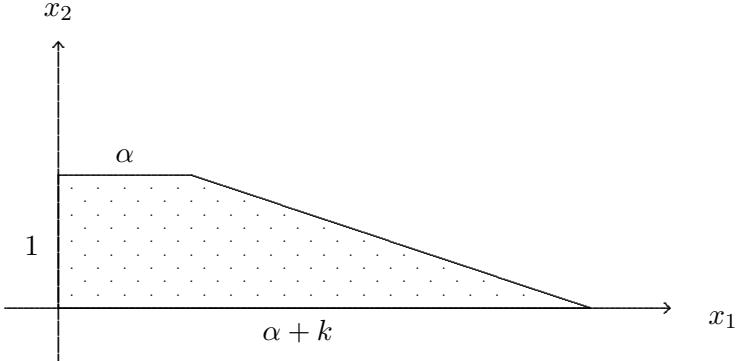


FIGURE 1

7. Hirzebruch Surfaces

As a simple illustration of Theorem 5.1, we now compute $\mathcal{A}(\Omega)$ for the Hirzebruch surfaces. Recall [7, 26] that, for any non-negative integer k , the k^{th} Hirzebruch surface \mathbb{F}_k is defined to be the \mathbb{CP}_1 -bundle $\mathbb{P}(\mathcal{O}(k) \oplus \mathcal{O})$ over \mathbb{CP}_1 ; that is, it is the complex surface obtained from line bundle $\mathcal{O}(k) \rightarrow \mathbb{CP}_1$ of Chern class k by adding a section at infinity. Calabi [11] explicitly constructed an extremal every Kähler metric in every Kähler class on each \mathbb{F}_k ; his direct assault on the problem proved feasible because the maximal compact subgroup $\mathbf{U}(2)/\mathbb{Z}_k$ of the automorphism group has orbits of real codimension 1, thereby reducing the relevant equation for the Kähler potential to an ODE. Because their automorphism groups all contain finite quotients of $\mathbf{U}(2)$, the Hirzebruch surfaces all admit actions of the 2-torus T^2 , and so are toric surfaces. Normalizing the fibers of $\mathbb{F}_k \rightarrow \mathbb{CP}_1$ to have area 1, the associated moment polygon becomes the trapezoid shown in Figure 1 and since $\mathcal{A}(\Omega)$ is unchanged by multiplying Ω by a positive constant, we may impose this normalization without loss of generality.

We will now apply Theorem 5.1 to calculate the Calabi energy of Calabi's extremal Kähler metrics; since Hwang and Simanca [30] have previously computed this quantity by other means, this exercise will, among other things, provide us with another useful double-check of equation (5.1). The area and λ -perimeter of the polygon are easily seen to be

$$|P| = \alpha + \frac{k}{2}, \quad |\partial P| = 2 + 2\alpha + k$$

and it is not difficult to calculate the barycenter of the interior

$$\bar{x} = \frac{(3\alpha^2 + 3k\alpha + k^2, 3\alpha + k)}{6|P|}$$

or boundary

$$\langle \vec{x} \rangle = \frac{(\alpha^2 + \alpha(k+1) + \frac{1}{2}k(k+1), \alpha+1)}{|\partial P|}$$

by hand. The vector

$$\vec{\mathfrak{D}} = \frac{k(2\alpha+k-1)}{12|\partial P||P|} (k, -2)$$

thus joins these two barycenter, and without too much work one can also check that the “moment-of inertia” matrix of P is given by

$$\Pi = \frac{1}{72|P|} \begin{bmatrix} 6\alpha^4 + 12\alpha^3k + 12\alpha^2k^2 + 6\alpha k^3 + k^4 & -\frac{k}{2}(6\alpha^2 + 6\alpha k + k^2) \\ -\frac{k}{2}(6\alpha^2 + 6\alpha k + k^2) & 6\alpha^2 + 6\alpha k + k^2 \end{bmatrix}$$

The Futaki contribution to \mathcal{A} is therefore encoded by the expression

$$\vec{\mathfrak{D}} \cdot \Pi^{-1} \vec{\mathfrak{D}} = \frac{2k^2(2\alpha+k-1)^2}{|P||\partial P|^2(6\alpha^2 + 6\alpha k + k^2)}$$

and the virtual action is thus given by

$$(7.1) \quad \mathcal{A}(\Omega) = \frac{2\alpha^3 + (4+3k)\alpha^2 + 2(1+k)^2\alpha + k(k^2+2)/2}{\alpha^2 + \alpha k + k^2/6}.$$

After multiplication by an overall constant and the change of variables $k = n$, $\alpha = (a-n)/2$, this agrees with the expression Hwang and Simanca [30, equation (3.2)] obtained for their “potential energy” via a different method.

For $k > 0$, the function $\mathcal{A}(\alpha)$ on the right-hand side of (7.1) extends smoothly across $\alpha = 0$, and satisfies

$$\left. \frac{d\mathcal{A}}{d\alpha} \right|_{\alpha=0} = -6 \frac{(k-2)^2}{k}$$

so $\mathcal{A}(\alpha)$ is a decreasing function for small α if $k \neq 2$. On the other hand, $\mathcal{A}(\alpha) \sim 2\alpha$ for $\alpha \gg 0$, so \mathcal{A} is increasing for large α . It follows that $\mathcal{A}(\alpha)$ has a minimum somewhere on \mathbb{R}^+ for any $k \neq 2$. Since Calabi’s construction [11] moreover shows that each Kähler class on a Hirzebruch surface is represented by an extremal Kähler metric, Proposition 4.1 tells us that, for $k \neq 2$, the Calabi metric g_k corresponding to the minimizing value of α is necessarily Bach-flat.

On the other hand, since

$$\mathcal{A}(\Omega) - \frac{3}{4}k = \frac{48\alpha^3 + (54k+96)\alpha^2 + (30k^2+96k+48)\alpha + 9k^3 + 24k}{4(6\alpha^2 + 6\alpha k + k^2)}$$

is positive for all $\alpha > 0$, it follows that

$$\min_{\Omega} \mathcal{A}(\Omega) > \frac{3}{4}k,$$

and we conclude that the corresponding Bach-flat Kähler metric g_k has

$$\mathcal{W}(g_k) > 2\pi^2 k$$

Since the Hirzebruch surface \mathbb{F}_k is diffeomorphic to $S^2 \times S^2$ when k is even, and is diffeomorphic to $\mathbb{CP}_2 \# \overline{\mathbb{CP}}_2$ when k is odd, the metrics g_k , first discovered by Hwang and Simanca [30], immediately give us the following:

PROPOSITION 7.1. *The smooth 4-manifolds $S^2 \times S^2$ and $\mathbb{CP}_2 \# \overline{\mathbb{CP}}_2$ both admit sequences of Bach-flat conformal classes $[g_{k_j}]$ with $\mathcal{W}([g_{k_j}]) \rightarrow +\infty$. Consequently, the moduli space of Bach-flat conformal metrics on either of these manifolds has infinitely many connected components.*

The metric g_1 on \mathbb{F}_1 has scalar curvature $s > 0$ everywhere, and its conformal rescaling $s^{-2}g_1$ was shown by Derdziński [19] to coincide with the Einstein metric on $\mathbb{CP}_2 \# \overline{\mathbb{CP}}_2$ discovered by Page [48]. For $k \geq 3$, the scalar curvature s of g_k instead vanishes along a hypersurface, which becomes the conformal infinity for the Einstein metric $s^{-2}g_k$; thus \mathbb{F}_k is obtained from two Poincaré-Einstein manifolds, glued along their conformal infinity. These two Einstein metrics are in fact isometric, in an orientation-reversing manner. Because of their $\mathbf{U}(2)$ symmetry, these Einstein metrics belong to the family first discovered by Bérard-Bergery [8], and later rediscovered by physicists, who call them AdS-Taub-bolt metrics [29].

8. The Two-Point Blow-Up of \mathbb{CP}_2

As a final illustration of Theorem 5.1, we now compute the virtual action for Kähler classes on the blow-up of \mathbb{CP}_2 at two distinct points. The present author has done this elsewhere by a more complicated method, and the details of the answer played an important role in showing [15, 37] that this manifold admits an Einstein metric, obtained by conformally rescaling a Bach-flat Kähler metric. Thus, repeating the computation by means of equation (5.1) provides yet another double-check of Theorem A.

Blowing up \mathbb{CP}_2 in two distinct points results in exactly the same complex surface as blowing $\mathbb{CP}_1 \times \mathbb{CP}_1$ in a single point [7, 26]. The latter

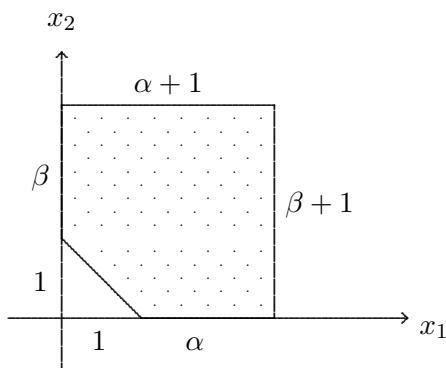


FIGURE 2

picture is actually useful in choosing a pair of generators for the torus action which makes the needed computations as simple as possible. The resulting moment polygon P then takes the form shown in Figure 2 after rescaling to give the blow-up divisor area 1. It is then easy to see that the area of the polygon and the λ -length of its boundary are given by

$$|P| = \frac{1}{2} + \alpha + \beta + \alpha\beta, \quad |\partial P| = 3 + 2\alpha + 2\beta$$

while the barycenter of the boundary

$$\langle \vec{x} \rangle = \frac{\left((1+\alpha)(2+\alpha+\beta), \quad (1+\beta)(2+\alpha+\beta) \right)}{|\partial P|}$$

and of the interior

$$\bar{\vec{x}} = \frac{\left(3(1+\alpha)^2(1+\beta) - 1, \quad 3(1+\alpha)(1+\beta)^2 - 1 \right)}{6|P|}$$

are not difficult to compute by hand. The vector joining these two barycenters is thus given by

$$\vec{\mathcal{D}} = \frac{\left(-\alpha + 2\beta + 3\alpha\beta + 3\alpha^2\beta, \quad -\beta + 2\alpha + 3\alpha\beta + 3\alpha\beta^2 \right)}{6|P| |\partial P|}$$

and the moment-of-inertia matrix

$$\Pi = \frac{1}{24} \begin{bmatrix} 8(1+\alpha)^3(1+\beta) - 2 & 6(1+\alpha)^2(1+\beta)^2 - 1 \\ 6(1+\alpha)^2(1+\beta)^2 - 1 & 8(1+\alpha)(1+\beta)^3 - 2 \end{bmatrix} - |P| \begin{bmatrix} \bar{x}_1^2 & \bar{x}_1 \bar{x}_2 \\ \bar{x}_1 \bar{x}_2 & \bar{x}_2^2 \end{bmatrix}$$

are also easily obtained without the use of a computer. According to (5.1), $\mathcal{A}(\Omega)$ is therefore given by

$$\begin{aligned} & 3 \left[3 + 28\beta + 96\beta^2 + 168\beta^3 + 164\beta^4 + 80\beta^5 + 16\beta^6 + 16\alpha^6(1+\beta)^4 + 16\alpha^5(5+24\beta + 43\beta^2 + 37\beta^3 + 15\beta^4 + 2\beta^5) + 4\alpha^4(41+228\beta + 478\beta^2 + 496\beta^3 + 263\beta^4 + 60\beta^5 + 4\beta^6) + 8\alpha^3(21+135\beta + 326\beta^2 + 392\beta^3 + 248\beta^4 + 74\beta^5 + 8\beta^6) + 4\alpha(7+58\beta + 176\beta^2 + 270\beta^3 + 228\beta^4 + 96\beta^5 + 16\beta^6) + 4\alpha^2(24+176\beta + 479\beta^2 + 652\beta^3 + 478\beta^4 + 172\beta^5 + 24\beta^6) \right] / \\ & \left[1 + 10\beta + 36\beta^2 + 64\beta^3 + 60\beta^4 + 24\beta^5 + 24\alpha^5(1+\beta)^5 + 12\alpha^4(1+\beta)^2(5+20\beta + 23\beta^2 + 10\beta^3) + 16\alpha^3(4+28\beta + 72\beta^2 + 90\beta^3 + 57\beta^4 + 15\beta^5) + 12\alpha^2(3+24\beta + 69\beta^2 + 96\beta^3 + 68\beta^4 + 20\beta^5) + 2\alpha(5+45\beta + 144\beta^2 + 224\beta^3 + 180\beta^4 + 60\beta^5) \right] \end{aligned}$$

as is most easily checked at this point using *Mathematica* or a similar program. After the substitution $\gamma = \alpha$, this agrees exactly with the answer obtained in [38, §2], where this explicit formula plays a key role in classifying compact Einstein 4-manifolds for which the metric is Hermitian with respect to some complex structure.

When $\alpha = \beta$, the above expression simplifies to become

$$\frac{9 + 96\alpha + 396\alpha^2 + 840\alpha^3 + 954\alpha^4 + 528\alpha^5 + 96\alpha^6}{1 + 12\alpha + 54\alpha^2 + 120\alpha^3 + 138\alpha^4 + 72\alpha^5 + 12\alpha^6}$$

which, after dividing by 3 and making the substitution $\alpha = 1/y$, coincides with the expression [36] first used to show that \mathcal{A} has a critical point, and later used again [15] to prove the existence of a conformally Einstein, Kähler metric on $\mathbb{CP}_2 \# 2\overline{\mathbb{CP}}_2$. For a second, conceptually simpler proof of this last fact, see [37].

Appendix A. Restricting the Futaki Invariant

In this appendix, we will prove Proposition 2.1. The key ingredient used in the proof is the following result of Nakagawa [45]:

PROPOSITION A.1 (Nakagawa). *Let (M, J) be a projective algebraic complex manifold, let \mathbf{H} be the identity component of its complex automorphism group, and suppose that the Jacobi homomorphism from \mathbf{H} to the Albanese torus of M is trivial. Let $L \rightarrow M$ be an ample line bundle for which the action of \mathbf{H} on M lifts to an action on $L \rightarrow M$, and let Ω be the Kähler class defined by $\Omega = c_1(L)$. Then the Futaki invariant $\mathfrak{F}(\Omega) \in \mathfrak{h}^*$ annihilates the Lie algebra $\mathfrak{r}_{\mathbf{u}}$ of the unipotent radical of \mathbf{H} .*

This generalizes a previous result of Mabuchi [42] concerning the case when L is the anti-canonical line bundle. Both of these results are proved using Tian's localization formula [53] for the Futaki invariant of a Hodge metric.

We will now extend Proposition A.1 to irrational Kähler classes on certain complex manifolds. In order to do this, we will first need the following observation:

LEMMA A.2. *Let (M, J) be a compact complex manifold with $b_1(M) = 0$, and let \mathbf{H} be the identity component of its complex automorphism group. If $L \rightarrow M$ is a positive line bundle, then the action of \mathbf{H} on M lifts to an action on $L^k \rightarrow M$ for some positive integer k .*

PROOF. By the Kodaira embedding theorem [26], L has a positive power L^ℓ for which there is a canonical holomorphic embedding $j : M \hookrightarrow \mathbb{P}(\mathbb{V})$ such that $j^*\mathcal{O}(-1) = L^{-\ell}$, where $\mathbb{V} := [H^0(M, \mathcal{O}(L^\ell))]^*$.

Now since (M, J) is of Kähler type and $H^1(M, \mathbb{C}) = 0$, the Hodge decomposition tells us that $H^{0,1}(M) = H^1(M, \mathcal{O}) = 0$, and the long exact sequence

$$\cdots \rightarrow H^1(M, \mathcal{O}) \rightarrow H^1(M, \mathcal{O}^\times) \rightarrow H^2(M, \mathbb{Z}) \rightarrow \cdots$$

therefore implies that holomorphic line bundles on M are classified by their first Chern classes. On the other hand, since \mathbf{H} is connected, each automorphism $\Phi : M \rightarrow M$, $\Phi \in \mathbf{H}$, is homotopic to the identity; and since Chern classes are homotopy invariants, we deduce that $c_1(\Phi^*L) = c_1(L)$ for all $\Phi \in \mathbf{H}$. Consequently, $\Phi^*L \cong L$ as a holomorphic line bundle for any $\Phi \in \mathbf{H}$. While the resulting isomorphism $\Phi^*L \cong L$ is not unique, any two such isomorphisms merely differ by an overall multiplicative constant, and the associated linear map $H^0(M, \mathcal{O}(L^\ell)) \rightarrow H^0(M, \mathcal{O}(L^\ell))$ induced by

Φ^* is therefore completely determined up to an overall scale factor. Thus, for every $\Phi \in \mathbf{H}$, there is a uniquely determined projective transformation $\mathbb{P}(\mathbb{V}) \rightarrow \mathbb{P}(\mathbb{V})$, where again $\mathbb{V} := [H^0(M, \mathcal{O}(L^\ell))]^*$. This gives us a faithful projective representation $\mathbf{H} \hookrightarrow \mathbf{PSL}(\mathbb{V})$ which acts on $M \subset \mathbb{P}(\mathbb{V})$ via the original action of \mathbf{H} .

Now consider the group $\mathbf{SL}(\mathbb{V})$ of unit-determinant linear endomorphisms of \mathbb{V} , and observe that there is a short exact sequence

$$0 \rightarrow \mathbb{Z}_n \rightarrow \mathbf{SL}(\mathbb{V}) \rightarrow \mathbf{PSL}(\mathbb{V}) \rightarrow 1$$

where $n = \dim \mathbb{V}$; that is, every projective transformation of $\mathbb{P}(\mathbb{V})$ arises from n different linear unit-determinant linear endomorphisms of \mathbb{V} , differing from each other merely by multiplication by an n^{th} root of unity. If $\tilde{\mathbf{H}} < \mathbf{SL}(\mathbb{V})$ is the inverse image of $\mathbf{H} < \mathbf{PSL}(\mathbb{V})$, then $\tilde{\mathbf{H}}$ acts on \mathbb{V} , and so also acts on the tautological line bundle $\mathcal{O}(-1)$ over $\mathbb{P}(\mathbb{V})$. Restricting $\mathcal{O}(-1)$ to M then gives us an action of $\tilde{\mathbf{H}}$ on $L^{-\ell}$ which lifts the action of \mathbf{H} on M , in such a manner that any two lifts of a given element only differ by multiplication of an n^{th} root of unity. The induced action of $\tilde{\mathbf{H}}$ on $L^{-n\ell}$ therefore descends to an action of \mathbf{H} , and passing to the dual line bundle $L^{n\ell}$ thus shows that the action of \mathbf{H} on M can be lifted to an action on $L^k \rightarrow M$ for $k = n\ell$. \square

PROPOSITION A.3. *Let (M, J) be a compact complex manifold of Kähler type, and suppose that M does not carry any non-trivial holomorphic 1- or 2-forms. Then, for any Kähler class Ω on M , the Futaki invariant $\mathfrak{F}(\Omega) \in \mathfrak{h}^*$ annihilates the unipotent radical $\mathfrak{r}_{\mathfrak{u}} \subset \mathfrak{h}$.*

PROOF. By hypothesis, $H^{1,0}(M) = H^{2,0}(M) = 0$. The Hodge decomposition therefore tells us that $b_1(M) = 0$ and that $H^{1,1}(M, \mathbb{R}) = H^2(M, \mathbb{R})$. Consequently, the Kähler cone $\mathcal{K} \subset H^{1,1}(M, \mathbb{R})$ is open in $H^2(M, \mathbb{R})$. Since $H^2(M, \mathbb{Q})$ is dense in $H^2(M, \mathbb{R})$, it follows that $H^2(M, \mathbb{Q}) \cap \mathcal{K}$ is dense in \mathcal{K} . In particular, $H^2(M, \mathbb{Q}) \cap \mathcal{K}$ is non-empty, and so, clearing denominators, we conclude that the Kähler cone \mathcal{K} must meet the integer lattice $H^2(M, \mathbb{Z})/\text{torsion} \subset H^2(M, \mathbb{R})$. This argument, due to Kodaira [33], shows that (M, J) carries Kähler metrics of Hodge type, and is therefore projective algebraic.

Pursuing this idea in the opposite direction, let Ψ now be an integral Kähler class, so that $\Psi = c_1(L)$ for some positive line bundle $L \rightarrow M$. By Lemma A.2, the action of \mathbf{H} on M then lifts to some positive power L^k of L . Since our hypotheses also imply that the Albanese torus is trivial, Proposition A.1 therefore implies that $\mathfrak{F}(k\Psi) \in \mathfrak{h}^*$ annihilates $\mathfrak{r}_{\mathfrak{u}}$. However, our expression (2.1) for the Futaki invariant implies that

$$\mathfrak{F}(\Xi, \lambda\Omega) = \lambda^m \mathfrak{F}(\Xi, \Omega)$$

for any $\lambda \in \mathbb{R}^+$, where m is the complex dimension, since rescaling a Kähler metric by $g \rightsquigarrow \lambda g$ results in $\omega \rightsquigarrow \lambda\omega$, $s \rightsquigarrow \lambda^{-1}s$, $f \rightsquigarrow \lambda f$, and $d\mu \rightsquigarrow \lambda^m d\mu$. By taking λ to be an arbitrary positive rational, we therefore see that $\mathfrak{F}(\Xi, \Omega) = 0$ whenever $\Xi \in \mathfrak{r}_{\mathfrak{u}}$ and $\Omega \in H^2(M, \mathbb{Q}) \cap \mathcal{K}$, where \mathcal{K} once again

denotes the Kähler cone. However, for any fixed Ξ , the right-hand-side of (2.1) clearly depends smoothly on the Kähler metric g , and $\mathfrak{F}(\Xi, \Omega)$ therefore is a smooth function of the Kähler class Ω . But $h^{2,0}(M) = 0$ implies that $H^2(M, \mathbb{Q}) \cap \mathcal{K}$ is dense in \mathcal{K} . Thus, for any $\Xi \in \mathfrak{r}_u$, we have shown that $\mathfrak{F}(\Xi, \Omega) = 0$ for a dense set of $\Omega \in \mathcal{K}$. Continuity therefore implies that $\mathfrak{F}(\Xi, \Omega) = 0$ for all $\Omega \in \mathcal{K}$. Hence $\mathfrak{F}(\Omega) \in \mathfrak{h}^*$ annihilates \mathfrak{r}_u for any Kähler class Ω on M . \square

Under the hypotheses of Proposition A.3, we can thus view $\mathfrak{F}(\Omega)$ as belonging to the complexified Lie coalgebra $\mathfrak{g}_{\mathbb{C}}^* = \mathfrak{g}^* \otimes \mathbb{C}$ of a maximal compact subgroup $\mathbf{G} \subset \mathbf{H}$. By averaging, let us now represent our given Kähler class Ω by a \mathbf{G} -invariant Kähler metric g . The Lie algebra of Killing fields of g then can be identified with the real holomorphy potentials of integral 0, which are their Hamiltonians; the Lie bracket on \mathfrak{g} then becomes the Poisson bracket $\{\cdot, \cdot\}$ on Hamiltonians. Since the scalar curvature s of g is also a real function, formula (2.1) thus tells us that $\mathfrak{F}(\Omega)$ is actually a *real* linear functional on \mathfrak{g} ; that is, $\mathfrak{F}(\Omega) \in \mathfrak{g}^*$. This proves Proposition 2.1.

References

- [1] M. ABREU, Kähler geometry of toric varieties and extremal metrics, *Internat. J. Math.* **9** (1998) 641–651.
- [2] C. AREZZO, F. PACARD, and M. SINGER, Extremal metrics on blowups, *Duke Math. J.* **157** (2011) 1–51.
- [3] M. F. ATIYAH, Convexity and commuting Hamiltonians, *Bull. London Math. Soc.* **14** (1982) 1–15.
- [4] T. AUBIN, Équations du type Monge-Ampère sur les variétés kähleriennes compactes, *C. R. Acad. Sci. Paris Sér. A-B* **283** (1976) Aiiii, A119–A121.
- [5] R. BACH, Zur Weylschen Relativitätstheorie und der Weylschen Erweiterung des Krümmungstensorbegriffs., *Math. Zeitschr.* **9** (1921) 110–135.
- [6] S. BANDO, An obstruction for Chern class forms to be harmonic, *Kodai Math. J.* **29** (2006) 337–345.
- [7] W. BARTH, C. PETERS, and A. VAN DE VEN, *Compact complex surfaces*, volume 4 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*, Springer-Verlag, Berlin, 1984.
- [8] L. BÉRARD-BERGERY, Sur de nouvelles variétés riemanniennes d’Einstein, in *Institut Élie Cartan, 6*, volume 6 of *Inst. Élie Cartan*, pp. 1–60, Univ. Nancy, Nancy, 1982.
- [9] A. L. BESSE, *Einstein manifolds*, volume 10 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*, Springer-Verlag, Berlin, 1987.
- [10] N. BUCHDAHL, On compact Kähler surfaces, *Ann. Inst. Fourier (Grenoble)* **49** (1999) 287–302.
- [11] E. CALABI, Extremal Kähler metrics, in *Seminar on Differential Geometry*, volume 102 of *Ann. Math. Studies*, pp. 259–290, Princeton Univ. Press, Princeton, N.J., 1982.
- [12] E. CALABI, Extremal Kähler metrics. II, in *Differential Geometry and Complex Analysis*, pp. 95–114, Springer, Berlin, 1985.
- [13] X. X. CHEN, Space of Kähler metrics. III. On the lower bound of the Calabi energy and geodesic distance, *Invent. Math.* **175** (2009) 453–503.
- [14] X. X. CHEN and G. TIAN, Geometry of Kähler metrics and foliations by holomorphic discs, *Publ. Math. Inst. Hautes Études Sci.* (2008) 1–107.
- [15] X. CHEN, C. LEBRUN, and B. WEBER, On conformally Kähler, Einstein manifolds, *J. Amer. Math. Soc.* **21** (2008) 1137–1168.

- [16] B. CHEN, A.-M. LI, and L. SHENG, Extremal Metrics on Toric Surfaces, e-print, arXiv:1008.2607v3 [math.DG], 2010.
- [17] C. CHEVALLEY, *Théorie des groupes de Lie. Tome III. Théorèmes généraux sur les algèbres de Lie*, Actualités Sci. Ind. no. 1226, Hermann & Cie, Paris, 1955.
- [18] T. DELZANT, Hamiltoniens périodiques et images convexes de l'application moment, *Bull. Soc. Math. France* **116** (1988) 315–339.
- [19] A. DERDZIŃSKI, Self-dual Kähler manifolds and Einstein manifolds of dimension four, *Compositio Math.* **49** (1983) 405–433.
- [20] S. K. DONALDSON, Scalar curvature and projective embeddings. I, *J. Differential Geom.* **59** (2001) 479–522.
- [21] S. K. DONALDSON, Scalar curvature and stability of toric varieties, *J. Differential Geom.* **62** (2002) 289–349.
- [22] A. FUJIKI, On automorphism groups of compact Kähler manifolds, *Invent. Math.* **44** (1978) 225–258.
- [23] W. FULTON, *Introduction to toric varieties*, volume 131 of *Annals of Mathematics Studies*, Princeton University Press, Princeton, NJ, 1993.
- [24] A. FUTAKI, An obstruction to the existence of Einstein Kähler metrics, *Invent. Math.* **73** (1983) 437–443.
- [25] A. FUTAKI and T. MABUCHI, Uniqueness and periodicity of extremal Kähler vector fields, in *Proceedings of GARC Workshop on Geometry and Topology '93 (Seoul, 1993)*, volume 18 of *Lecture Notes Ser.*, pp. 217–239, Seoul, 1993, Seoul Nat. Univ.
- [26] P. GRIFFITHS and J. HARRIS, *Principles of Algebraic Geometry*, Wiley-Interscience, New York, 1978.
- [27] V. GUILLEMIN, *Moment maps and combinatorial invariants of Hamiltonian T^n -spaces*, volume 122 of *Progress in Mathematics*, Birkhäuser Boston Inc., Boston, MA, 1994.
- [28] V. GUILLEMIN and S. STERNBERG, Convexity properties of the moment mapping, *Invent. Math.* **67** (1982) 491–513.
- [29] S. W. HAWKING, C. J. HUNTER, and D. N. PAGE, NUT charge, anti-de Sitter space, and entropy, *Phys. Rev. D (3)* **59** (1999) 044033, 6.
- [30] A. D. HWANG and S. R. SIMANCA, Extremal Kähler metrics on Hirzebruch surfaces which are locally conformally equivalent to Einstein metrics, *Math. Ann.* **309** (1997) 97–106.
- [31] J. KIM, C. LEBRUN, and M. PONTECORVO, Scalar-flat Kähler surfaces of all genera, *J. Reine Angew. Math.* **486** (1997) 69–95.
- [32] S. KOBAYASHI, Fixed points of isometries, *Nagoya Math. J.* **13** (1958) 63–68.
- [33] K. KODAIRA, On compact complex analytic surfaces. I, *Ann. of Math. (2)* **71** (1960) 111–152.
- [34] C. LEBRUN, Explicit self-dual metrics on $\mathbb{CP}_2 \# \cdots \# \mathbb{CP}_2$, *J. Differential Geom.* **34** (1991) 223–253.
- [35] C. LEBRUN, Anti-self-dual metrics and Kähler geometry, in *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994)*, pp. 498–507, Basel, 1995, Birkhäuser.
- [36] C. LEBRUN, Einstein metrics on complex surfaces, in *Geometry and Physics (Aarhus, 1995)*, volume 184 of *Lecture Notes in Pure and Appl. Math.*, pp. 167–176, Dekker, New York, 1997.
- [37] C. LEBRUN, Einstein Manifolds and Extremal Kähler Metrics, to appear in *Crelle*; e-print arXiv:1009.1270 [math.DG], 2010.
- [38] C. LEBRUN, On Einstein, Hermitian 4-Manifolds, *J. Differential Geom.* **90** (2012) 277–302.
- [39] C. LEBRUN and S. R. SIMANCA, On the Kähler classes of extremal metrics, in *Geometry and Global Analysis (Sendai, 1993)*, pp. 255–271, Tohoku Univ., Sendai, 1993.

- [40] C. LEBRUN and S. R. SIMANCA, Extremal Kähler metrics and complex deformation theory, *Geom. Funct. Anal.* **4** (1994) 298–336.
- [41] T. MABUCHI, Einstein-Kähler forms, Futaki invariants and convex geometry on toric Fano varieties, *Osaka J. Math.* **24** (1987) 705–737.
- [42] T. MABUCHI, An algebraic character associated with the Poisson brackets, in *Recent topics in differential and analytic geometry*, volume 18 of *Adv. Stud. Pure Math.*, pp. 339–358, Academic Press, Boston, MA, 1990.
- [43] T. MABUCHI, Uniqueness of extremal Kähler metrics for an integral Kähler class, *Internat. J. Math.* **15** (2004) 531–546.
- [44] J. MILNOR, *Morse Theory*, volume 51 of *Ann. Math. Studies*, Princeton University Press, Princeton, N.J., 1963, Based on lecture notes by M. Spivak and R. Wells.
- [45] Y. NAKAGAWA, Bando-Calabi-Futaki characters of Kähler orbifolds, *Math. Ann.* **314** (1999) 369–380.
- [46] Y. NAKAGAWA, Bando-Calabi-Futaki character of compact toric manifolds, *Tohoku Math. J. (2)* **53** (2001) 479–490.
- [47] Y. ODAKA, C. SPOTTI, and S. SUN, Compact Moduli Spaces of Del Pezzo Surfaces and Kähler-Einstein Metrics, e-print arXiv:1210.0858 [math.DG], 2012.
- [48] D. PAGE, A Compact Rotating Gravitational Instanton, *Phys. Lett.* **79B** (1979) 235–238.
- [49] E. SHELUKHIN, Remarks on invariants of Hamiltonian loops, *J. Topol. Anal.* **2** (2010) 277–325.
- [50] Y. SIU, Every K3 Surface is Kähler, *Inv. Math.* **73** (1983) 139–150.
- [51] C. H. TAUBES, The existence of anti-self-dual conformal structures, *J. Differential Geom.* **36** (1992) 163–253.
- [52] G. TIAN, On Calabi’s conjecture for complex surfaces with positive first Chern class, *Invent. Math.* **101** (1990) 101–172.
- [53] G. TIAN, Kähler-Einstein metrics on algebraic manifolds, in *Transcendental methods in algebraic geometry (Cetraro, 1994)*, volume 1646 of *Lecture Notes in Math.*, pp. 143–185, Springer, Berlin, 1996.
- [54] S. T. YAU, On the Ricci curvature of a compact Kähler manifold and the complex Monge-Ampère equation. I, *Comm. Pure Appl. Math.* **31** (1978) 339–411.

DEPARTMENT OF MATHEMATICS, SUNY AT STONY BROOK, STONY BROOK, NY 11794-3651, USA

E-mail address: claude@math.sunysb.edu

Mean curvature flows and isotopy problems

Mu-Tao Wang

ABSTRACT. In this note, we discuss the mean curvature flow of graphs of maps between Riemannian manifolds. Special emphasis will be placed on estimates of the flow as a non-linear parabolic system of differential equations. Several global existence theorems and applications to isotopy problems in geometry and topology will be presented. The results are based on joint works of the author with his collaborators I. Medoš, K. Smoczyk, and M.-P. Tsui.

1. Introduction

We start with classical minimal surfaces in \mathbb{R}^3 (see for example [26]). Suppose a surface Σ is given as the graph of a function $f = f(x, y)$ over a domain $\Omega \subset \mathbb{R}^2$:

$$\Sigma = \{(x, y, f(x, y)) \mid (x, y) \in \Omega\}.$$

The area $A(\Sigma)$ is given by the formula

$$A(\Sigma) = \int_{\Omega} \sqrt{1 + |\nabla f|^2}.$$

The Euler-Lagrange equation for the area functional is derived to be

$$(1.1) \quad \operatorname{div}\left(\frac{\nabla f}{\sqrt{1 + |\nabla f|^2}}\right) = 0.$$

Equation (1.1), so called the minimal surface equation, is one of the most studied nonlinear elliptic PDE and there are many beautiful classical results such as the celebrated Bernstein's conjecture for entire solutions [3, 4]. The Dirichlet problem is uniquely solvable as long as the mean curvature of the boundary $\partial\Omega$ is positive [20]. In addition, any Lipschitz solution is smooth and analytic [24, 23].

The author was partially supported by the National Science Foundation under grant DMS-1105483. He would like to thank his collaborators I. Medoš, K. Smoczyk, and M.-P. Tsui.

The corresponding parabolic equation is called the mean curvature flow. Here we have a time-dependent surface Σ_t , given as the graph of a function $f = f(x, y, t)$ for each t , and f satisfies

$$\frac{\partial f}{\partial t} = \sqrt{1 + |\nabla f|^2} \operatorname{div}\left(\frac{\nabla f}{\sqrt{1 + |\nabla f|^2}}\right),$$

This is the negative gradient flow of the area functional. In fact, the normal component of the velocity vector of the graph of $f(x, y, t)$ in \mathbb{R}^3 is exactly the mean curvature vector.

The equation has been extensively studied by many authors such as Huisken [15, 16], Ecker-Huisken [9, 10], Ilmanen [19], Andrews [1], White [41, 42], Huisken-Sinestrari, [17, 18] X.-J. Wang [40], Colding-Minicozzi [8], etc. Note that though the elliptic equation (1.1) is in divergence form, the parabolic equation is not. Therefore, standard results from parabolic PDE theory do not readily apply.

We can also consider the equations in parametric form. Suppose the surface is given by an embedding: $\vec{X}(u, v) = (X_1(u, v), X_2(u, v), X_3(u, v)) \in \mathbb{R}^3$. The minimal surface equation (1.1) is equivalent to

$$(\Delta_\Sigma X_1, \Delta_\Sigma X_2, \Delta_\Sigma X_3) = (0, 0, 0)$$

where Σ is the image surface of \vec{X} and Δ_Σ is the Laplace operator with respect to the induced metric on Σ . In fact, $\vec{H} = \Delta_\Sigma \vec{X}$ is the mean curvature vector of Σ . However, this elegant form has a disadvantage that it is invariant under reparametrization and thus represents a degenerate elliptic system for (X_1, X_2, X_3) . The same phenomenon is encountered for any curvature equation in which the diffeomorphism group appears as the symmetry group.

The corresponding parabolic equation for a family of time-dependent embeddings $\vec{X}(u, v, t)$ is

$$\frac{\partial \vec{X}}{\partial t} = \Delta_\Sigma \vec{X}.$$

For this, the mean curvature flow is often referred as the heat equation for submanifolds, just as the Ricci flow is the heat equation for Riemannian metrics. However, it is clear that the equation is of nonlinear nature as Δ_Σ depends on first derivatives of \vec{X} .

Our subject of study in this note is a submanifold of “higher codimension”, such as a 2-surface in a 4-dimensional space given by the graph of a vector value function (f, g) :

$$\Sigma = \{(x, y, f(x, y), g(x, y)) \mid (x, y) \in \Omega\}.$$

The area of Σ is then

$$A(\Sigma) = \int_{\Omega} \sqrt{1 + |\nabla f|^2 + |\nabla g|^2 + (f_x g_y - f_y g_x)^2}$$

and the Euler-Lagrange equation is a non-linear elliptic system for f and g (see the next paragraph).

In general, we consider a vector-valued function $\vec{f} : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ and Σ is the graph of $\vec{f} = (f^1, \dots, f^m)$ in \mathbb{R}^{n+m} . Denote the induced metric on Σ by

$$g_{ij} = \delta_{ij} + \sum_{\alpha=1}^m \frac{\partial f^\alpha}{\partial x^i} \frac{\partial f^\alpha}{\partial x^j}.$$

The volume of Σ is

$$\int_\Omega \sqrt{\det g_{ij}}$$

and the Euler-Lagrange equation, which is often referred as the minimal surface system, is

$$\sum_{i,j=1}^n g^{ij} \frac{\partial^2 f^\alpha}{\partial x^i \partial x^j} = 0, \quad \alpha = 1, \dots, m,$$

where $g^{ij} = (g_{ij})^{-1}$ is the inverse of g_{ij} .

The corresponding parabolic equation is the mean curvature flow

$$\frac{\partial f^\alpha}{\partial t} = \sum_{i,j=1}^n g^{ij} \frac{\partial^2 f^\alpha}{\partial x^i \partial x^j}, \quad \alpha = 1, \dots, m.$$

There is no reason to stop there and we can consider the even more general situation when $\mathbf{f} : M_1 \rightarrow M_2$ is a differentiable map between Riemannian manifolds, and Σ is the graph of \mathbf{f} in $M_1 \times M_2$ for M_1 an n -dimensional Riemannian manifold and M_2 an m -dimensional one.

In contrast to the codimensional one case, in an article by Lawson-Osserman [21] entitled “Non-existence, non-uniqueness and irregularity of solutions to the minimal surface system”, the undesirable features of the system mentioned in the title are demonstrated. The codimension one case, i.e. $m = 1$, is essentially a scalar equation. In addition, the normal bundle of an oriented hypersurface is always trivial. On the other hand, $m > 1$ corresponds to a genuine systems and the components f^1, \dots, f^m interact with each other. Moreover, the geometry of normal bundle can be rather complicated.

Nevertheless, we managed to obtain estimates and prove several global existence theorems for higher-codimensional mean curvature flows with appropriate initial data. I shall discuss the methods in the next section before presenting the results.

2. Method of proofs

Let us start with the C^1 estimate. In the codimension-one case (see [9] for the equation in a slightly different but equivalent form), $m = 1$, an important equation satisfied by $J_1 = \frac{1}{\sqrt{1+|\nabla f|^2}}$ is

$$\frac{d}{dt} J_1 = \Delta_\Sigma J_1 + R_1(\nabla f, \nabla^2 f).$$

The term $R_1 > 0$ is quadratic in $\nabla^2 f$.

Let us look at the $m = 2, n = 2$ case. We can similarly take

$$J_2 = \frac{1}{\sqrt{1 + |\nabla f|^2 + |\nabla g|^2 + (f_x g_y - f_y g_x)^2}}$$

and compute the evolution equation:

$$\frac{d}{dt} J_2 = \Delta_\Sigma J_2 + R_2(\nabla f, \nabla g, \nabla^2 f, \nabla^2 g).$$

It is observed that R_2 is quadratic in $\nabla^2 f$ and $\nabla^2 g$ and is positive if $|f_x g_y - f_y g_x| \leq 1$ (The is can be found in [33], though in a somewhat more complicated form).

A natural idea is to investigate how the quantity $f_x g_y - f_y g_x$, or the Jacobian of the map (f, g) changes along the flow. Together with the maximal principle, it was shown in [32, 33] that:

(1) $f_x g_y - f_y g_x = 1$ is “preserved” along the mean curvature flow (area preserving).

(2) $|f_x g_y - f_y g_x| < 1$ is “preserved” along the mean curvature flow (area decreasing).

Here a condition is “preserved” means if the condition holds initially, it remains true later as long as the flow exists smoothly.

Combining with the evolution equation of J_2 and applying the maximum principle again show that J_2 has a lower bound, which in turn gives a C^1 estimate of f and g . Notice that J_2 can be regarded as the Jacobian of the projection map onto the first factor of \mathbb{R}^2 . Thus by the inverse function theorem, the graphical condition is also preserved.

Such a condition indeed corresponds to the Gauss map of the submanifold lies in a totally geodesic or geodesically convex subset of the Grassmannian [36]. The underlying fact for this calculation is based on the observation [36] that the Gauss map of the mean curvature flow is a (nonlinear) harmonic map heat flow.

In codimension one case, the higher derivatives estimates follows from the C^1 estimates [9]. The elliptic analogue is the theorem of Moser which states that any Lipschitz solution of the minimal surface equation is smooth. The scenario is totally different in the higher codimension case. Lawson-Osserman constructed minimal cones in higher codimensions and thus a Lipschitz solution to the minimal surface system with $m > 1$ may not be smooth at all.

Here we use “blow-up analysis” for geometric evolution equations. An important tool is Huisken-White’s monotonicity formula [15, 41] which characterizes central blow-up profiles as solutions of the elliptic equation:

$$\vec{H} = -\vec{X}.$$

In general, singularity profiles for parabolic equations are soliton (self-similar) solutions of the equation. In the case of mean curvature flows, soliton (self-similar) solutions are moved by homothety or translations of

the ambient space. Exclusion of self-similar “area-preserving” or “area-decreasing” singularity profiles and the ϵ regularity theorems of White [42] give the desired C^2 estimates.

Two major difficulties remain to be overcome:

(1) Boundary value problem. This was addressed in [37]. More sophisticated barriers that are adapted to the boundary geometry are needed in order to obtain sharper result to cover the area-decreasing case.

(2) Effective estimates in time as $t \rightarrow \infty$. So far, convergence results rely on the sign of the curvature of the ambient space. The C^2 estimates obtained through blow-up analysis usually deteriorate in time.

In the next section, we present the statements of results which are cleanest when M_1 and M_2 are closed Riemannian manifolds with suitable curvature conditions. We remark that there have been several global existence and convergence theorems on higher codimensional graphical mean curvature flows such as [29, 30, 34, 7, 2], etc. Here we focus on those theorems that have implications on isotopy problems.

3. Statements of results related to isotopy problems

3.1. Symplectomorphisms of Riemann surfaces. Let (M_1, g_1) and (M_2, g_2) be Riemann surfaces with metrics of the same constant curvature. We can normalize so the curvature is $-1, 0$ or 1 . Let $f : M_1 \rightarrow M_2$ be an oriented area-preserving map and Σ be the graph of f in $M_1 \times M_2$. A oriented area-preserving map is also a symplectomorphism, i.e. $f^*\omega_2 = \omega_1$ where ω_1 and ω_2 are the area forms (or symplectic forms) of g_1 and g_2 , respectively. The area $A(f)$ of the graph of f is a symmetric function on the symplectomorphism group, i.e. $A(f) = A(f^{-1})$ and the mean curvature flow gives a deformation retract of this group to a finite dimensional one.

THEOREM 1. ([32, 33, 35], see also [38]) *Suppose Σ_0 is the graph of a symplectomorphism $f_0 : M_1 \rightarrow M_2$. The mean curvature flow Σ_t exists for all $t \in [0, \infty)$ and converges smoothly to a minimal submanifold as $t \rightarrow \infty$. Σ_t is the graph of a symplectic isotopy f_t from f_0 to a canonical minimal map f_∞ .*

Since any diffeomorphism is isotopic to an area preserving diffeomorphism, this gives a new proof of Smale’s theorem [27] that $O(3)$ is the deformation retract of the diffeomorphism group of S^2 . For a positive genus Riemann surface, this implies the identity component of the diffeomorphism group is contractible.

The result for the positive genus case was also obtained by Smoczyk [29] under an extra angle condition.

In this case, the graph of the symplectomorphism is indeed a Lagrangian submanifolds in the product space. There have been important recent progresses on the Lagrangian minimal surface equation, we refer to the excellent survey article of Brendle [6] in this direction.

For an area-decreasing map f , i.e. $|f^*\omega_2| < \omega_1$, the mean curvature flow exists for all time and converges to the graph of a constant map, see [33].

3.2. Area-decreasing maps in higher dimensions. The area-decreasing condition, which turns out to be rather natural for the mean curvature flow, can be generalized to higher dimensions. A Lipschitz map $f : M_1 \rightarrow M_2$ between Riemannian manifolds is area-decreasing if the 2-dilation $|\Lambda^2 df|_p < 1$ for each $p \in M_1$. Here $\Lambda^2 df|_p : \Lambda^2 T_p M_1 \rightarrow \Lambda^2 T_{f(p)} M_2$ is the map induced by the differential $df|_p : T_p M_1 \rightarrow T_{f(p)} M_2$.

Equivalently, in local orthonormal coordinate systems on the domain and the target, we ask

$$\left| \frac{\partial f^\alpha}{\partial x^i} \frac{\partial f^\beta}{\partial x^j} - \frac{\partial f^\alpha}{\partial x^i} \frac{\partial f^\beta}{\partial x^j} \right| < 1$$

for $\alpha \neq \beta$, $i \neq j$. This is also the same as $H^2(f(D)) \leq H^2(D)$ for any $D \subset M_1$ of finite two-dimensional Hausdorff measure $H^2(\cdot)$.

In [31], we proved that area decreasing condition is preserved along the mean curvature flow for the graph of a smooth map $f : S^n \rightarrow S^m$ between spheres of constant curvature 1. In addition,

THEOREM 2. [31] *Suppose $n, m \geq 2$. If $f : S^n \rightarrow S^m$ is an area-decreasing smooth map, the mean curvature flow of the graph of f exists for all time, remains a graph, and converges smoothly to a constant map as $t \rightarrow \infty$.*

The most difficult part of the proof is to express the area-decreasing condition as the two-positivity condition (i.e. the sum of the two smallest eigenvalues is positive) for a Lorentzian metric of signature (n, m) and compute the evolution equation of the induced metric.

A simple corollary is the following:

COROLLARY 3. *If $n, m \geq 2$, every area-decreasing map $f : S^n \rightarrow S^m$ is homotopically trivial.*

Gromov [12] shows that for each pair (n, m) , there exists a number $\epsilon(n, m) > 0$, so that any map from S^n to S^m with $|\Lambda^2 df| < \epsilon(n, m)$ is homotopically trivial, where $\epsilon(n, m) \ll 1$. In general, we may consider the k -Jacobian $\Lambda^k df : \Lambda^k TM_1 \rightarrow \Lambda^k TM_2$, whose supreme norm $|\Lambda^k df|$ is called the k -dilation ($k = 1$ is the Lipschitz norm). Guth [13] constructed homotopically non-trivial maps from S^n to S^m with arbitrarily small 3-dilation. It is amazing that 2-dilation is sharp here as it arises naturally from a completely different consideration of the Gauss map of the mean curvature flow (see last section).

3.3. Symplectomorphisms of complex projective spaces. In this section, we consider the generalization of the theorem for symplectomorphisms of Riemann surfaces to higher dimensional manifolds. Let M_1 and M_2 be Kähler manifolds equipped with Kähler-Einstein metrics of the same constant scalar curvature. Let $f : M_1 \rightarrow M_2$ be a symplectomorphism. As

was remarked in the last section, we can consider the graph of f as a Lagrangian submanifold Σ in the product space $M_1 \times M_2$ and deform it by the mean curvature flow. A theorem of Smoczyk [28] (see also [25]) implies that the mean curvature flow Σ_t remains a Lagrangian submanifold. If we can show Σ_t remains graphical as well, it will corresponds to a symplectic isotopy $f_t : M_1 \rightarrow M_2$. The simplest case to be considered in higher dimension is $M_1 = M_2 = \mathbb{CP}^n$ with the Fubini-Study metric. In a joint work with Medoš, we proved the following pinching theorem.

THEOREM 4. [22] *There exists an explicitly computable constant $\Lambda > 1$ depending only on n , such that any symplectomorphism $f : \mathbb{CP}^n \rightarrow \mathbb{CP}^n$ with*

$$\frac{1}{\Lambda}g \leq f^*g \leq \Lambda g$$

is symplectically isotopic to a biholomorphic isometry of \mathbb{CP}^n through the mean curvature flow.

A theorem of Gromov [11] shows that, when $n = 2$, the statement holds true without any pinching condition by the method of pseudoholomorphic curves. Our theorem is not strong enough to give an analytic proof of Gromov's theorem for $n = 2$. However, for $n \geq 3$, this seems to be the first known result.

Unlike previous theorems, Grassmannian geometry does not quite help here, as the subset that corresponds to biholomorphic isometries does not have any convex neighborhood in the Grassmannian. The integrability condition, or the Gauss-Codazzi equations, is used in an essential way to overcome this difficulty.

References

- [1] B. Andrews, *Contraction of convex hypersurfaces in Riemannian spaces*. J. Differential Geom. 39 (1994), no. 2, 407–431.
- [2] B. Andrews and C. Baker, *Mean curvature flow of pinched submanifolds to spheres*. J. Differential Geom. 85 (2010), no. 3, 357–395.
- [3] E. Bombieri, E. De Giorgi and M. Miranda, *Una maggiorazione a priori relativa alle ipersuperficie minimali non parametriche*. (Italian) Arch. Rational Mech. Anal. 32 (1969) 255–267.
- [4] E. Bombieri, E. De Giorgi, and E. Giusti, *Minimal cones and the Bernstein problem*. Invent. Math. 7 (1969) 243–68.
- [5] K. A. Brakke, *The motion of a surface by its mean curvature*. Mathematical Notes, 20. Princeton University Press, Princeton, N.J., 1978.
- [6] S. Brendle, *On the Lagrangian minimal surface equation and related problems*. arXiv:1108.0148v1.
- [7] J.-Y. Chen, J.-Y. Li, and G. Tian, *Two-dimensional graphs moving by mean curvature flow*. Acta Math. Sin. (Engl. Ser.) 18 (2002), no. 2, 209–24.
- [8] T. H. Colding and W. P. Minicozzi, II, *Sharp estimates for mean curvature flow of graphs*. J. Reine Angew. Math. 574 (2004), 187–195.
- [9] K. Ecker and G. Huisken, *Mean curvature evolution of entire graphs*. Ann. of Math. (2) 130 (1989), no. 3, 453–471.
- [10] K. Ecker and G. Huisken, *Interior estimates for hypersurfaces moving by mean curvature*. Invent. Math. 105 (1991), no. 3, 547–569.

- [11] M. Gromov, *Pseudoholomorphic curves in symplectic manifolds*. Invent. Math. 82 (1985), no. 2, 307–347.
- [12] M. Gromov, *Metric Structures for Riemannian and Non-Riemannian Spaces*. Birkhauser, Boston, 1998.
- [13] L. Guth, *Homotopically non-trivial maps with small k-dilation*. Available at <http://xxx.lanl.gov/pdf/0709.1241>
- [14] R. Hamilton, *Four-manifolds with positive curvature operator*. J. Differential Geom. 24 (1986), no. 2, 153–179.
- [15] G. Huisken, *Asymptotic behavior for singularities of the mean curvature flow*. J. Differential Geom. 31 (1990), no. 1, 285–299, MR1030675, Zbl 0694.53005.
- [16] G. Huisken, *Flow by mean curvature of convex surfaces into spheres*. J. Differential Geom. 20 (1984), no. 1, 237–266.
- [17] G. Huisken and C. Sinestrari, *Mean curvature flow singularities for mean convex surfaces*. Calc. Var. Partial Differential Equations 8 (1999), no. 1, 1–14.
- [18] G. Huisken and C. Sinestrari, *Convexity estimates for mean curvature flow and singularities of mean convex surfaces*. Acta Math. 183 (1999), no. 1, 45–70.
- [19] T. Ilmanen, *Elliptic regularization and partial regularity for motion by mean curvature*. Mem. Amer. Math. Soc. 108 (1994), no. 520.
- [20] H. Jenkins and J. Serrin, *The Dirichlet problem for the minimal surface equation in higher dimensions*. J. Reine Angew. Math. 229 (1968) 170–87.
- [21] H. B. Lawson and R. Osserman, *Non-existence, non-uniqueness and irregularity of solutions to the minimal surface system*. Acta Math. 139 (1977), no. 1-2, 1–17.
- [22] I. Medoš and M.-T. Wang, *Deforming symplectomorphisms of complex projective spaces by the mean curvature flow*. J. Differential Geom. 87 (2011), no. 2, 309–342.
- [23] C. B. Morrey, Jr. *Multiple integrals in the calculus of variations*. Die Grundlehren der mathematischen Wissenschaften, Band 130 Springer-Verlag New York, Inc., New York 1966.
- [24] J. Moser, *A new proof of De Giorgi's theorem concerning the regularity problem for elliptic differential equations*. Comm. Pure Appl. Math. 13 (1960) 457–68.
- [25] Y.-G. Oh, *Mean curvature vector and symplectic topology of Lagrangian submanifolds in Einstein-Kähler manifolds*. Math. Z. 216 (1994), no. 3, 471–482.
- [26] R. Osserman, *A survey of minimal surfaces. Second edition*. Dover Publications, Inc., New York, 1986.
- [27] S. Smale, *Diffeomorphisms of the 2-sphere*. Proc. Amer. Math. Soc. 10 1959, 621–626.
- [28] K. Smoczyk, *A canonical way to deform a Lagrangian submanifold*. preprint, dg-ga/9605005.
- [29] K. Smoczyk, *Angle theorems for the Lagrangian mean curvature flow*. Math. Z. 240 (2002), no. 4, 849–883.
- [30] K. Smoczyk and M.-T. Wang, *Mean curvature flows of Lagrangian submanifolds with convex potentials*. J. Differential Geom. 62 (2002), no. 2, 243–257.
- [31] M.-P. Tsui and M.-T. Wang, *Mean curvature flows and isotopy of maps between spheres*. Comm. Pure. Appl. Math. 57 (2004), no. 8 , 1110–1126.
- [32] M.-T. Wang, *Deforming area preserving diffeomorphism of surfaces by mean curvature flow*. Math. Res. Lett. 8 (2001), no.5-6, 651–662.
- [33] M.-T. Wang, *Mean curvature flow of surfaces in Einstein Four-Manifolds*. J. Differential Geom. 57 (2001), no.2, 301–338.
- [34] M.-T. Wang, *Long-time existence and convergence of graphic mean curvature flow in arbitrary codimension*. Invent. math. 148 (2002) 3, 525–543.
- [35] M.-T. Wang, *A convergence result of the Lagrangian mean curvature flow*. Third International Congress of Chinese Mathematicians. Part 1, 2, 291–295, AMS/IP Stud. Adv. Math., 42, pt. 1, 2, Amer. Math. Soc., Providence, RI, 2008.
- [36] M.-T. Wang, *Gauss maps of the mean curvature flow*. Math. Res. Lett. 10 (2003), no. 2-3, 287–299.

- [37] M.-T. Wang, *Lectures on mean curvature flows in higher codimensions*. Handbook of geometric analysis. No. 1, 525–543, Adv. Lect. Math. (ALM), 7, Int. Press, Somerville, MA, 2008.
- [38] M.-T. Wang, *Some recent developments in Lagrangian mean curvature flows*. Surveys in differential geometry. Vol. XII. Geometric flows, 333–347, Surv. Differ. Geom., 12, Int. Press, Somerville, MA, 2008.
- [39] M.-T. Wang, *The Dirichlet problem for the minimal surface system in arbitrary codimension*. Comm. Pure. Appl. Math. 57 (2004), no. 2, 267–281.
- [40] X.-J. Wang, *Convex solutions to the mean curvature flow*. Ann. of Math. (2) 173 (2011), no. 3, 1185–1239.
- [41] B. White, *The nature of singularities in mean curvature flow of mean-convex sets*. J. Amer. Math. Soc. 16 (2003), no. 1, 123–138.
- [42] B. White, *A local regularity theorem for classical mean curvature flow*. Ann. of Math. (2) **161** (2005), no. 3, 1487–1519, MR2180405, Zbl 1091.53045.

DEPARTMENT OF MATHEMATICS, COLUMBIA UNIVERSITY, 2990 BROADWAY, NEW YORK, NY 10027, USA

This page intentionally left blank

Eigenfunctions and nodal sets

Steve Zelditch

ABSTRACT. This is a survey of recent results on nodal sets of eigenfunctions of the Laplacian on Riemannian manifolds. The emphasis is on complex nodal sets of analytic continuations of eigenfunctions.

Let (M, g) be a (usually compact) Riemannian manifold of dimension n , and let $\{\varphi_j\}$ denote an orthonormal basis of eigenfunctions of its Laplacian,

$$(1) \quad \Delta_g \varphi_j = -\lambda_j^2 \varphi_j \quad \langle \varphi_j, \varphi_k \rangle = \delta_{jk}.$$

Here $\langle u, v \rangle = \int_M u v dV_g$ where dV_g is the volume form of (M, g) . If $\partial M \neq 0$ we impose Dirichlet or Neumann boundary conditions. When (M, g) is compact, the spectrum of Δ is discrete and can be put in non-decreasing order $\lambda_0 < \lambda_1 \leq \lambda_2 \uparrow \infty$. The eigenvalues λ_j^2 are often termed energies while their square roots λ_j are often termed the frequencies. The nodal set of an eigenfunction φ_λ is the zero set

$$(2) \quad Z_{\varphi_\lambda} = \{x \in M : \varphi_\lambda(x) = 0\}.$$

The aim of this survey is to review some recent results on the \mathcal{H}^{n-1} -surface measure and on the yet more difficult problem of the spatial distribution of the nodal sets, i.e. the behavior of the integrals

$$(3) \quad \frac{1}{\lambda_j} \int_{Z_{\varphi_\lambda}} f dS_{\lambda_j}, \quad (f \in C(M))$$

as $\lambda \rightarrow \infty$. Here, $dS_\lambda = d\mathcal{H}^{n-1}$ denotes the Riemannian hypersurface volume form on Z_{φ_λ} . More generally, we consider the same problems for any level set

$$(4) \quad \mathcal{N}_{\varphi_\lambda}^c := \{\varphi_\lambda = c\},$$

where c is a constant (which in general may depend on λ). Nodal sets are special level sets and much more attention has been devoted to them than

Research partially supported by NSF grant # DMS-0904252.

¹In difference references we use either the notation Z or \mathcal{N} for the nodal set. Sometimes we use the subscript φ_λ and sometimes only λ .

other level sets, but it is often of interest to study general level sets and in particular ‘high level’ sets or excursion sets.

We have recently written surveys [Z5, Z6] on the global harmonic analysis of eigenfunctions, which include some discussion of nodal sets and critical point sets. To the extent possible, we hope to avoid repeating what is written there, but inevitably there will be some overlap. We refer there and [H] for background on well-established results. We also decided to cover some results of research in progress (especially from [Z3], but also on L^∞ quantum ergodic theory). We generally refer to the results as ‘Conjectures’ even when detailed arguments exist, since they have not yet been carefully examined by others.

There are two basic intuitions underlying many of the conjectures and results on eigenfunctions:

- Eigenfunctions of Δ_g -eigenvalue $-\lambda^2$ are similar to polynomials of degree λ . In particular, Z_λ is similar to a real algebraic variety of degree λ .

Of course, this intuition is most reliable when (M, g) is real analytic. It is quite unclear at this time how reliable it is for general C^∞ metrics, although there are some recent improvements on volumes and equidistribution in the smooth case.

- High frequency behavior of eigenfunctions reflects the dynamics of the geodesic flow $G^t : S^*M \rightarrow S^*M$ of M . Here, S^*M is the unit co-sphere bundle of (M, g) .

When the dynamics is “chaotic” (highly ergodic), then eigenfunctions are de-localized and behave like Gaussian random waves of almost fixed frequency. This motivates the study of Gaussian random wave models for eigenfunctions, and suggests that in the ‘chaotic case’ nodal sets should be asymptotically uniformly distributed.

When G^t is completely integrable, model eigenfunctions are highly localized and their nodal sets are often exhibit quite regular patterns. The latter heuristic is not necessarily expected when there exist high multiplicities, as for rational flat tori, and then some weaker randomness can enter.

Both of these general intuitions lead to predictions about nodal sets and critical point sets. Most of the predictions are well beyond current or foreseeable techniques to settle. A principal theme of this survey is that the analogues of such ‘wild’ predictions can sometimes be proved for real analytic (M, g) if one analytically continues eigenfunctions to the complexification of M and studies complex nodal sets instead of real ones.

As with algebraic varieties, nodal sets in the real analytic case are better behaved in the complex domain than the real domain. That is, zero sets of analytic continuations of eigenfunctions to the complexification of M behave

like complex algebraic varieties and also reflect the dynamics of the geodesic flow.

It is well-known that the complexification of M can be identified with a neighborhood of the zero-section of the phase space T^*M . That is one reason why dynamics of the geodesic flow has greater impact on the complex nodal set.

We will exhibit a number of relatively recent results (some unpublished elsewhere) which justify this viewpoint:

- Theorem 8.4, which shows that complex methods can be used to give upper bounds on the number of nodal components of Dirichlet or Neumann eigenfunctions which “touch the boundary” of a real analytic plane domain.
- Theorem 9.1 on the limit distribution of the normalized currents of integration

$$\frac{1}{\lambda_{j_k}}[Z_{\varphi_{j_k}^c}]$$

over the complex zero sets of “ergodic eigenfunctions” in the complex domain.

- Theorem 11.2 and Corollary 11.1, which show that the similar currents for analytic continuations of “Riemannian random waves” tend to the same limit almost surely. Thus, the prediction that zero sets of ergodic eigenfunctions agrees with that of random waves is correct in the complex domain.
- Sharper results on the distribution of intersections points of nodal sets and geodesics on complexified real analytic surfaces (Theorem 10.1).

Our analysis of nodal sets in the complex domain is based on the use of complex Fourier integral techniques (i.e. generalized Paley-Wiener theory). The principal tools are the analytic continuation of the Poisson-wave kernel and the Szegö kernel in the complex domain. They become Fourier integral operators with complex phase and with wave fronts along the complexified geodesic flow. One can read off the growth properties of complexified eigenfunctions from mapping properties of such operators. Log moduli of complexified spectral projectors are asymptotically extremal plurisubharmonic functions for all (M, g) . These ideas are the basis of the articles [Z2, TZ, Z3, Z4, Z8, Z9, He]. Such ideas have antecedents in work of S. Bernstein, Baouendi- Goulaouic, and Donnelly-Fefferman, Guillemin, F.H. Lin (among others) .

We note that the focus on complex nodal sets only makes sense for real analytic (M, g) . It is possible that one can study “almost analytic extensions” of eigenfunctions for general C^∞ metrics in a similar spirit, but this is just a speculation and certain key methods break down when g is not real analytic. Hence the results in the C^∞ case are much less precise than in the real analytic case.

It should also be mentioned that much work on eigenfunctions concerns ground states, i.e. the first and second eigenfunctions. Unfortunately, we do not have the space or expertise to review the results on ground states in this survey. For a sample we refer to [Me]. Further, many if not all of the techniques and results surveyed here have generalizations to Schrödinger operators $-\hbar^2 \Delta + V$. For the sake of brevity we confine the discussion to the Laplacian.

0.1. Notation. The first notational issue is whether to choose Δ_g to be the positive or negative Laplacian. The traditional choice

$$(5) \quad \Delta_g = \frac{1}{\sqrt{g}} \sum_{i,j=1}^n \frac{\partial}{\partial x_i} \left(g^{ij} \sqrt{g} \frac{\partial}{\partial x_j} \right).$$

makes Δ_g is negative, but many authors call $-\Delta_g$ the Laplacian to avoid the minus signs. Also, the metric g is often fixed and is dropped from the notation.

A less traditional choice is to denote eigenvalues by λ^2 rather than λ . It is a common convention in microlocal analysis and so we adopt it here. But we warn that λ is often used to denote Δ -eigenvalues as is [DF, H].

We sometimes denote eigenfunctions of eigenvalue $-\lambda^2$ by φ_λ when we only wish to emphasize the corresponding eigenvalue and do not need φ_λ to be part of an orthonormal basis. For instance, when Δ_g has multiplicities as on the standard sphere or rational torus, there are many possible orthonormal bases. But estimates on $\mathcal{H}^{n-1}(Z_{\varphi_\lambda})$ do not depend on whether φ_λ is included in the orthonormal basis.

Acknowledgments. Thanks to D. Mangoubi, G. Rivière, C. D. Sogge and B. Shiffman for helpful comments/improvements on the exposition, and to S. Dyatlov for a stimulating discussion of L^∞ quantum ergodicity.

1. Basic estimates of eigenfunctions

We start by collecting some classical elliptic estimates and their applications to eigenfunctions.

First, the general Sobolev estimate: Let $w \in C_0^\infty(\Omega)$ where $\Omega \subset \mathbb{R}^n$ with $n \geq 3$. Then there exists $C > 0$:

$$\left(\int_{\Omega} |w|^{\frac{2n}{n-2}} \right)^{\frac{n-2}{n}} \leq C \int_{\Omega} |\nabla w|^2.$$

Next, we recall the Bernstein gradient estimates:

THEOREM 1.1. [DF3] *Local eigenfunctions of a Riemannian manifold satisfy:*

(1) L^2 Bernstein estimate:

$$(6) \quad \left(\int_{B(p,r)} |\nabla \varphi_\lambda|^2 dV \right)^{1/2} \leq \frac{C\lambda}{r} \left(\int_{B(p,r)} |\varphi_\lambda|^2 dV \right)^{1/2}.$$

(2) L^∞ Bernstein estimate: There exists $K > 0$ so that

$$(7) \quad \max_{x \in B(p,r)} |\nabla \varphi_\lambda(x)| \leq \frac{C\lambda^K}{r} \max_{x \in B(p,r)} |\varphi_\lambda(x)|.$$

(3) Dong's improved bound:

$$\max_{B_r(p)} |\nabla \varphi_\lambda| \leq \frac{C_1 \sqrt{\lambda}}{r} \max_{B_r(p)} |\varphi_\lambda|$$

for $r \leq C_2 \lambda^{-1/4}$.

Another well-known estimate is the doubling estimate:

THEOREM 1.2. (Donnelly-Fefferman, Lin) and [H] (Lemma 6.1.1) Let φ_λ be a global eigenfunction of a $C^\infty(M, g)$ there exists $C = C(M, g)$ and r_0 such that for $0 < r < r_0$,

$$\frac{1}{Vol(B_{2r}(a))} \int_{B_{2r}(a)} |\varphi_\lambda|^2 dV_g \leq e^{C\lambda} \frac{1}{Vol(B_r(a))} \int_{B_r(a)} |\varphi_\lambda|^2 dV_g.$$

Further,

$$(8) \quad \max_{B(p,r)} |\varphi_\lambda(x)| \leq \left(\frac{r}{r'} \right)^{C\lambda} \max_{x \in B(p,r')} |\varphi_\lambda(x)|, \quad (0 < r' < r).$$

The doubling estimates imply the vanishing order estimates. Let $a \in M$ and suppose that $u(a) = 0$. By the vanishing order $\nu(u, a)$ of u at a is meant the largest positive integer such that $D^\alpha u(a) = 0$ for all $|\alpha| \leq \nu$.

THEOREM 1.3. Suppose that M is compact and of dimension n . In the case of a global eigenfunction, $\nu(\varphi_\lambda, a) \leq C(M, g)\lambda$.

We now recall quantitative lower bound estimates. They follow from doubling estimates and also from Carleman inequalities.

THEOREM 1.4. Suppose that M is compact and that φ_λ is a global eigenfunction, $\Delta \varphi_\lambda = \lambda^2 \varphi_\lambda$. Then for all p, r , there exist $C, C' > 0$ so that

$$\max_{x \in B(p,r)} |\varphi_\lambda(x)| \geq C'e^{-C\lambda}.$$

Local lower bounds on $\frac{1}{\lambda} \log |\varphi_\lambda|$ follow from doubling estimates. They imply that there exists $A, \delta > 0$ so that, for any $\zeta_0 \in \overline{M_{\tau/2}}$,

$$(9) \quad \sup_{\zeta \in B_\delta(\zeta_0)} |\varphi_\lambda(\zeta)| \geq Ce^{-A\lambda}.$$

To see how doubling estimates imply Theorem 1.4, we observe that there exists a point $x_0 \in M$ so that $|\varphi_\lambda(x_0)| \geq 1$. Any point of $\overline{M_{\tau/2}}$ can be linked

to this point by a smooth curve of uniformly bounded length. We then choose δ sufficiently small so that the δ -tube around the curve lies in M_τ and link $B_\delta(\zeta)$ to $B_\delta(x_0)$ by a chain of δ -balls in M_τ where the number of links in the chain is uniformly bounded above as ζ varies in M_τ . If the balls are denoted B_j we have $\sup_{B_{j+1}} |\varphi_\lambda| \leq e^{\beta\lambda} \sup_{B_j} |\varphi_\lambda|$ since $B_{j+1} \subset 2B_j$. The growth estimate implies that for any ball B , $\sup_{2B} |\varphi_\lambda| \leq e^{C\lambda} \sup_B |\varphi_\lambda|$. Since the number of balls is uniformly bounded,

$$1 \leq \sup_{B_\delta(x_0)} |\varphi_\lambda| \leq e^{A\lambda} \sup_{B_\delta(\zeta)} |\varphi_\lambda|.$$

proving Theorem 1.4.

As an illustration, Gaussian beams such as highest weight spherical harmonics decay at a rate $e^{-C\lambda d^2(x,\gamma)}$ away from a stable elliptic orbit γ . Hence if the closure of an open set is disjoint from γ , one has a uniform exponential decay rate which saturate the lower bounds.

We now recall sup-norm estimates of eigenfunctions which follow from the local Weyl law:

$$\Pi_\lambda(x, x) := \sum_{\lambda_\nu \leq \lambda} |\varphi_\nu(x)|^2 = (2\pi)^{-n} \int_{p(x,\xi) \leq \lambda} d\xi + R(\lambda, x)$$

with uniform remainder bounds

$$|R(\lambda, x)| \leq C\lambda^{n-1}, \quad x \in M.$$

Since the integral in the local Weyl law is a continuous function of λ and since the spectrum of the Laplacian is discrete, this immediately gives

$$\sum_{\lambda_\nu = \lambda} |\varphi_\nu(x)|^2 \leq 2C\lambda^{n-1}$$

which in turn yields

$$(10) \quad \|\varphi_\lambda\|_{C^0} = O(\lambda^{\frac{n-1}{2}})$$

on any compact Riemannian manifold.

1.1. L^p estimates. The classical Sogge estimates state that, for any compact Riemannian manifold of dimension n , we have

$$(11) \quad \frac{\|\varphi_\lambda\|_p}{\|\varphi_\lambda\|_2} = O(\lambda^{\delta(p)}), \quad 2 \leq p \leq \infty,$$

where

$$(12) \quad \delta(p) = \begin{cases} n\left(\frac{1}{2} - \frac{1}{p}\right) - \frac{1}{2}, & \frac{2(n+1)}{n-1} \leq p \leq \infty \\ \frac{n-1}{2}\left(\frac{1}{2} - \frac{1}{p}\right), & 2 \leq p \leq \frac{2(n+1)}{n-1}. \end{cases}$$

Since we often use surfaces as an illustration, we note that in dimension 2 one has for $\lambda \geq 1$,

$$(13) \quad \|\varphi_\lambda\|_{L^p(M)} \leq C\lambda^{\frac{1}{2}\left(\frac{1}{2} - \frac{1}{p}\right)} \|\varphi_\lambda\|_{L^2(M)}, \quad 2 \leq p \leq 6,$$

and

$$(14) \quad \|\varphi_\lambda\|_{L^p(M)} \leq C\lambda^{2\left(\frac{1}{2} - \frac{1}{p}\right) - \frac{1}{2}} \|\varphi_\lambda\|_{L^2(M)}, \quad 6 \leq p \leq \infty.$$

These estimates are also sharp for the round sphere S^2 . The first estimate, (13), is saturated by highest weight spherical harmonics. The second estimate, (14), is sharp due to the zonal functions on S^2 , which concentrate at points. We go over these examples in §3.2.

2. Volume and equidistribution problems on nodal sets and level sets

We begin the survey by stating some of the principal problems and results regarding nodal sets and more general level sets. Some of the problems are intentionally stated in vague terms that admit a number of rigorous formulations.

2.1. Hypersurface areas of nodal sets. One of the principal problems on nodal sets is to measure their hypersurface volume. In the real analytic case, Donnelly-Fefferman ([DF] (see also [Lin])) proved:

THEOREM 2.1. *Let (M, g) be a compact real analytic Riemannian manifold, with or without boundary. Then there exist c_1, C_2 depending only on (M, g) such that*

$$c_1\lambda \leq \mathcal{H}^{m-1}(Z_{\varphi_\lambda}) \leq C_2\lambda, \quad (\Delta\varphi_\lambda = \lambda^2\varphi_\lambda; c_1, C_2 > 0).$$

The bounds were conjectured by S. T. Yau [Y1, Y2] for all $C^\infty(M, g)$, but this remains an open problem. The lower bound was proved for all C^∞ metrics for surfaces, i.e. for $n = 2$ by Brüning [Br]. For general C^∞ metrics the sharp upper and lower bounds are not known, although there has been some recent progress that we consider below.

The nodal hypersurface bounds are consistent with the heuristic that φ_λ is the analogue on a Riemannian manifold of a polynomial of degree λ , since the hypersurface volume of a real algebraic variety is bounded by its degree.

2.2. Equidistribution of nodal sets in the real domain. The equidistribution problem for nodal sets is to study the behavior of the integrals (3) of general continuous functions f over the nodal set. Here, we normalize the delta-function on the nodal set by the conjectured surface volume of §2.1. More precisely:

Problem Find the weak* limits of the family of measures $\{\frac{1}{\lambda_j}dS_{\lambda_j}\}$.

Note that in the C^∞ case we do not even know if this family has uniformly bounded mass. The high-frequency limit is the semi-classical limit and generally signals increasing complexity in the ‘topography’ of eigenfunctions.

Heuristics from quantum chaos suggests that eigenfunctions of quantum chaotic systems should behave like random waves. The random wave model is defined and studied in [Z4] (see §11), and it is proved (see Theorem 11.1)

that if one picks a random sequence $\{\psi_{\lambda_j}\}$ of random waves of increasing frequency, then almost surely

$$(15) \quad \frac{1}{\lambda_j} \int_{\mathcal{H}_{\psi_{\lambda_j}}} f dS_{\lambda_j} \rightarrow \frac{1}{Vol(M)} \int_M f dV_g,$$

i.e. their nodal sets become equidistributed with respect to the volume form on M . Hence the heuristic principle leads to the conjecture that nodal sets of eigenfunctions of quantum chaotic systems should become equidistributed according to the volume form.

The conjecture for eigenfunctions (rather than random waves) is far beyond any current techniques and serves mainly as inspiration for studies of equidistribution of nodal sets.

A yet more speculative conjecture in quantum chaos is that the nodal sets should tend to CLE_6 curves in critical percolation. CLE refers to conformal loop ensembles, which are closed curves related to SLE curves. As above, this problem is motivated by a comparison to random waves, but for these the problem is also completely open. In §12 we review the heuristic principles which started in condensed matter physics [**KH**, **KHS**, **Isi**, **IsiK**, **Wei**] before migrating to quantum chaos [**BS**, **BS2**, **FGS**, **BGS**, **SS**, **EGJS**]. It is dubious that such speculative conjectures can be studied rigorously in the foreseeable future, but we include them to expose the reader to the questions that are relevant to physicists.

2.3. L^1 norms and nodal sets. Besides nodal sets it is of much current interest to study L^p norms of eigenfunctions globally on (M, g) and also of their restrictions to submanifolds. In fact, recent results show that nodal sets and L^p norms are related. For instance, in §4 we will use the identity

$$(16) \quad \|\varphi_\lambda\|_{L^1} = \frac{1}{\lambda^2} \int_{Z_{\varphi_\lambda}} |\nabla \varphi_\lambda| dS$$

relating the L^1 norm of φ_λ to a weighted integral over Z_{φ_λ} to obtain lower bounds on $\mathcal{H}^{n-1}(Z_{\varphi_\lambda})$. See (21).

Obtaining lower bounds on L^1 norms of eigenfunctions is closely related to finding upper bounds on L^4 norms. The current bounds are not sharp enough to improve nodal set bounds.

2.4. Critical points and values. A closely related problem in the ‘topography’ of Laplace eigenfunctions φ_λ is to determine the asymptotic distribution of their critical points

$$C(\varphi_\lambda) = \{x : \nabla \varphi_\lambda(x) = 0\}.$$

This problem is analogous to that of measuring the hypersurface area $\mathcal{H}^{n-1}(Z_\lambda)$ of the nodal (zero) set of φ_λ , but it is yet more complicated due to the instability of the critical point set as the metric varies. For a generic metric, all eigenfunctions are Morse functions and the critical point set is

discrete. One may ask to count the number of critical points asymptotically as $\lambda \rightarrow \infty$. But there exist metrics (such as the flat metric on the torus, or the round metric on the sphere) for which the eigenfunctions have critical hypersurfaces rather than points. To get around this obstruction, we change the problem from counting critical points to counting critical values

$$CV(\varphi_\lambda) = \{\varphi_\lambda(x) : \nabla \varphi_\lambda(x) = 0\}.$$

Since a real analytic function on a compact real analytic manifold has only finitely many critical values, eigenfunctions of real analytic Riemannian manifolds (M, g) have only finitely many critical values and we can ask to count them. See Conjecture 6.2 for an apparently plausible bound. Moreover for generic real analytic metrics, all eigenfunctions are Morse functions and there exists precisely one critical point for each critical value. Thus, in the generic situation, counting critical values is equivalent to counting critical points. To our knowledge, there are no results on this problem, although it is possible to bound the \mathcal{H}^{n-1} -measure of $C(\varphi_\lambda)$ (see Theorem [Ba]). However $\mathcal{H}^{n-1}(C(\varphi_\lambda)) = 0$ in the generic case and in special cases where it is not zero the method is almost identical to bounds on the nodal set. Thus, such results bypass all of the difficulties in counting critical values. We will present one new (unpublished) result which generalizes (16) to critical points. But the resulting identity is much more complicated than for zeros.

Singular points are critical points which occur on the nodal sets. We recall (see [H, HHL, HHON]) that the singular set

$$\Sigma(\varphi_\lambda) = \{x \in Z_{\varphi_\lambda} : \nabla \varphi_\lambda(x) = 0\}$$

satisfies $\mathcal{H}^{n-2}(\Sigma(\varphi_\lambda)) < \infty$. Thus, outside of a codimension one subset, Z_{φ_λ} is a smooth manifold, and the Riemannian surface measure $dS = \iota_{\frac{\nabla \varphi_\lambda}{|\nabla \varphi_\lambda|}} dV_g$ on Z_{φ_λ} is well-defined. We refer to [HHON, H, HHL, HS] for background.

2.5. Inradius. It is known that in dimension two, the minimal possible area of a nodal domain of a Euclidean eigenfunction is $\pi(\frac{j_1}{\lambda})^2$. This follows from the two-dimensional Faber-Krahn inequality,

$$\lambda_k(\Omega) \text{Area}(D) = \lambda_1(D) \text{Area}(D) \geq \pi j_1^2$$

where D is a nodal domain in Ω . In higher dimensions, the Faber-Krahn inequality shows that on any Riemannian manifold the volume of any nodal domain is $\geq C\lambda^{-n}$ [EK].

Another size measure of a nodal domain is its inradius r_λ , i.e. the radius of the largest ball contained inside the nodal domain. As can be seen from computer graphics (see e.g. [HEJ]), there are a variety of ‘types’ of nodal components. In [Man3], Mangoubi proves that

$$(17) \quad \frac{C_1}{\lambda} \geq r_\lambda \geq \frac{C_2}{\lambda^{\frac{1}{2}k(n)} (\log \lambda)^{2n-4}},$$

where $k(n) = n^2 - 15n/8 + 1/4$; note that eigenvalues in [Man] are denoted λ while here we denote them by λ^2 . In dimension 2, it is known (loc.cit.)

that

$$(18) \quad \frac{C_1}{\lambda} \geq r_\lambda \geq \frac{C_2}{\lambda}.$$

2.6. Decompositions of M with respect to φ_λ . There are two natural decompositions (partitions) of M associated to an eigenfunction (or any smooth function).

(i) Nodal domain decomposition.

First is the decomposition of M into nodal domains of φ_λ . As in [PS] we denote the collection of nodal domains by $\mathcal{A}(\varphi_\lambda)$ and denote a nodal domain by A . Thus,

$$M \setminus Z_{\varphi_\lambda} = \bigcup_{A \in \mathcal{A}(\varphi_\lambda)} A.$$

When 0 is a regular value of φ_λ the level sets are smooth hypersurfaces and one can ask how many components of Z_{φ_λ} occur, how many components of the complement, the topological types of components or the combinatorics of the set of domains. When 0 is a singular value, the nodal set is a singular hypersurface and can be connected but one may ask similar questions taking multiplicities of the singular points into account.

To be precise, let

$$\mu(\varphi_\lambda) = \#\mathcal{A}(\varphi_\lambda), \quad \nu(\varphi_\lambda) = \# \text{ components of } Z(\varphi_\lambda).$$

The best-known problem is to estimate $\mu(\varphi_\lambda)$. According to the Courant nodal domain theorem, $\mu(\varphi_{\lambda_n}) \leq n$. In the case of spherical harmonics, where many orthonormal bases are possible, it is better to estimate the number in terms of the eigenvalue, and the estimate has the form $\mu(\varphi_\lambda) \leq C(g)\lambda^m$ where $m = \dim M$ and $C(g) > 0$ is a constant depending on g . In dimension 2, Pleijel used the Faber-Krahn theorem to improve the bound to

$$\limsup_{\lambda \rightarrow \infty} \frac{\mu(\varphi_\lambda)}{\lambda^2} \leq \frac{4}{j_0^2} < 0.69$$

where j_0 is the smallest zero of the J_0 Bessel function.

A wide variety of behavior is exhibited by spherical harmonics of degree N . We review the definitions below. The even degree harmonics are equivalent to real projective plane curves of degree N . But each point of \mathbb{RP}^2 corresponds to a pair of points of S^2 and at most one component of the nodal set is invariant under the anti-podal map. For other components, the anti-podal map takes a component to a disjoint component. Thus there are essentially twice the number of components in the nodal set as components of the associated plane curve.

As discussed in [Ley], one has:

- Harnack's inequality: the number of components of any irreducible real projective plane curve is bounded by $g+1$ where g is the genus of the curve.

- If p is a real projective plane curve of degree N then its genus is given by Noether's formula

$$g = \frac{(N-1)(N-2)}{2} - \sum_{\text{singular points } x} \frac{\text{ord}_p(x)(\text{ord}_p(x) - 1)}{2}$$

where $\text{ord}_p(x)$ is the order of vanishing of φ_λ at x . Thus, the number of components is $\leq \frac{(N-1)(N-2)}{2} + 1$ for a non-singular irreducible plane curve of degree N .

Curves which achieve the maximum are called M -curves. Also famous are Harnack curves, which are M curves for which there exist three distinct lines ℓ_j of \mathbb{RP}^2 and three distinct arcs a_j of the curve on one component so that $\#a_j \cap \ell_j = N$. It follows from Pleijel's bound that nodal sets of spherical harmonics cannot be maximal for large N , since half of the Pleijel bound is roughly $.35N^2$ which is below the threshold $.5N^2 + O(N)$ for maximal curves.

Associated to the collection of nodal domains is its incidence graph Γ_λ , which has one vertex for each nodal domain, and one edge linking each pair of nodal domains with a common boundary component. Here we assume that 0 is a regular value of φ_λ so that the nodal set is a union of embedded submanifolds. The Euler characteristic of the graph is the difference between the number of nodal domains and nodal components. In the non-singular case, one can convert the nodal decomposition into a cell decomposition by attaching a one cell between two adjacent components, and then one has $\mu(\varphi_\lambda) = \nu(\varphi_\lambda) + 1$ (see Lemma 8 of [Ley]).

The possible topological types of arrangements of nodal components of spherical harmonics is studied in [EJN]. They prove that for any $m \leq N$ with $N - m$ even and for every set of m disjoint closed curves whose union is invariant with respect to the antipodal map, there exists an eigenfunction whose nodal set has the topological type of the union of curves. Note that these spherical harmonics have relatively few nodal domains compared to the Pleijel bound. It is proved in [NS] that random spherical harmonics have aN^2 nodal components for some (undetermined) $a > 0$.

Morse-Smale decomposition

For generic metrics, all eigenfunctions are Morse functions [U]. Suppose that $f : M \rightarrow \mathbb{R}$ is a Morse function. For each critical point p let $W^s p$ (the descending cell through p) denote the union of the downward gradient flow lines which have p as their initial point, i.e. their α -limit point. Then W_p is a cell of dimension $\lambda_p = \text{number of negative eigenvalues of } H_p f$. By the Morse-Smale decomposition we mean the decomposition

$$M = \bigcup_{p:df(p)=0} W_p^s$$

It is not a good cell decomposition in general. If we change f to $-f$ we get the decomposition into ascending cells $M = \bigcup_{p:df(p)=0} W_p^u$. If the intersections

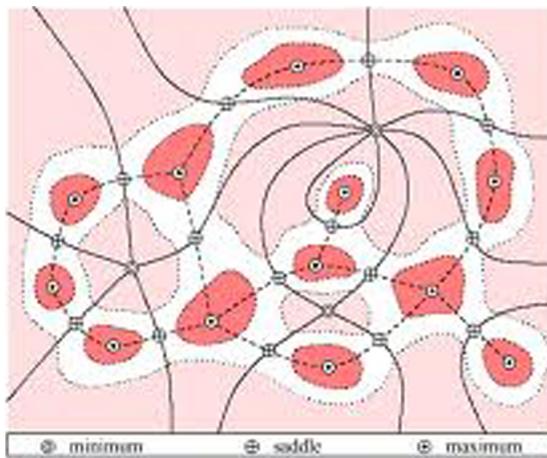


FIGURE 1. A Morse complex with solid stable 1-manifolds and dashed unstable 1-manifolds. In drawing the dotted isolines we assume that all saddles have height between all minima and all maxima.

$W_p^s \cap W_q u$ are always transversal then ∇f is said to be transversal. In this case $\dim(W_p^s \cap W_q u) = \lambda_p - \lambda_q + 1$ and the number of gradient curves joining two critical points whose Morse index differs by 1 is finite.

We are mainly interested in the stable cells of maximum dimension, i.e. basins of attraction of the gradient flow to each local minimum. We then have the partition

$$(19) \quad M = \bigcup_{p \text{ a local min}} W_p^u.$$

This decomposition is sometimes used in condensed matter physics (see e.g. [Wei]) and in computational shape analysis [Reu]. In dimension two, the surface is partitioned into ‘polygons’ defined by the basins of attraction of the local minima of φ . The boundaries of these polygons are gradient lines of φ which emanate from saddle points. The vertices occur at local maxima.

An eigenfunction is a Neumann eigenfunction in each basin since the boundary is formed by integral curves of $\nabla \varphi_\lambda$. Possibly it is ‘often’ the first non-constant Neumann eigenfunction (analogously to φ_λ being the lowest Dirichlet eigenfunction in each nodal domain), but this does not seem obvious. Hence it is not clear how to relate the global eigenvalue λ^2 to the Neumann eigenvalues of the basins, which would be useful in understanding the areas or diameters of these domains. Note that

$$\int_{W_p^u} \varphi_j dV = \int_{\partial W_p^u} \nabla \varphi_\lambda \cdot \nu dS = 0,$$

where ν is the unit normal to ∂W_p^u , since $\nabla \varphi_\lambda$ is tangent to the boundary. In particular, the intersection $Z_{\varphi_\lambda} \cap W_p^u$ is non-empty and is a connected

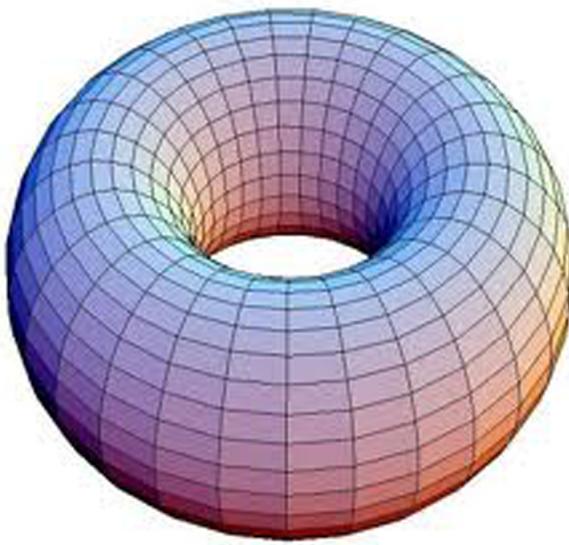


FIGURE 2

hypersurface which separates W_p^s into two components on which φ_λ has a fixed sign. To our knowledge, there do not exist rigorous results bounding the number of local minima from above or below, i.e. there is no analogue of the Courant upper bound for the number of local minima basins. It is possible to obtain statistical results on the asymptotic expected number of local minima, say for random spherical harmonics of degree N . The methods of [DSZ] adapt to this problem if one replaces holomorphic Szegö kernels by spectral projections (see also [Nic].) Thus, in a statistical sense it is much simpler to count the number of “Neumann domains” or Morse-Smale basins than to count nodal domains as in [NS].

3. Examples

Before proceeding to rigorous results, we go over a number of explicitly solvable examples. Almost by definition, they are highly non-generic and in fact represent the eigenfunctions of quantum integrable systems. Aside from being explicitly solvable, the eigenfunctions of this section are extremals for a number of problems.

3.1. Flat tori. The basic real valued eigenfunctions are $\varphi_k(x) = \sin\langle k, x \rangle$ or $\cos\langle k, x \rangle$ ($k \in \mathbb{Z}^n$) on the flat torus $\mathbf{T} = \mathbb{R}^n/\mathbb{Z}^n$. The zero set consists of the hyperplanes $\langle k, x \rangle = 0 \bmod 2\pi$ or in other words $\langle x, \frac{k}{|k|} \rangle \in \frac{1}{2\pi|k|}\mathbb{Z}$. Thus the normalized delta function $\frac{1}{|k|}dS|_{Z_{\varphi_k}}$ tends to uniform distribution along rays in the lattice \mathbb{Z}^n . The lattice arises as the joint spectrum of the commuting operators $D_j = \frac{\partial}{i\partial x_j}$ and is a feature of quantum integrable systems.

The critical point equation for $\cos\langle k, x \rangle$ is $k \sin\langle k, x \rangle = 0$ and is thus the same as the nodal equation. In particular, the critical point sets are hypersurfaces in this case. There is just one critical value = 1.

Instead of the square torus we could consider \mathbb{R}^n/L where $L \subset \mathbb{R}^n$ is a lattice of full rank. Then the joint spectrum becomes the dual lattice L^* and the eigenfunctions are $\cos\langle k, x \rangle, \sin\langle k, x \rangle$ with $k \in L^*$.

The real eigenspace $\mathcal{H}_\lambda = \mathbb{R} - \text{span}\{\sin\langle k, x \rangle, \cos\langle k, x \rangle : |k| = \lambda\}$ is of multiplicity 2 for generic L but has unbounded multiplicity in the case of $L = \mathbb{Z}^n$ and other rational lattices. In that case, one may take linear combinations of the basic eigenfunctions and study their nodal and critical point sets. For background, some recent results and further references we refer to [BZ].

3.2. Spherical harmonics on S^2 . The spectral decomposition for the Laplacian is the orthogonal sum of the spaces of spherical harmonics of degree N ,

$$(20) \quad L^2(S^2) = \bigoplus_{N=0}^{\infty} V_N, \quad \Delta|_{V_N} = \lambda_N Id.$$

The eigenvalues are given by $\lambda_N^{S^2} = N(N + 1)$ and the multiplicities are given by $m_N = 2N + 1$. A standard basis is given by the (complex valued) spherical harmonics Y_m^N which transform by $e^{im\theta}$ under rotations preserving the poles.

The Y_m^N are complex valued, so we study the nodal sets of their real and imaginary parts. They are separable, i.e. factor as $C_{N,m} P_m^N(r) \sin(m\theta)$ (resp. $\cos(m\theta)$) where P_m^N is an associated Legendre function. Thus the nodal sets of these special eigenfunctions form a checkerboard pattern that can be explicitly determined from the known behavior of zeros of associated Legendre functions. See the first image in the illustration below.

Among the basic spherical harmonics, there are two special ones: the zonal spherical harmonics (i.e. the rotationally invariant harmonics) and the highest weight spherical harmonics. Their nodal sets and intensity plots are graphed in the bottom two images, respectively.

Since the zonal spherical harmonics Y_0^N on S^2 are real-valued and rotationally invariant, their zero sets consist of a union of circles, i.e. orbits of the S^1 rotation action around the third axis. It is well known that $Y_0^N(r) = \sqrt{\frac{(2N+1)}{2\pi}} P_N(\cos r)$, where P_N is the N th Legendre function and the normalizing constant is chosen so that $\|Y_0^N\|_{L^2(S^2)} = 1$, i.e. $4\pi \int_0^{\pi/2} |P_N(\cos r)|^2 dv(r) = 1$, where $dv(r) = \sin r dr$ is the polar part of the area form. Thus the circles occur at values of r so that $P_N(\cos r) = 0$. All zeros of $P_N(x)$ are real and it has N zeros in $[-1, 1]$. It is classical that the zeros r_1, \dots, r_N of $P_N(\cos r)$ in $(0, \pi)$ become uniformly distributed with

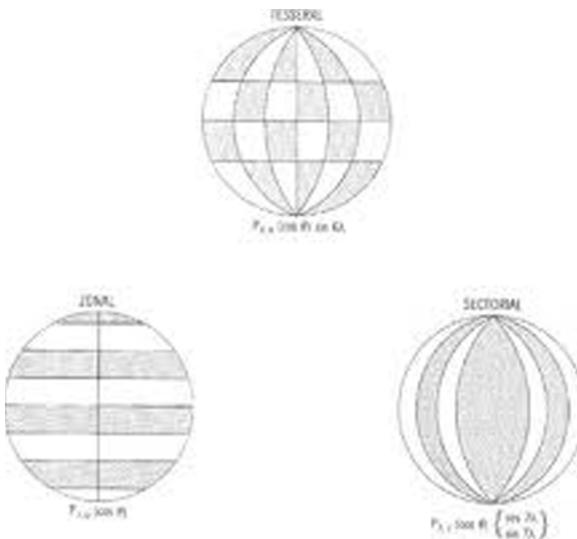


FIGURE 3. Examples of the different kinds of spherical harmonics.

respect to dr [Sz]. It is also known that P_N has $N - 1$ distinct critical points [C, Sz2] and so the critical points of Y_0^N is a union of $N - 1$ latitude circles.

We now consider real or imaginary parts of highest weight spherical harmonics Y_N^N . Up to a scalar multiple, $Y_N(x_1, x_2, x_3) = (x_1 + ix_2)^N$ as a harmonic polynomial on \mathbb{R}^3 . It is an example of a Gaussian beams along a closed geodesic γ (such as exist on equators of convex surfaces of revolution). See [R] for background on Gaussian beams on Riemannian manifolds.

The real and imaginary parts are of the form $P_N^N(\cos r) \cos N\theta$, $P_N^N(\cos r) \sin N\theta$ where $P_N^N(x)$ is a constant multiple of $(1 - x^2)^{N/2}$ so $P_N^N(\cos r) = (\sin r)^N$. The factors $\sin N\theta, \cos N\theta$ have N zeros on $(0, 2\pi)$. The Legendre functions satisfy the recursion relation $P_{\ell+1}^{\ell+1} = -(2\ell + 1)\sqrt{1 - x^2}P_\ell^\ell(x)$ with $P_0^0 = 1$ and therefore have no real zeros away from the poles. Thus, the nodal set consists of N circles of longitude with equally spaced intersections with the equator.

The critical points are solutions of the pair of equations $\frac{d}{dr}P_N^N(r) \cos N\theta = 0, P_N^N \sin N\theta = 0$. Since P_N^N has no zeros away from the poles, the second equation forces the zeros to occur at zeros of $\sin N\theta$. But then $\cos N\theta \neq 0$ so the zeros must occur at the zeros of $\frac{d}{dr}P_N^N(r)$. The critical points only occur when $\sin r = 0$ or $\cos r = 0$ on $(0, \pi)$. There are critical points at the poles where Y_N^N vanishes to order N and there is a local maximum at the value $r = \frac{\pi}{2}$ of the equator. Thus, $\operatorname{Re} Y_N^N$ has N isolated critical points on the equator and multiple critical points at the poles.

We note that $|\operatorname{Re} Y_N^N|^2$ is a Gaussian bump with peak along the equator in the radial direction. Its radial Gaussian decay implies that it extremely small outside a $N^{\frac{1}{2}}$ tube around the equator. The complement of this tube

is known in physics as the classically forbidden region. We see that the nodal set stretches a long distance into the classically forbidden region. This creates problems for nodal estimates since exponentially small values (in terms of the eigenvalue) are hard to distinguish from zeros. On the other hand, it has only two (highly multiple) critical points away from the equator.

3.3. Random spherical harmonics and chaotic eigenfunctions.

The examples above exhibit quite disparate behavior but all are eigenfunctions of quantum integrable systems. We do not review the general results in this case but plan to treat this case in an article in preparation [Z9].

Figure 4 contrasts the nodal set behavior with that of random spherical harmonics (left) and a chaotic billiard domain (the graphics are due to E. J. Heller).

4. Lower bounds on hypersurface areas of nodal sets and level sets in the C^∞ case

In this section we review the lower bounds on $\mathcal{H}^{n-1}(Z_{\varphi_\lambda})$ from [CM, SoZ, SoZa, HS, HW]. Here

$$\mathcal{H}^{n-1}(Z_{\varphi_\lambda}) = \int_{Z_{\varphi_\lambda}} dS$$

is the Riemannian surface measure, where dS denotes the Riemannian volume element on the nodal set, i.e. the insert $\text{iota}_n dV_g$ of the unit normal into the volume form of (M, g) . The main result is:

THEOREM 4.1. *Let (M, g) be a C^∞ Riemannian manifold. Then there exists a constant C independent of λ such that*

$$C\lambda^{1-\frac{n-1}{2}} \leq \mathcal{H}^{n-1}(Z_{\varphi_\lambda}).$$

We sketch the proof of Theorem 4.1 from [SoZ, SoZa]. The starting point is an identity from [SoZ] (inspired by an identity in [Dong]):

PROPOSITION 4.2. *For any $f \in C^2(M)$,*

$$(21) \quad \int_M |\varphi_\lambda| (\Delta_g + \lambda^2) f \, dV_g = 2 \int_{Z_{\varphi_\lambda}} |\nabla_g \varphi_\lambda| f \, dS,$$

This identity can be used to obtain some rudimentary but non-trivial information on the limit distribution of nodal sets in the C^∞ case; see §4.9. For the moment we only use it to study hypersurface measures of nodal sets. When $f \equiv 1$ we obtain

COROLLARY 4.3.

$$(22) \quad \lambda^2 \int_M |\varphi_\lambda| \, dV_g = 2 \int_{Z_{\varphi_\lambda}} |\nabla_g \varphi_\lambda| \, dS,$$

The lower bound of Theorem 4.1 follows from the identity in Corollary 4.3 and the following lemma:

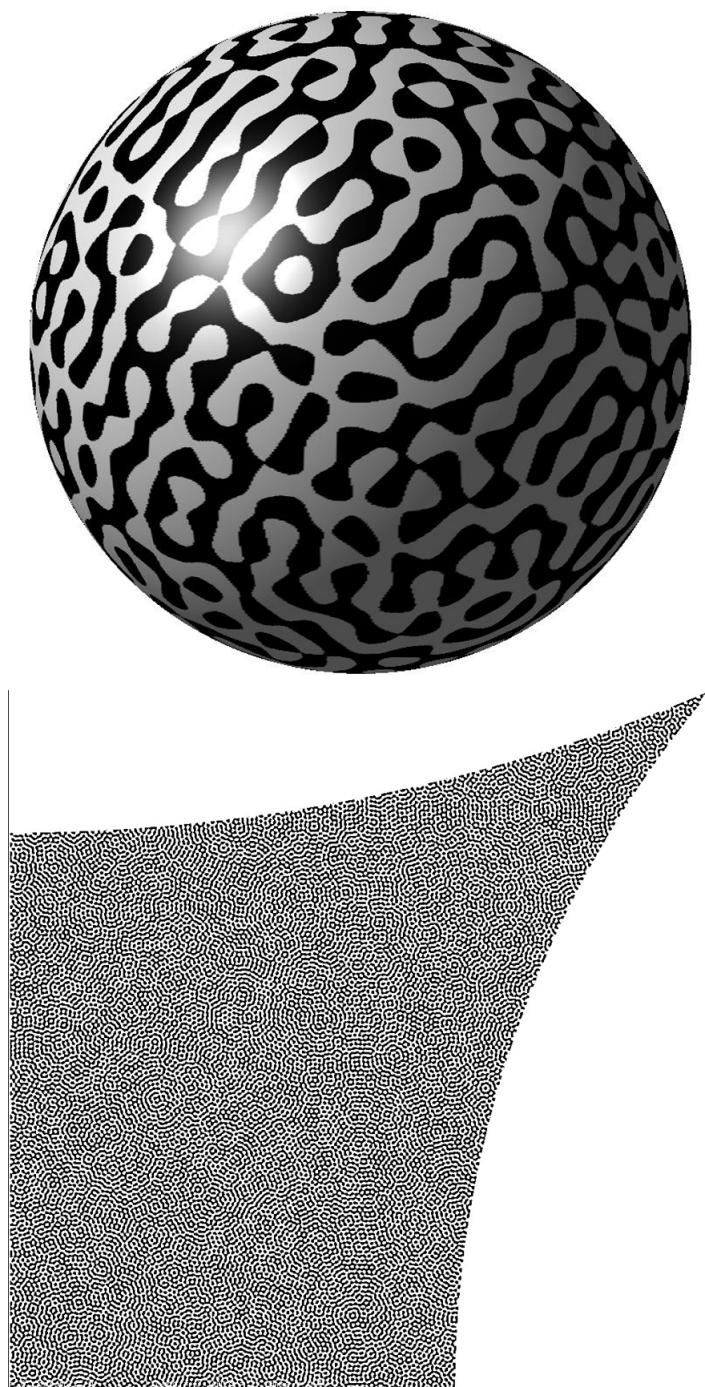


FIGURE 4

LEMMA 4.4. *If $\lambda > 0$ then*

$$(23) \quad \|\nabla_g \varphi_\lambda\|_{L^\infty(M)} \lesssim \lambda^{1+\frac{n-1}{2}} \|\varphi_\lambda\|_{L^1(M)}$$

Here, $A(\lambda) \lesssim B(\lambda)$ means that there exists a constant independent of λ so that $A(\lambda) \leq CB(\lambda)$.

By Lemma 4.4 and Corollary 4.3, we have

$$(24) \quad \begin{aligned} \lambda^2 \int_M |\varphi_\lambda| dV &= 2 \int_{Z_\lambda} |\nabla_g \varphi_\lambda|_g dS \leq 2 \mathcal{H}^{n-1}(Z_\lambda) \|\nabla_g \varphi_\lambda\|_{L^\infty(M)} \\ &\lesssim 2 \mathcal{H}^{n-1}(Z_\lambda) \lambda^{1+\frac{n-1}{2}} \|\varphi_\lambda\|_{L^1(M)}. \end{aligned}$$

Thus Theorem 4.1 follows from the somewhat curious cancellation of $\|\varphi_\lambda\|_{L^1}$ from the two sides of the inequality.

4.1. Proof of Proposition 4.2. We begin by recalling the co-area formula: Let $f : M \rightarrow \mathbb{R}$ be Lipschitz. Then for any continuous function u on M ,

$$\int_M u(x) dV = \int_{\mathbb{R}} \left(\int_{f^{-1}(y)} u \frac{dV}{df} \right) dy.$$

Equivalently,

$$\int_M u(x) \|\nabla f\| dV = \int_{\mathbb{R}} \left(\int_{f^{-1}(y)} u d\mathcal{H}^{n-1} \right) dy.$$

We refer to $\frac{dV}{df}$ as the “Leray form” on the level set $\{f = y\}$. Unlike the Riemannian surface measure $dS = d\mathcal{H}^{n-1}$ it depends on the choice of defining function f . The surface measures are related by $d\mathcal{H}^{n-1} = |\nabla f| \frac{dV}{df}$. For background, see Theorem 1.1 of [HL].

There are several ways to prove the identity of Lemma 4.2. One way to see it is that $d\mu_\lambda := (\Delta + \lambda^2)|\varphi_\lambda|dV = 0$ away from $\{\varphi_\lambda = 0\}$. Hence this distribution is a positive measure supported on Z_{φ_λ} . To determine the coefficient of the surface measure dS we calculate the limit as $\delta \rightarrow 0$ of the integral

$$\int_M f(\Delta + \lambda^2)|\varphi_\lambda| dV = \int_{|\varphi_\lambda| \leq \delta} f(\Delta + \lambda^2)|\varphi_\lambda| dV.$$

Here $f \in C^2(M)$ and with no loss of generality we may assume that δ is a regular value of φ_λ (by Sard’s theorem). By the Gauss-Green theorem,

$$\begin{aligned} &\int_{|\varphi_\lambda| \leq \delta} f(\Delta + \lambda^2)|\varphi_\lambda| dV - \int_{|\varphi_\lambda| \leq \delta} |\varphi_\lambda|(\Delta + \lambda^2)f dV \\ &= \int_{|\varphi_\lambda| = \delta} (f \partial_\nu |\varphi_\lambda| - |\varphi_\lambda| \partial_\nu f) dS. \end{aligned}$$

Here, ν is the outer unit normal and ∂_ν is the associated directional derivative. For $\delta > 0$, we have

$$(25) \quad \nu = \frac{\nabla \varphi_\lambda}{|\nabla \varphi_\lambda|} \text{ on } \{\varphi_\lambda = \delta\}, \quad \nu = -\frac{\nabla \varphi_\lambda}{|\nabla \varphi_\lambda|} \text{ on } \{\varphi_\lambda = -\delta\}.$$

Letting $\delta \rightarrow 0$ (through the sequence of regular values) we get

$$\int_M f(\Delta + \lambda^2)|\varphi_\lambda|dV = \lim_{\delta \rightarrow 0} \int_{|\varphi_\lambda| \leq \delta} f(\Delta + \lambda^2)|\varphi_\lambda|dV = \lim_{\delta \rightarrow 0} \int_{|\varphi_\lambda| = \delta} f \partial_\nu |\varphi_\lambda| dS.$$

Since $|\varphi_\lambda| = \pm \varphi_\lambda$ on $\{\varphi_\lambda = \pm \delta\}$ and by (25), we see that

$$\begin{aligned} \int_M f(\Delta + \lambda^2)|\varphi_\lambda|dV &= \lim_{\delta \rightarrow 0} \int_{|\varphi_\lambda| = \delta} f \frac{\nabla |\varphi_\lambda|}{|\nabla |\varphi_\lambda||} \cdot \nabla |\varphi_\lambda| dS \\ &= \lim_{\delta \rightarrow 0} \sum_{\pm} \int_{\varphi_\lambda = \pm \delta} f |\nabla \varphi_\lambda| dS \\ &= 2 \int_{Z_{\varphi_\lambda}} f |\nabla \varphi_\lambda| dS. \end{aligned}$$

The Gauss-Green formula and limit are justified by the fact that the singular set Σ_{φ_λ} has codimension two. We refer to [SoZ] for further details.

4.2. Proof of Lemma 4.4.

PROOF. The main idea is to construct a designer reproducing kernel for φ_λ of the form

$$(26) \quad \hat{\rho}(\lambda - \sqrt{-\Delta_g})f = \int_{-\infty}^{\infty} \rho(t)e^{-it\lambda} e^{it\sqrt{-\Delta_g}} f dt,$$

with $\rho \in C_0^\infty(\mathbb{R})$. It has the spectral expansion,

$$(27) \quad \chi_\lambda f = \sum_{j=0}^{\infty} \hat{\rho}(\lambda - \lambda_j) E_j f,$$

where $E_j f$ is the projection of f onto the λ_j -eigenspace of $\sqrt{-\Delta_g}$. Then (26) reproduces φ_λ if $\hat{\rho}(0) = 1$. We denote the kernel of χ_λ by $K_\lambda(x, y)$, i.e.

$$\chi_\lambda f(x) = \int_M K_\lambda(x, y) f(y) dV(y), \quad (f \in C(M)).$$

Assuming $\hat{\rho}(0) = 1$, then

$$\int_M K_\lambda(x, y) \varphi_\lambda(y) dV(y) = \varphi_\lambda(x).$$

To obtain Lemma 4.4, we choose ρ so that the reproducing kernel $K_\lambda(x, y)$ is uniformly bounded by $\lambda^{\frac{n-1}{2}}$ on the diagonal as $\lambda \rightarrow +\infty$. It suffices to choose ρ so that $\rho(t) = 0$ for $|t| \notin [\varepsilon/2, \varepsilon]$, with $\varepsilon > 0$ less than the injectivity radius of (M, g) , then it is proved in Lemma 5.1.3 of [Sog3] that

$$(28) \quad K_\lambda(x, y) = \lambda^{\frac{n-1}{2}} a_\lambda(x, y) e^{i\lambda r(x, y)},$$

where $a_\lambda(x, y)$ is bounded with bounded derivatives in (x, y) and where $r(x, y)$ is the Riemannian distance between points. This WKB formula for $K_\lambda(x, y)$ is known as a parametrix.

It follows from (28) that

$$(29) \quad |\nabla_g K_\lambda(x, y)| \leq C\lambda^{1+\frac{n-1}{2}},$$

and therefore,

$$\begin{aligned} \sup_{x \in M} |\nabla_g \chi_\lambda f(x)| &= \sup_x \left| \int f(y) \nabla_g K_\lambda(x, y) dV \right| \\ &\leq \|\nabla_g K_\lambda(x, y)\|_{L^\infty(M \times M)} \|f\|_{L^1} \\ &\leq C\lambda^{1+\frac{n-1}{2}} \|f\|_{L^1}. \end{aligned}$$

To complete the proof of Lemma 4.4, we set $f = \varphi_\lambda$ and use that $\chi_\lambda \varphi_\lambda = \varphi_\lambda$. \square

We view $K_\lambda(x, y)$ as a designer reproducing kernel, because it is much smaller on the diagonal than kernels of the spectral projection operators $E_{[\lambda, \lambda+1]} = \sum_{j: \lambda_j \in [\lambda, \lambda+1]} E_j$. The restriction on the support of ρ removes the big singularity on the diagonal at $t = 0$. As discussed in [SoZa], it is possible to use this kernel because we only need it to reproduce one eigenfunction and not a whole spectral interval of eigenfunctions.

4.3. Modifications. After an initial modification in [HW], an interesting application of Proposition 4.2 was used in [HS] to prove

THEOREM 4.5. [HS] *For any C^∞ compact Riemannian manifold, the L^2 -normalized eigenfunctions satisfy*

$$\mathcal{H}^{n-1}(Z_{\varphi_\lambda}) \geq C \lambda \|\varphi_\lambda\|_{L^1}^2.$$

They first apply the Schwarz inequality to get

$$(30) \quad \lambda^2 \int_M |\varphi_\lambda| dV_g \leq 2(\mathcal{H}^{n-1}(Z_{\varphi_\lambda}))^{1/2} \left(\int_{Z_{\varphi_\lambda}} |\nabla_g \varphi_\lambda|^2 dS \right)^{1/2}.$$

They then use the test function

$$(31) \quad f = (1 + \lambda^2 \varphi_\lambda^2 + |\nabla_g \varphi_\lambda|^2_g)^{\frac{1}{2}}$$

in Proposition 4.2 to show that

$$(32) \quad \int_{Z_{\varphi_\lambda}} |\nabla_g \varphi_\lambda|^2 dS \leq \lambda^3.$$

A simpler approach to the last step was suggested by W. Minicozzi, who pointed out that the result also follows from the identity

$$(33) \quad 2 \int_{Z_\lambda} |\nabla_g e_\lambda|^2 dS_g = - \int_M \operatorname{sgn}(\varphi_\lambda) \operatorname{div}_g (|\nabla_g e_\lambda| \nabla_g e_\lambda) dV_g.$$

This approach is used in [Ar] to generalize the nodal bounds to Dirichlet and Neumann eigenfunctions of bounded domains.

Theorem 4.5 shows that Yau's conjectured lower bound would follow for a sequence of eigenfunctions satisfying $\|\varphi_\lambda\|_{L^1} \geq C > 0$ for some positive constant C .

4.4. More general identities. It is possible to further generalize the identity of Proposition 4.2 and we pause to record an obvious one. For any function χ , we have

$$\Delta\chi(\varphi) = \chi''(\varphi)|\nabla\varphi|^2 - \lambda^2\chi'(\varphi)\varphi.$$

We then take χ to be the meromorphic family of homogeneous distribution x_+^s . We recall that for $\operatorname{Re} a > -1$,

$$x_+^a := \begin{cases} x^a, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

The family extends to $a \in \mathbb{C}$ as a meromorphic family of distributions with simple poles at $a = -1, -2, \dots, -k, \dots$ using the equation $\frac{d}{dx}x_+^s = sx_+^{s-1}$ to extend it one unit strip at a time. One can convert x_+^s to the holomorphic family

$$\chi_+^\alpha = \frac{x_+^\alpha}{\Gamma(\alpha + 1)}, \quad \text{with } \chi_+^{-k} = \delta_0^{(k-1)}.$$

The identity we used above belongs to the family,

$$(34) \quad (\Delta + s\lambda^2)\varphi_+^s = s(s-1)|\nabla\varphi|^2\varphi_+^{s-2}.$$

Here $\varphi_+^s = \varphi^*x_+^s$ has poles at $s = -1, -2, \dots$. The calculation in (21) used $|\varphi|$ but is equivalent to using (34) when $s = 1$. Then φ_+^{s-2} has a pole when $s = 1$ with residue $\delta_0(\varphi) = \frac{dS}{|\nabla\varphi|}dS|_{Z_{\varphi_\lambda}}$; it is cancelled by the factor $s-1$ and we obtain (21). This calculation is formal because the pullback formulae are only valid when $d\varphi \neq 0$ when $\varphi = 0$, but as above they can be justified because the singular set has codimension 2. The right side also has a pole at $s = 0$ and we get $\Delta\varphi_+^0 = -|\nabla\varphi|^2\delta'(\varphi)$, which is equivalent to the divergence identity above. There are further poles at $s = -1, -2, \dots$ but they now occur on both sides of the formulae. It is possible that they have further uses.

Such identities appear to be related to the Bernstein-Kashiwara theorem that for any real analytic function f one may meromorphically extend f_+ to \mathbb{C} by constructing a family $P_s(D)$ of differential operators with analytic coefficients and a meromorphic function $b(s)$ so that $P_s(D)f^{s+1} = b(s)f^s$. In the case $f = \varphi_\lambda$, the operator $|\nabla\varphi|^{-2}(\Delta + s\lambda^2)$ accomplishes something like this, although it does not have analytic coefficients due to poles at the critical points of φ .

4.5. Other level sets. These results generalize easily to any level set $\mathcal{N}_{\varphi_\lambda}^c := \{\varphi_\lambda = c\}$. Let $\text{sgn}(x) = \frac{x}{|x|}$.

PROPOSITION 4.6. *For any C^∞ Riemannian manifold, and any $f \in C(M)$ we have,*

$$(35) \quad \int_M f(\Delta + \lambda^2) |\varphi_\lambda - c| dV + \lambda^2 c \int f \text{sgn}(\varphi_\lambda - c) dV = 2 \int_{\mathcal{N}_{\varphi_\lambda}^c} f |\nabla \varphi_\lambda| dS.$$

This identity has similar implications for $\mathcal{H}^{n-1}(\mathcal{N}_{\varphi_\lambda}^c)$ and for the equidistribution of level sets. Note that if $c > \sup |\varphi_\lambda(x)|$ then indeed both sides are zero.

COROLLARY 4.7. *For $c \in \mathbb{R}$*

$$\lambda^2 \int_{\varphi_\lambda \geq c} \varphi_\lambda dV = \int_{\mathcal{N}_{\varphi_\lambda}^c} |\nabla \varphi_\lambda| dS \leq \lambda^2 \text{Vol}(M)^{1/2}.$$

Consequently, if $c > 0$

$$\mathcal{H}^{n-1}(\mathcal{N}_{\varphi_\lambda}^c) + \mathcal{H}^{n-1}(\mathcal{N}_{\varphi_\lambda}^{-c}) \geq C_g \lambda^{2-\frac{n+1}{2}} \int_{|\varphi_\lambda| \geq c} |\varphi_\lambda| dV.$$

The Corollary follows by integrating Δ by parts, and by using the identity,

$$(36) \quad \begin{aligned} \int_M |\varphi_\lambda - c| + c \text{sgn}(\varphi_\lambda - c) dV &= \int_{\varphi_\lambda > c} \varphi_\lambda dV - \int_{\varphi_\lambda < c} \varphi_\lambda dV \\ &= 2 \int_{\varphi_\lambda > c} \varphi_\lambda dV, \end{aligned}$$

since $0 = \int_M \varphi_\lambda dV = \int_{\varphi_\lambda > c} \varphi_\lambda dV + \int_{\varphi_\lambda < c} \varphi_\lambda dV$.

4.6. Examples. The lower bound of Theorem 4.1 is far from the lower bound conjectured by Yau, which by Theorem 2.1 is correct at least in the real analytic case. In this section we go over the model examples to understand why the methods are not always getting sharp results.

4.7. Flat tori. We have, $|\nabla \sin \langle k, x \rangle|^2 = \cos^2 \langle k, x \rangle |k|^2$. Since $\cos \langle k, x \rangle = 1$ when $\sin \langle k, x \rangle = 0$ the integral is simply $|k|$ times the surface volume of the nodal set, which is known to be of size $|k|$. Also, we have $\int_{\mathbf{T}} |\sin \langle k, x \rangle| dx \geq C$. Thus, our method gives the sharp lower bound $\mathcal{H}^{n-1}(Z_{\varphi_\lambda}) \geq C \lambda^1$ in this example.

So the upper bound is achieved in this example. Also, we have $\int_{\mathbf{T}} |\sin \langle k, x \rangle| dx \geq C$. Thus, our method gives the sharp lower bound $\mathcal{H}^{n-1}(Z_{\varphi_\lambda}) \geq C \lambda^1$ in this example. Since $\cos \langle k, x \rangle = 1$ when $\sin \langle k, x \rangle = 0$ the integral is simply $|k|$ times the surface volume of the nodal set, which is known to be of size $|k|$.

4.8. Spherical harmonics on S^2 . The L^1 of Y_0^N norm can be derived from the asymptotics of Legendre polynomials

$$P_N(\cos \theta) = \sqrt{2}(\pi N \sin \theta)^{-\frac{1}{2}} \cos \left(\left(N + \frac{1}{2} \right) \theta - \frac{\pi}{4} \right) + O(N^{-3/2})$$

where the remainder is uniform on any interval $\epsilon < \theta < \pi - \epsilon$. We have

$$\|Y_0^N\|_{L^1} = 4\pi \sqrt{\frac{(2N+1)}{2\pi}} \int_0^{\pi/2} |P_N(\cos r)| dv(r) \sim C_0 > 0,$$

i.e. the L^1 norm is asymptotically a positive constant. Hence $\int_{Z_{Y_0^N}} |\nabla Y_0^N| ds \simeq C_0 N^2$. In this example $|\nabla Y_0^N|_{L^\infty} = N^{\frac{3}{2}}$ saturates the sup norm bound. The length of the nodal line of Y_0^N is of order λ , as one sees from the rotational invariance and by the fact that P_N has N zeros. The defect in the argument is that the bound $|\nabla Y_0^N|_{L^\infty} = N^{\frac{3}{2}}$ is only obtained on the nodal components near the poles, where each component has length $\simeq \frac{1}{N}$.

Gaussian beams

Gaussian beams are Gaussian shaped lumps which are concentrated on $\lambda^{-\frac{1}{2}}$ tubes $\mathcal{T}_{\lambda^{-\frac{1}{2}}}(\gamma)$ around closed geodesics and have height $\lambda^{\frac{n-1}{4}}$. We note that their L^1 norms decrease like $\lambda^{-\frac{(n-1)}{4}}$, i.e. they saturate the L^p bounds of [Sog] for small p . In such cases we have $\int_{Z_{\varphi_\lambda}} |\nabla \varphi_\lambda| dS \simeq \lambda^2 \|\varphi_\lambda\|_{L^1} \simeq \lambda^{2-\frac{n-1}{4}}$. Gaussian beams are minimizers of the L^1 norm among L^2 -normalized eigenfunctions of Riemannian manifolds. Also, the gradient bound $\|\nabla \varphi_\lambda\|_{L^\infty} = O(\lambda^{\frac{n+1}{2}})$ is far off for Gaussian beams, the correct upper bound being $\lambda^{1+\frac{n-1}{4}}$. If we use these estimates on $\|\varphi_\lambda\|_{L^1}$ and $\|\nabla \varphi_\lambda\|_{L^\infty}$, our method gives $\mathcal{H}^{n-1}(Z_{\varphi_\lambda}) \geq C \lambda^{1-\frac{n-1}{2}}$, while λ is the correct lower bound for Gaussian beams in the case of surfaces of revolution (or any real analytic case). The defect is again that the gradient estimate is achieved only very close to the closed geodesic of the Gaussian beam. Outside of the tube $\mathcal{T}_{\lambda^{-\frac{1}{2}}}(\gamma)$ of radius $\lambda^{-\frac{1}{2}}$ around the geodesic, the Gaussian beam and all of its derivatives decay like $e^{-\lambda d^2}$ where d is the distance to the geodesic. Hence $\int_{Z_{\varphi_\lambda}} |\nabla \varphi_\lambda| dS \simeq \int_{Z_{\varphi_\lambda} \cap \mathcal{T}_{\lambda^{-\frac{1}{2}}}(\gamma)} |\nabla \varphi_\lambda| dS$. Applying the gradient bound for Gaussian beams to the latter integral gives $\mathcal{H}^{n-1}(Z_{\varphi_\lambda} \cap \mathcal{T}_{\lambda^{-\frac{1}{2}}}(\gamma)) \geq C \lambda^{1-\frac{n-1}{2}}$, which is sharp since the intersection $Z_{\varphi_\lambda} \cap \mathcal{T}_{\lambda^{-\frac{1}{2}}}(\gamma)$ cuts across γ in $\simeq \lambda$ equally spaced points (as one sees from the Gaussian beam approximation).

4.9. Non-scarring of nodal sets on (M, g) with ergodic geodesic flow. The identity of Lemma 4.2 for general $f \in C^2(M)$ can be used to investigate the equidistribution of nodal sets equipped with the surface measure $|\nabla \varphi_\lambda| dS$. We denote the normalized measure by $\lambda^{-2} |\nabla \varphi_\lambda| dS|_{Z_{\varphi_\lambda}}$.

We first prove a rather simple (unpublished) result on nodal sets when the geodesic flow of (M, g) is ergodic. Since there exist many expositions of quantum ergodic eigenfunctions, we only briefly recall the main facts and definitions and refer to [Z5, Z6] for further background.

Quantum ergodicity concerns the semi-classical (large λ) asymptotics of eigenfunctions in the case where the geodesic flow G^t of (M, g) is ergodic. We recall that the geodesic flow is the Hamiltonian flow of the Hamiltonian $H(x, \xi) = |\xi|_g^2$ (the length squared) and that ergodicity means that the only G^t -invariant subsets of the unit cosphere bundle S^*M have either full Liouville measure or zero Liouville measure (Liouville measure is the natural measure on the level set $H = 1$ induced by the symplectic volume measure of T^*M).

We will say that a sequence $\{\varphi_{j_k}\}$ of L^2 -normalized eigenfunctions is *quantum ergodic* if

$$(37) \quad \langle A\varphi_{j_k}, \varphi_{j_k} \rangle \rightarrow \frac{1}{\mu(S^*M)} \int_{S^*M} \sigma_A d\mu, \quad \forall A \in \Psi^0(M).$$

Here, $\Psi^s(M)$ denotes the space of pseudodifferential operators of order s , σ_A denotes the principal symbol of A , and $d\mu$ denotes Liouville measure on the unit cosphere bundle S^*M of (M, g) . More generally, we denote by $d\mu_r$ the (surface) Liouville measure on ∂B_r^*M , defined by

$$(38) \quad d\mu_r = \frac{\omega^m}{d|\xi|_g} \text{ on } \partial B_r^*M.$$

We also denote by α the canonical action 1-form of T^*M .

The main result is that there exists a subsequence $\{\varphi_{j_k}\}$ of eigenfunctions whose indices j_k have counting density one for which $\rho_{j_k}(A) := \langle A\varphi_{j_k}, \varphi_{j_k} \rangle \rightarrow \omega(A)$ (where as above $\omega(A) = \frac{1}{\mu(S^*M)} \int_{S^*M} \sigma_A d\mu$ is the normalized Liouville average of σ_A). The key quantities to study are the quantum variances

$$(39) \quad V_A(\lambda) := \frac{1}{N(\lambda)} \sum_{j: \lambda_j \leq \lambda} |\langle A\varphi_j, \varphi_j \rangle - \omega(A)|^2.$$

The following result is the culmination of the results in [Sh.1, Z1, CV, ZZw, GL].

THEOREM 4.8. *Let (M, g) be a compact Riemannian manifold (possibly with boundary), and let $\{\lambda_j, \varphi_j\}$ be the spectral data of its Laplacian Δ . Then the geodesic flow G^t is ergodic on $(S^*M, d\mu)$ if and only if, for every $A \in \Psi^0(M)$, we have:*

- (1) $\lim_{\lambda \rightarrow \infty} V_A(\lambda) = 0$.
- (2) $(\forall \epsilon)(\exists \delta) \limsup_{\lambda \rightarrow \infty} \frac{1}{N(\lambda)} \sum_{\substack{j \neq k: \lambda_j, \lambda_k \leq \lambda \\ |\lambda_j - \lambda_k| < \delta}} |(A\varphi_j, \varphi_k)|^2 < \epsilon$

Since all the terms in (1) are positive, no cancellation is possible, hence (1) is equivalent to the existence of a subset $\mathcal{S} \subset \mathbb{N}$ of density one such that $\mathcal{Q}_{\mathcal{S}} := \{d\Phi_k : k \in \mathcal{S}\}$ has only ω as a weak* limit point.

We now consider nodal sets of quantum ergodic eigenfunctions. The following result says that if we equip nodal sets with the measure $\frac{1}{\lambda_j^2} |\nabla \varphi_{\lambda_j}| dS$, then nodal sets cannot ‘scar’, i.e. concentrate singularly as $\lambda_j \rightarrow \infty$.

PROPOSITION 4.9. *Suppose that $\{\varphi_{\lambda_j}\}$ is a quantum ergodic sequence. Then any weak limit of $\{\frac{1}{\lambda_j^2} |\nabla \varphi_{\lambda_j}| dS\}$ must be absolutely continuous with respect to dV_g .*

(In fact this will be improved below in Corollary 4.13).

The first point is the following

LEMMA 4.10. *The weak * limits of the sequence $\{\lambda^{-2} |\nabla \varphi_{\lambda_j}| dS|_{Z_{\varphi_{\lambda_j}}}\}$ of bounded positive measures are the same as the weak * limits of $\{|\varphi_{\lambda_j}|\}$ (against $f \in C(M)$).*

We let $f \in C^2(M)$ and multiply the identity of Proposition 4.2 by λ^{-2} . We then integrate by parts to put Δ on f . This shows that for $f \in C^2(M)$, we have

$$\int_M f |\varphi_{\lambda_j}| dV = \lambda^{-2} \int_{Z_{\varphi_{\lambda_j}}} f |\nabla \varphi_{\lambda_j}| dS + O(\lambda^{-2}).$$

Letting $f = 1$, we see that the family of measures $\{\lambda^{-2} |\nabla \varphi_{\lambda_j}|^2 \delta(\varphi_{\lambda_j})\}$ is bounded. By uniform approximation of $f \in C(M)$ by elements of $C^2(M)$, we see that the weak* limit formula extends to $C(M)$.

LEMMA 4.11. *Suppose that $\{\varphi_{\lambda_j}\}$ is a quantum ergodic sequence. Then any weak limit of $\{|\varphi_{\lambda_j}| dS\}$ must be absolutely continuous with respect to dV .*

We recall that a sequence of measures μ_n converges weak* to μ if $\int_M f d\mu_n \rightarrow \int f d\mu$ for all continuous f . A basic fact about weak* convergence of measures is that $\int f d\mu_n \rightarrow \int f d\mu$ for all $f \in C(M)$ implies that $\mu_n(E) \rightarrow \mu(E)$ for all sets E with $\mu(\partial E) = 0$ (Portmanteau theorem).

We also recall that a sequence of eigenfunctions is called quantum ergodic (in the base) if

$$(40) \quad \int f |\varphi_{\lambda_j}|^2 dV \rightarrow \frac{1}{Vol(M)} \int_M f dV.$$

In other words, $\varphi_{\lambda_j}^2 \rightarrow 1$ in the weak* topology, i.e. the vague topology on measures. We now prove Lemma 4.11.

PROOF. Suppose that $|\varphi_{\lambda_{j_k}}| dV \rightarrow d\mu$ and assume that $d\mu = cdV + d\nu$ where $d\nu$ is singular with respect to dV . Let $\Sigma = \text{supp } \nu$, and let $\sigma = \mu(\Sigma) = \nu(\Sigma)$. Let T_ϵ be the ϵ -tube around Σ . Then

$$\lim_{k \rightarrow \infty} \int_{T_\epsilon} |\varphi_{\lambda_{j_k}}| dV = cVol(T_\epsilon) + \nu(\Sigma) = \sigma + O(\epsilon).$$

But for any set $\Omega \subset M$, $\int_{\Omega} |\varphi_{\lambda_j}| dV \leq \sqrt{Vol(\Omega)} \sqrt{\int_{\Omega} |\varphi_{\lambda_j}|^2 dV}$. Hence if $Vol(\partial\Omega) = 0$, $\limsup_{j \rightarrow \infty} \int_{\Omega} |\varphi_{\lambda_j}| dV \leq Vol(\Omega)$. Letting $\Omega = \mathcal{T}_\epsilon(\Sigma)$ we get $\sigma + O(\epsilon) \leq Vol(\mathcal{T}_\epsilon(\Sigma)) = O(\epsilon)$ since $\lim_{k \rightarrow \infty} \int_{\mathcal{T}_\epsilon} |\varphi_{\lambda_{j_k}}|^2 dV = Vol(\mathcal{T}_\epsilon) = O(\epsilon)$. Letting $\epsilon \rightarrow 0$ gives a contradiction. \square

Of course, it is possible that the only weak* limit is zero.

4.10. A stronger non-scarring result. G. Rivière [Ri] pointed out some improvements to Proposition 4.9.

PROPOSITION 4.12. *Suppose that $\{\varphi_{j_k}\}$ is a sequence of L^2 normalized eigenfunctions satisfying the following ‘weak quantum ergodic’ condition:*

$$(WQE) : |\varphi_{j_k}|^2 dV_g \rightarrow \rho dV_g \text{ weak*}, \text{ with } \rho \in L^\infty(M, dV_g).$$

Suppose also that

$$|\varphi_{j_k}| dV_g \rightarrow d\mu,$$

where μ is a probability measure on M . Then $\mu = F dV_g$ with $F \in L^\infty(M, dV_g)$.

In fact, the same is true for weak limits of $|\varphi_{j_k}|^p dV_g$ for any $1 \leq p < 2$, but we only treat the case $p = 1$.

PROOF. If $f \in C(M)$ then

$$\left| \int_M f |\varphi_{j_k}| dV_g \right| \leq \left| \int_M (|f|^{\frac{1}{2}} |\varphi_{j_k}|) |f|^{\frac{1}{2}} dV_g \right| \leq \left| \int_M |f| |\varphi_{j_k}|^2 dV_g \right|^{\frac{1}{2}} \|f\|_{L^1}^{\frac{1}{2}}.$$

Let $k \rightarrow \infty$ and we get

$$\left| \int_M f d\mu \right| \leq \|\rho\|_{L^\infty} \|f\|_{L^1}.$$

Hence $f \rightarrow \int f d\mu$ is a continuous linear functional on L^1 and must have the form $\mu = F dV_g$ where $\|F\|_\infty \leq \|\rho\|_\infty$. \square

COROLLARY 4.13. *Suppose that $\{\varphi_{\lambda_j}\}$ is a quantum ergodic sequence. Then any weak limit of $\{\frac{1}{\lambda_j^2} |\nabla \varphi_{\lambda_j}| dS\}$ must be of the form $F dV_g$ with $|F| \leq 1$.*

4.11. Weak* limits for L^∞ quantum ergodic sequences. To our knowledge, the question whether the limit (4.8) holds $f \in L^\infty$ when (M, g) has ergodic geodesic flow has not been studied. It is equivalent to strengthening the Portmanteau statement to all measurable sets E , and is equivalent to the statement that $\{\varphi_{\lambda_j}^2\} \rightarrow 1$ weakly in L^1 . We call such sequences L^∞ quantum ergodic on the base. The term ‘on the base’ refers to the fact that we only demand quantum ergodicity for the projections of the ‘microlocal lifts’ to the base M . For instance, the exponential eigenfunctions of flat tori are L^∞ quantum ergodic in this sense.

LEMMA 4.14. Suppose that $\{\varphi_j\}$ is an L^∞ - quantum ergodic sequence. Then there exists $\epsilon > 0$ so that $\|\varphi_j\|_{L^1} \geq \epsilon > 0$ for all j .

PROOF. We argue by contradiction. If the conclusion were false, there would exist a subsequence $\varphi_{j_k} \rightarrow 0$ strongly in L^1 , but with $\varphi_{j_k}^2 dV \rightarrow dV$ weakly in L^1 . The first assumption implies the existence of a subsequence (which we continue to denote by φ_{j_k}) satisfying $\varphi_{j_k} \rightarrow 0$ a.e. dV . But L^1 has the weak Banach-Saks property: any weakly convergent sequence in L^1 has a subsequence whose arithmetic means converge strongly (Szlenk's weak Banach-Saks theorem for L^1). We choose such a subsequence for φ_{j_k} and continue to denote it as φ_{j_k} . This subsequence has the properties that

- (1) $\varphi_{j_k} \rightarrow 0$ a.e.
- (2) $\psi_N := \frac{1}{N} \sum_{k \leq N} \varphi_{j_k}^2 \rightarrow 1$ strongly in L^1 .

But $\psi_N(x) \rightarrow 0$ on the same set where $\varphi_{j_k}(x) \rightarrow 0$, hence by (1) $\psi_N \rightarrow 0$ a.s. This contradicts (2) and completes the proof. \square

Combining with the above, we have

COROLLARY 4.15. Suppose that $\{\varphi_{\lambda_j}\}$ is an L^∞ quantum ergodic sequence on the base. Then the conjectured Yau lower bound holds: $\mathcal{H}^{n-1}(Z_{\varphi_\lambda}) \geq C_g \lambda$ for some $C_g > 0$.

We also see that the limits in Proposition 4.9 are non-zero:

COROLLARY 4.16. Suppose that $\{\varphi_{\lambda_j}\}$ is an L^∞ quantum ergodic sequence on the base. Then there exists $C > 0$ so that any weak limit of the sequence $\frac{1}{\lambda^2} |\nabla \varphi_{\lambda_j}| dS|_{Z_{\varphi_{\lambda_j}}}$ has mass $\geq C > 0$.

Of course, such an abstract functional analysis argument only serves a purpose if we can prove that eigenfunctions of Δ are L^∞ quantum ergodic on the base in interesting cases. It is natural to conjecture that this condition holds on negatively curved manifolds, since the expected L^1 norm of a random wave is bounded below by a positive constant. The main problem is that $L^\infty(M)$ is a non-separable Banach space. The standard quantum ergodicity arguments show that (when quantum ergodicity is valid), for any Borel set E there exists a subsequence \mathcal{S}_E of density one so that

$$(41) \quad \lim_{k \rightarrow \infty, j_k \in \mathcal{S}_E} \int_E \varphi_{j_k}^2 dV = Vol(E).$$

However, the non-separability of $L^\infty(M)$ means that one cannot use the diagonalization argument of [Z1, CV] to show that there exists a density one subsequence independent of E so that (41) holds. If L^∞ quantum ergodicity fails, then zero-density subsequences of eigenfunctions would ‘scar’ along Cantor sets C of positive measure. That is, the mass $\int_C \varphi_{j_k}^2 dV$ may tend to a larger value than $Vol(C)$.

Equidistributed sums of Gaussian beams and quantum ergodicity

We briefly consider the question whether it is possible to have a quantum ergodic sequence of eigenfunctions for which $\|\varphi_j\|_{L^1} \rightarrow 0$.

First, we observe that there do exist sequences of quantum ergodic functions (not eigenfunctions) with this property: $\sum_{j=1}^{M(n)} \sqrt{\frac{n}{M(n)}} \chi_{[x_j(n), x_j(n) + \frac{1}{n}]} \rightarrow 0$ in $L^1([0, 1], dx)$ as long as $M(n) = o(n)$. But its square is the probability measure $\frac{1}{M(n)} \sum_{j=1}^{M(n)} n \chi_{[x_j(n), x_j(n) + \frac{1}{n}]}$ and if the $\{x_j(n)\}$ are uniformly distributed in $[0, 1]$ (w.r.t. dx), this tends weakly to dx .

It is tempting to construct sequences of eigenfunctions with the same property: a Gaussian beam Y_γ^N on the standard S^2 associated to a closed geodesic γ (i.e. a rotate of Y_N^N) is of height $\lambda^{\frac{1}{2}}$ in a tube of radius $\sqrt{\lambda}$ around γ . If we let $M(N) = o(N^{\frac{1}{2}})$ and choose $M(N)$ closed geodesics which are $\frac{1}{\sqrt{M(N)}}$ -separated, and become equidistributed in the space of closed geodesics, then $\varphi_N = \frac{1}{\sqrt{M(N)}} \sum_{j=1}^{M(N)} Y_{\gamma_j}^N$ is an eigenfunction whose L^1 -norm tends to zero like $\sqrt{M(N)} N^{-\frac{1}{4}}$ but whose L^2 norm is asymptotic to 1 and whose modulus square tends weak* to 1. More precisely, $\frac{1}{M} \sum_{j=1}^{M(N)} |Y_{\gamma_j}^N|^2 \rightarrow 1$ weakly. To prove that $|\varphi_N|^2 \rightarrow 1$ requires proving that $\frac{1}{M(N)} \sum_{j \neq k} Y_{\gamma_j}^N \overline{Y_{\gamma_k}^N} \rightarrow 0$. The sum is over $\sim M(N)^2$ terms which are exponentially outside the tube intersections $T_{\lambda^{-\frac{1}{2}}}(\gamma_j) \cap T_{\lambda^{-\frac{1}{2}}}(\gamma_k)$. In the sum we may fix $j = j_0$ and multiply by $M(N)$. So we need then to show that $\sum_{k \neq j_0} |\langle Y_{\gamma_{j_0}}^N, Y_{\gamma_k}^N \rangle| \rightarrow 0$. The geodesics are well-separated if the distance in the space of geodesics between them is $\geq \frac{1}{\sqrt{M(N)}}$, which means that the angle between γ_j and γ_k is at least this amount. When the angle is $\geq \epsilon$ then the inner product $|\langle Y_{\gamma_j}^N, Y_{\gamma_k}^N \rangle| \leq \frac{1}{\epsilon} N^{-1}$ since the area of $T_{\lambda^{-\frac{1}{2}}}(\gamma_j) \cap T_{\lambda^{-\frac{1}{2}}}(\gamma_k)$ is bounded by this amount. For any ϵ the sum over geodesics separated by ϵ is $O(\frac{1}{\epsilon} M(N) N^{-1})$. The remaining number of terms is $O(\epsilon^2 M(N))$. So if $\epsilon = o(\sqrt{M(N)})$ both terms tend to zero.

4.12. Intersections of nodal sets of orthogonal eigenfunctions.

A related question is whether nodal sets of orthogonal eigenfunctions of the same eigenvalue must intersect. Of course, this question only arises when the eigenvalue has multiplicity > 1 . A result of this kind was obtained by V. Gichev under a topological condition on M .

THEOREM 4.17. [Gi] *Suppose that $H^1(M) = 0$ and that $\varphi_{\lambda,1}, \varphi_{\lambda,2}$ are orthogonal eigenfunctions with the same eigenvalue λ^2 . Then $Z_{\varphi_{\lambda,1}} \cap Z_{\varphi_{\lambda,2}} \neq \emptyset$.*

We briefly sketch the proof: Let \mathcal{A}_1 resp. \mathcal{A}_2 be the family of nodal domains of $\varphi_{\lambda,1}$ resp. $\varphi_{\lambda,2}$. Each union $\bigcup_{W \in \mathcal{A}_j} W$ covers M up to the

nodal set of $\varphi_{\lambda,j}$. If the nodal sets do not intersect then the nodal set of $\varphi_{\lambda,2}$ is contained in $\bigcup_{W \in \mathcal{A}_1} W$, for instance; similarly if the indices are reversed. Hence the nodal sets have empty intersection if and only if $\bigcup_{W \in \mathcal{A}_1} W \cup \bigcup_{W \in \mathcal{A}_2} W$ covers M . Under this condition, Gichev constructs a closed 1-form which is not exact by showing that the incidence graph of the cover obtained from the union of the nodal domains of $\varphi_{\lambda,1}$ and $\varphi_{\lambda,2}$ contains a cycle. He then considers a nodal domain U of $\varphi_{\lambda,1}$ and a nodal domain V of $\varphi_{\lambda,2}$ which intersect. Let $Q = \partial U \cap V$. Since $Q \cap \partial V \neq \emptyset$ there exists a smooth function f on M such that $f \equiv 1$ in a neighborhood of Q and $f = 0$ near $\partial U \setminus Q$. Let η be the one form which equals df on U and 0 on the complement of U . Clearly η is closed and it is verified in [Gi] that η is not exact.

Givech also proves that for S^2 , if 0 is a regular value of $\varphi_{\lambda,1}$ then $\#Z_{\varphi_{\lambda,1}} \cap Z_{\varphi_{\lambda,2}} \geq 2$ for every orthogonal eigenfunction $\varphi_{\lambda,2}$ with the same eigenvalue. The proof is simply to use Green's formula for a nodal domain for $\varphi_{\lambda,1}$ and note that the integral of $\varphi_{\lambda,2} \frac{\partial}{\partial \nu} \varphi_{\lambda,1}$ equals zero on its boundary.

A related observation is the curious identity of [SoZ], which holds for any (M, g) : for any pair of eigenfunctions,

$$(\lambda_j^2 - \lambda_k^2) \int_M \varphi_{\lambda_k} |\varphi_{\lambda_j}| dV = 2 \int_{Z_{\varphi_{\lambda_j}}} \varphi_{\lambda_k} |\nabla \varphi_{\lambda_j}| dS.$$

Hence for a pair of orthogonal eigenfunctions of the same eigenvalue,

$$\int_{Z_{\varphi_{\lambda_j}}} \varphi_{\lambda_k} |\nabla \varphi_{\lambda_j}| dS = 0.$$

5. Norms and nodal sets

Studies of nodal sets often involve dual studies of L^p norms of eigenfunctions. In this section, we review a number of relatively recent results on L^p norms, both in the global manifold M and for restrictions of eigenfunctions to submanifolds.

5.1. Polterovich-Sodin on norms and nodal sets. Let $\mathcal{A}(\varphi_\lambda)$ denote the collection of nodal domains of φ_λ . For $A \in \mathcal{A}(\varphi_\lambda)$ let $m_A = \max_A |\varphi_\lambda|$. In [PS] the following is proved (see Corollary 1.7):

THEOREM 5.1. [PS] *Let (M, g) be a C^∞ Riemannian surface. For every φ_λ with $\|\varphi_\lambda\| = 1$,*

$$\sum_{A \in \mathcal{A}} m_A^6 \leq k_g \lambda^3.$$

Hence, for each $a > 0$, the number of nodal domains A of φ_λ where the maximal bound $m_A \geq a\lambda^{1/2}$ is achieved in order of magnitude does not exceed $k_g a^{-6}$. In particular, for fixed a , it remains bounded as $\lambda \rightarrow \infty$.

The proof uses a certain Banach indicatrix, the Sogge L^6 bounds, and estimates on the inradius of nodal domains. For a continuous function $u \in C(\mathbb{R})$, the generalized Banach indicatrix is defined by

$$B(u, f) = \int_{-\infty}^{+\infty} u(c)\beta(c, f)dc,$$

where for a regular value $c \in \mathbb{R}$ of f , $\beta(c, f)$ is the number of connected components of $f^{-1}(c)$. In [PS], the integral $B(u, f)$ is bounded from above through the L^2 -norms of the function f and Δf . I.e.. in Theorem 1.3. For any $f \in \mathcal{F}_\lambda$ and any continuous function u on \mathbb{R} ,

$$B(u, f) \leq k_g \|u \circ f\| (\|f\| + \|\Delta f\|).$$

The proof is roughly as follows: Let p_i be a point of A_i where the maximum is achieved. By the inradius bound [Man3], there exists $\mu > 0$ so that the disc $D(p_j, \frac{\mu}{\lambda}) \subset A_i$. One can then express φ_λ in $D(p_j, \frac{\mu}{\lambda})$ by the sum of a Green's integral and Poisson integral with respect to the Euclidean Dirichlet Green's function of a slightly smaller disc. In particular one may express $\varphi_\lambda(p_j)$ by such an integral. Apply Hölder's inequality one gets

$$m_j^6 \leq k_g \lambda^2 \int_{D(p_j, r)} \varphi_\lambda^6 dV, \quad (r = \mu \lambda^{-\frac{1}{2}}).$$

Since the discs are disjoint one can sum in j and apply the Sogge L^6 bound to include the proof. Thus, the only fact one used about nodal domains was lower bound on the inradius.

This result bears a curious comparison to the results of [STZ] giving new constraints on (M, g) which are of maximal eigenfunction growth, i.e. possess eigenfunctions such that $m_A \geq C\lambda^{\frac{1}{2}}$ for some sequence of eigenfunctions φ_{λ_j} with $\lambda_j \rightarrow \infty$. The result (building on older results of Sogge and the author) states that such a sequence can exist only if (M, g) possesses a ‘pole’ p for which the set of geodesic loops \mathcal{L}_p based at p has positive measure in S_p^*M (with respect to the natural spherical volume measure) and such that the first return map has a recurrence property. In fact, the only known surfaces where the bounds are achieved are surfaces of revolution, and in this case the first return map is the identity. It is quite plausible that if (M, g) has maximal eigenfunction growth, then the first return map must be the identity map on a set of positive measure in \mathcal{L}_p .

Combined with the Polterovich-Sodin result above, we see that such ‘poles’ p , when they exist, can only occur in a uniformly bounded number of nodal domains of a surface. It would be interesting to know if there can exist only a finite number of such points at all if one additionally assumes that the set of smoothly closed geodesics has measure zero. For instance, in that case, there might be a unique pole in each of the finite number of possible nodal domains. This finitude problem would be useful in strengthening the condition on (M, g) of maximal eigenfunction growth.

5.2. Norms of restrictions. A problem of current interest is to consider L^p norms of restrictions of eigenfunctions to hypersurfaces or higher codimension submanifolds. For expository purposes we only consider geodesics on surfaces here. Following earlier work of A. Reznikov, Burq, Gérard and Tzvetkov [BGT] proved

THEOREM 5.2. [BGT] *Suppose that (M, g) is a compact surface, then there exists $\lambda_0(\epsilon), C > 0$ so that, for any geodesic segment γ of length L_γ and any eigenfunction φ_λ with $\lambda \geq \lambda_0$ we have*

$$(42) \quad \frac{1}{L_\gamma} \int_\gamma |\varphi_\lambda|^2 ds \leq C\lambda^{\frac{1}{2}} \|\varphi_\lambda\|^2$$

Their estimate is sharp for the round sphere S^2 because of the highest weight spherical harmonics. They also showed that for all geodesic segments γ of unit length,

$$\left(\frac{1}{L_\gamma} \int_\gamma |\varphi_\lambda|^4 ds \right)^{1/4} \leq C\lambda^{\frac{1}{4}} \|e_\lambda\|_{L^2(M)},$$

The estimate is only known to be achieved when the geodesic is elliptic, and quite likely it can be improved if the geodesic is hyperbolic. A result in this direction is:

THEOREM 5.3. [SoZ2] *Suppose that (M, g) is a compact surface of non-positive curvature. Then for all ϵ , there exists $\lambda_0(\epsilon), C > 0$ so that, for any geodesic segment γ of length L_γ and any eigenfunction φ_λ with $\lambda \geq \lambda_0(\epsilon)$, we have*

$$(43) \quad \frac{1}{L_\gamma} \int_\gamma |\varphi_\lambda|^2 ds \leq C\epsilon\lambda^{\frac{1}{2}} \|\varphi_\lambda\|^2$$

A related result on L^4 norms is,

THEOREM 5.4. [SoZ3] *Let (M, g) be a surface and assume that the set*

$$(44) \quad \mathcal{P} = \{(x, \xi) \in S^*M : g^t(x, \xi) = (x, \xi), \text{ some } t > 0\}$$

*of periodic points has Liouville measure zero in S^*M . Then there is a subsequence of eigenvalues λ_{j_k} of density one so that*

$$(45) \quad \|e_{\lambda_{j_k}}\|_{L^4(M)} = o(\lambda_{j_k}^{1/8}).$$

The results are based in part on a relatively new Kakeya-Nikodym maximal function estimate of Bourgain [Bourg], as improved by Sogge [Sog2]. We believe that it can be improved the following phase space Kakeya-Nikodym theorem. Let $T_\delta(\gamma)$ be the tube of radius δ around a geodesic arc in M , and let $\chi_{\delta, \gamma}$ be a smooth cutoff to a phase space tube of its lift to S^*M . Then for all ϵ , there exists $\delta(\epsilon)$ such that

$$\limsup_{\lambda \rightarrow \infty} \frac{1}{N(\lambda)} \sum_{\lambda_j \leq \lambda} \sup_{\gamma \in \Pi} \int_{T_{\delta(\epsilon)}(\gamma)} |\varphi_\lambda|^2 ds < \epsilon.$$

We expect the sup occurs when γ is the orbit of (x, ξ) . But then it is easy to estimate the right side and one should be able to get a quantitative improvement of Theorem 5.4.

5.3. Quantum ergodic restriction (QER) theorems. In this section we briefly review a recent series of results [**TZ2**, **TZ3**, **DZ**, **CTZ**] on quantum ergodic restriction theorems. They are used in section §10 to determine the limit distribution of intersections of nodal lines and geodesics on real analytic surfaces (in the complex domain).

Let $H \subset M$ be a hypersurface and consider the Cauchy data $(\varphi_j|_H, \lambda_j^{-1} \partial_\nu \varphi_j|_H)$ of eigenfunctions along H ; here ∂_ν is the normal derivative. We refer to $\varphi_j|_H$ as the Dirichlet data and to $\lambda_j^{-1} \partial_\nu \varphi_j|_H$ as the Neumann data. A QER (quantum ergodic restriction) theorem seeks to find limits of matrix elements of this data along H with respect to pseudo-differential operators $O_{ph}(a)$ on H . The main idea is that $S_H^* M$, the set of unit covectors with footpoints on H , is a cross-section to the geodesic flow and the first return map of the geodesic flow for $S_H^* M$ is ergodic. The Cauchy data should be the quantum analogue of such a cross section and therefore should be quantum ergodic on H .

For applications to nodal sets and other problems, it is important to know if the Dirichlet data alone satisfies a QER theorem. The answer is obviously ‘no’ in general. For instance if (M, g) has an isometric involution and with a hypersurface H of fixed points, then any eigenfunction which is odd with respect to the involution vanishes on H . But in [**TZ2**, **TZ3**] a sufficient condition is given for quantum ergodic restriction, which rules out this and more general situations. The symmetry condition is that geodesics emanating from the ‘left side’ of H have a different return map from geodesics on the ‘right side’ when the initial conditions are reflections of each other through TH . To take the simplest example of the circle, the restriction of $\sin kx$ to a point is never quantum ergodic but the full Cauchy data $(\cos kx, \sin kx)$ of course satisfies $\cos^2 kx + \sin^2 kx = 1$. In [**CTZ**] it is proved that Cauchy data always satisfies QER for any hypersurface. This has implications for (at least complex) zeros of even or odd eigenfunctions along an axis of symmetry, e.g. for the case of Maass forms for the modular domain $SL(2, \mathbb{Z})/\mathcal{H}^2$ (see §10).

To state the QER theorem, we introduce some notation. We put

$$(46) \quad T_H^* M = \{(q, \xi) \in T_q^* M, q \in H\}, \quad T^* H = \{(q, \eta) \in T_q^* H, q \in H\}.$$

We further denote by $\pi_H : T_H^* M \rightarrow T^* H$ the restriction map,

$$(47) \quad \pi_H(x, \xi) = \xi|_{TH}.$$

For any orientable (embedded) hypersurface $H \subset M$, there exists two unit normal co-vector fields ν_\pm to H which span half ray bundles $N_\pm = \mathbb{R}_+ \nu_\pm \subset N^* H$. Infinitesimally, they define two ‘sides’ of H , indeed they are the two components of $T_H^* M \setminus T^* H$. We use Fermi normal coordinates (s, y_n)

along H with $s \in H$ and with $x = \exp_x y_n \nu$ and let σ, η_n denote the dual symplectic coordinates. For $(s, \sigma) \in B^*H$ (the co-ball bundle), there exist two unit covectors $\xi_{\pm}(s, \sigma) \in S_H^*M$ such that $|\xi_{\pm}(s, \sigma)| = 1$ and $\xi|_{T_s H} = \sigma$. In the above orthogonal decomposition, they are given by

$$(48) \quad \xi_{\pm}(s, \sigma) = \sigma \pm \sqrt{1 - |\sigma|^2} \nu_+(s).$$

We define the reflection involution through T^*H by

$$(49) \quad r_H : T_H^*M \rightarrow T_H^*M, \quad r_H(s, \mu \xi_{\pm}(s, \sigma)) = (s, \mu \xi_{\mp}(s, \sigma)), \quad \mu \in \mathbb{R}_+.$$

Its fixed point set is T^*H .

We denote by G^t the homogeneous geodesic flow of (M, g) , i.e. Hamiltonian flow on $T^*M - 0$ generated by $|\xi|_g$. We define the *first return time* $T(s, \xi)$ on S_H^*M by,

$$(50) \quad T(s, \xi) = \inf\{t > 0 : G^t(s, \xi) \in S_H^*M, \quad (s, \xi) \in S_H^*M\}.$$

By definition $T(s, \xi) = +\infty$ if the trajectory through (s, ξ) fails to return to H . Inductively, we define the j th return time $T^{(j)}(s, \xi)$ to S_H^*M and the j th return map Φ^j when the return times are finite.

We define the first return map on the same domain by

$$(51) \quad \Phi : S_H^*M \rightarrow S_H^*M, \quad \Phi(s, \xi) = G^{T(s, \xi)}(s, \xi)$$

When G^t is ergodic, Φ is defined almost everywhere and is also ergodic with respect to Liouville measure $\mu_{L,H}$ on S_H^*M .

Definition: We say that H has a positive measure of microlocal reflection symmetry if

$$\mu_{L,H} \left(\bigcup_{j \neq 0}^{\infty} \{(s, \xi) \in S_H^*M : r_H G^{T^{(j)}(s, \xi)}(s, \xi) = G^{T^{(j)}(s, \xi)} r_H(s, \xi)\} \right) > 0.$$

Otherwise we say that H is asymmetric with respect to the geodesic flow.

The QER theorem we state below holds for both poly-homogeneous (Kohn-Nirenberg) pseudo-differential operators as in [HoI-IV] and also for semi-classical pseudo-differential operators on H [Zw] with essentially the same proof. To avoid confusion between pseudodifferential operators on the ambient manifold M and those on H , we denote the latter by $Oph(a)$ where $a \in S_{cl}^0(T^*H)$. By Kohn-Nirenberg pseudo-differential operators we mean operators with classical poly-homogeneous symbols $a(s, \sigma) \in C^\infty(T^*H)$,

$$a(s, \sigma) \sim \sum_{k=0}^{\infty} a_{-k}(s, \sigma), \quad (a_{-k} \text{ positive homogeneous of order } -k)$$

as $|\sigma| \rightarrow \infty$ on T^*H as in [HoI-IV]. By semi-classical pseudo-differential operators we mean h -quantizations of semi-classical symbols $a \in S^{0,0}(T^*H \times$

$(0, h_0]$) of the form

$$a_h(s, \sigma) \sim \sum_{k=0}^{\infty} h^k a_{-k}(s, \sigma), \quad (a_{-k} \in S_{1,0}^0(T^*H))$$

as in **[Zw, HZ, TZ]**.

We further introduce the zeroth order homogeneous function

$$(52) \quad \gamma(s, y_n, \sigma, \eta_n) = \frac{|\eta_n|}{\sqrt{|\sigma|^2 + |\eta_n|^2}} = (1 - \frac{|\sigma|^2}{r^2})^{\frac{1}{2}}, \quad (r^2 = |\sigma|^2 + |\eta_n|^2)$$

on T_H^*M and also denote by

$$(53) \quad \gamma_{B^*H} = (1 - |\sigma|^2)^{\frac{1}{2}}$$

its restriction to $S_H^*M = \{r = 1\}$.

For homogeneous pseudo-differential operators, the QER theorem is as follows:

THEOREM 5.5. [TZ, TZ2, DZ] *Let (M, g) be a compact manifold with ergodic geodesic flow, and let $H \subset M$ be a hypersurface. Let $\varphi_{\lambda_j}; j = 1, 2, \dots$ denote the L^2 -normalized eigenfunctions of Δ_g . If H has a zero measure of microlocal symmetry, then there exists a density-one subset S of \mathbb{N} such that for $\lambda_0 > 0$ and $a(s, \sigma) \in S_{cl}^0(T^*H)$*

$$\lim_{\lambda_j \rightarrow \infty; j \in S} \langle Oph(a)\gamma_H \varphi_{\lambda_j}, \gamma_H \varphi_{\lambda_j} \rangle_{L^2(H)} = \omega(a),$$

where

$$\omega(a) = \frac{2}{vol(S^*M)} \int_{B^*H} a_0(s, \sigma) \gamma_{B^*H}^{-1}(s, \sigma) ds d\sigma.$$

Alternatively, one can write $\omega(a) = \frac{1}{vol(S^*M)} \int_{S_H^*M} a_0(s, \pi_H(\xi)) d\mu_{L,H}(\xi)$. Note that $a_0(s, \sigma)$ is bounded but is not defined for $\sigma = 0$, hence $a_0(s, \pi_H(\xi))$ is not defined for $\xi \in N^*H$ if $a_0(s, \sigma)$ is homogeneous of order zero on T^*H . The analogous result for semi-classical pseudo-differential operators is:

THEOREM 5.6. [TZ, TZ2, DZ] *Let (M, g) be a compact manifold with ergodic geodesic flow, and let $H \subset M$ be a hypersurface. If H has a zero measure of microlocal symmetry, then there exists a density-one subset S of \mathbb{N} such that for $a \in S^{0,0}(T^*H \times [0, h_0))$,*

$$\lim_{h_j \rightarrow 0^+; j \in S} \langle Oph_j(a)\gamma_H \varphi_{h_j}, \gamma_H \varphi_{h_j} \rangle_{L^2(H)} = \omega(a),$$

where

$$\omega(a) = \frac{2}{vol(S^*M)} \int_{B^*H} a_0(s, \sigma) \gamma_{B^*H}^{-1}(s, \sigma) ds d\sigma.$$

Examples of asymmetric curves on surfaces in the case where (M, g) is a finite area hyperbolic surface are the following:

- H is a geodesic circle;

- H is a closed horocycle of radius $r < \text{inj}(M, g)$, the injectivity radius.
- H is a generic closed geodesic or an arc of a generic non-closed geodesic.

6. Critical points

In this section, we briefly discuss some analogues of (16) and (21) for critical points on surfaces. To be sure, it is not hard to generate many identities; the main problem is to derive information from them.

We denote the gradient of a function φ by $\nabla\varphi$ and its Hessian by $\nabla^2\varphi := \nabla d\varphi$, where ∇ is the Riemannian connection. We also denote the area form by dA and the scalar curvature by K . The results are based on unpublished work in progress of the author. It is often said that measuring critical point sets and values is much more difficult than measuring nodal sets; the identities reflect this difficulty in that the identities become signed:

PROPOSITION 6.1. *Suppose that (M, g) is a Riemannian surface, and that φ is a Morse eigenfunction with $(\Delta + \lambda^2)\varphi = 0$. Let $V \in C^2(M)$. Then*

(54)

$$\begin{aligned} 2\pi \sum_{p:d\varphi(p)=0} \text{sign}(\det \nabla^2\varphi(p)) V(p) &= 2\lambda^2 \int_M \frac{\varphi}{|\nabla\varphi|} \frac{\nabla V \cdot \nabla\varphi}{|\nabla\varphi|} dA + 2 \int_M KV dA \\ &\quad - \int_M (\Delta V) \log |\nabla\varphi|^2 dA. \end{aligned}$$

Here, $\text{sign}(\det \nabla^2\varphi(p)) = 1$ if p is a local maximum or minimum and $= -1$ if p is a saddle point. When $V \equiv 1$, the identity reduces to the Gauss-Bonnet theorem $\int K dA = 2\pi\chi(M)$ and the Hopf index formula $\chi(M) = \sum_{x:\nabla\varphi(x)=0} \text{sign}(\det \nabla^2\varphi(p))$. As this indicates, the main problem with applying the identity to counting critical points is that the left side is an alternating sum over critical points rather than a positive sum. In [Dong] a related identity using $|\nabla\varphi|^2 + \lambda^2\varphi^2$ produced a sum of constant sign over the singular points of φ , but singular points are always saddle points of index -1 and hence of constant sign. Note that under the Morse assumption, $\log |\nabla\varphi|, |\nabla\varphi|^{-1} \in L^1(M, dA)$, so that the right side is a well defined measure integrated against V .

We now make some interesting choices of V . As mentioned above, (weighted) counting of critical values should be simpler than weighted counting of critical points. Hence we put $V = f(\varphi)$ for smooth f . This choice does give cancellation of the ‘bad factor’ $|\nabla\varphi|^{-1}$ and (using that $\Delta f(\varphi) = f''(\varphi)|\nabla\varphi|^2 - f'(\varphi)\lambda^2\varphi$) we get

COROLLARY 1. *With the assumptions of Proposition 6.1, if $f \in C^2(\mathbb{R})$, then*

$$(55) \quad 2\pi \sum_{p:d\varphi(p)=0} \text{sign}(\det \nabla^2 \varphi(p)) f(\varphi(p)) = 2\lambda^2 \int_M \varphi f'(\varphi) dA + 2 \int_M Kf(\varphi) dA - \int_M (f''(\varphi) |\nabla \varphi|^2 - f'(\varphi) \lambda^2 \varphi) \log |\nabla \varphi|^2 dA.$$

Of course, this still has the defect that the left side is an oscillating sum, and the factor $f(\varphi)$ in the sum damps out the critical points in regions of exponential decay. To illustrate, if $f(x) = x$ we get

$$(56) \quad 2\pi \sum_{p:d\varphi(p)=0} \text{sign}(\det \nabla^2 \varphi(p)) \varphi(p) = 2 \int_M K\varphi dA + \lambda^2 \int_M \varphi \log |\nabla \varphi|^2 dA.$$

To highlight the sign issue, we break up the sum into the sub-sum over maxima/minima and the sub-sum over saddle points, denoting the set of local maxima (resp. minima) by \max (resp. \min) and the set of saddle points by Sad . Of course we have $\#(\max \cup \min) - \#Sad = \chi(M)$. Then (55) is equivalent to

$$(57) \quad 2\pi \sum_{p \in \max \cup \min} f(\varphi(p)) = 2\pi \sum_{p \in Sad} f(\varphi(p)) + 2\lambda^2 \int_M \varphi f'(\varphi) dA + 2 \int_M Kf(\varphi) dA - \int_M (f''(\varphi) |\nabla \varphi|^2 - f'(\varphi) \lambda^2 \varphi) \log |\nabla \varphi|^2 dA.$$

We write $\log r = \log_+ r - \log_- r$ where $\log_+ r = \max\{\log r, 0\}$. We note that on any compact Riemannian manifold, $\log_+ |\nabla \varphi|^2 = O(\log \lambda)$ uniformly in x as $\lambda \rightarrow \infty$ while $\log_- |\nabla \varphi|^2$ can be quite complicated to estimate. When $f = x^2$ we get,

$$(58) \quad 2\pi \sum_{p \in \max, \min} \varphi^2(p) = 2\pi \sum_{p \in Sad} \varphi(p)^2 + 4\lambda^2 + 2 \int_M (\lambda^2 \varphi^2 - |\nabla \varphi|^2) \log |\nabla \varphi|^2 dA + 2 \int_M K\varphi^2 dA.$$

Assuming φ is a Morse eigenfunction, this implies

$$(59) \quad \sum_{p \in \max, \min} \varphi^2(p) \leq \sum_{p \in Sad} \varphi(p)^2 + O(\lambda^2 \log \lambda).$$

To get rid of the signs in the sum, we could choose $V = W \det \nabla^2 \varphi$, where the determinant is defined by the metric. Since $(\text{sign} \det \nabla^2 \varphi) \det \nabla^2 \varphi =$

$|\det \nabla^2 \varphi|$ we obtain

$$(60) \quad 2\pi \sum_{p:d\varphi(p)=0} |\det \nabla^2 \varphi(p)| W(p) = 2\lambda^2 \int_M \frac{\varphi}{|\nabla \varphi|} \frac{\nabla(W \det \nabla^2 \varphi) \cdot \nabla \varphi}{|\nabla \varphi|} dA$$

$$(61) \quad + 2 \int_M KW \det \nabla^2 \varphi dA$$

$$- \int_M (\Delta W \det \nabla^2 \varphi) \log |\nabla \varphi|^2 dA.$$

But the first term appears to be difficult to estimate.

The optimist might conjecture the following Bézout bound for the number of critical values of eigenfunctions in the real analytic case:

CONJECTURE 6.2. *If (M, g) is real analytic then the number $\#CV(\varphi_\lambda)$ of critical values of φ_λ satisfies $\#CV(\varphi_\lambda) \leq C_g \lambda^n$ (where $n = \dim M$).*

Note that the critical point set could have codimension 1 (e.g. rotationally invariant eigenfunctions on a surface of revolution), so that in general we cannot count critical points. The number of critical values is generically the same as the number of connected components, although there could exist high multiplicities in the number of components of a give critical level.

The conjecture is motivated by Bézout's theorem for the number of intersection points of n real algebraic varieties of degree λ in dimension n . But it is difficult to control intersections in the real analytic case and it is not very clear at present how plausible the conjecture is.

7. Analytic continuation of eigenfunctions for real analytic (M, g)

We now take up the theme mentioned in the introduction of analytically continuing eigenfunctions on real analytic (M, g) to the complex domain. In the next sections we apply the analytic continuation to the study of nodal of eigenfunctions in the real analytic case. For background we refer to [LS1, LS2, GS1, GS2, GLS, Z8].

A real analytic manifold M always possesses a unique complexification $M_{\mathbb{C}}$ generalizing the complexification of \mathbb{R}^m as \mathbb{C}^m . The complexification is an open complex manifold in which M embeds $\iota : M \rightarrow M_{\mathbb{C}}$ as a totally real submanifold (Bruhat-Whitney). As examples, we have:

- $M = \mathbb{R}^m / \mathbb{Z}^m$ is $M_{\mathbb{C}} = \mathbb{C}^m / \mathbb{Z}^m$.
- The unit sphere S^n defined by $x_1^2 + \cdots + x_{n+1}^2 = 1$ in \mathbb{R}^{n+1} is complexified as the complex quadric $S_{\mathbb{C}}^2 = \{(z_1, \dots, z_n) \in \mathbb{C}^{n+1} : z_1^2 + \cdots + z_{n+1}^2 = 1\}$.
- The hyperboloid model of hyperbolic space is the hypersurface in \mathbb{R}^{n+1} defined by

$$\mathbb{H}^n = \{x_1^2 + \cdots + x_n^2 - x_{n+1}^2 = -1, \quad x_n > 0\}.$$

Then,

$$H_{\mathbb{C}}^n = \{(z_1, \dots, z_{n+1}) \in \mathbb{C}^{n+1} : z_1^2 + \cdots + z_n^2 - z_{n+1}^2 = -1\}.$$

- Any real algebraic subvariety of \mathbb{R}^m has a similar complexification.
- Any Lie group G (or symmetric space) admits a complexification $G_{\mathbb{C}}$.

The Riemannian metric determines a special kind of distance function on $M_{\mathbb{C}}$ known as a Grauert tube function. It is the plurisubharmonic function $\sqrt{\rho} = \sqrt{\rho}_g$ on $M_{\mathbb{C}}$ defined as the unique solution of the Monge-Ampère equation

$$(\partial\bar{\partial}\sqrt{\rho})^m = \delta_{M_{\mathbb{R}}, dV_g}, \quad \iota^*(i\partial\bar{\partial}\rho) = g.$$

Here, $\delta_{M_{\mathbb{R}}, dV_g}$ is the delta-function on the real M with respect to the volume form dV_g , i.e. $f \rightarrow \int_M f dV_g$. In fact, it is observed in [GS1, GLS] that the Grauert tube function is obtained from the distance function by setting $\sqrt{\rho}(\zeta) = i\sqrt{r^2(\zeta, \bar{\zeta})}$ where $r^2(x, y)$ is the squared distance function in a neighborhood of the diagonal in $M \times M$.

One defines the Grauert tubes $M_{\tau} = \{\zeta \in M_{\mathbb{C}} : \sqrt{\rho}(\zeta) \leq \tau\}$. There exists a maximal τ_0 for which $\sqrt{\rho}$ is well defined, known as the Grauert tube radius. For $\tau \leq \tau_0$, M_{τ} is a strictly pseudoconvex domain in $M_{\mathbb{C}}$.

The complexified exponential map $(x, \xi) \rightarrow \exp_x i\xi$ defines a diffeomorphism from $B_{\tau}^* M$ to M_{τ} and pulls back $\sqrt{\rho}$ to $|\xi|_g$. The one-complex dimensional null foliation of $\partial\bar{\partial}\sqrt{\rho}$, known as the ‘Monge-Ampère’ or Riemann foliation, are the complex curves $t + i\tau \rightarrow \tau\dot{\gamma}(t)$, where γ is a geodesic, where $\tau > 0$ and where $\tau\dot{\gamma}(t)$ denotes multiplication of the tangent vector to γ by τ . We refer to [LS1, GLS, Z8] for further discussion.

7.1. Poisson operator and analytic Continuation of eigenfunctions. The half-wave group of (M, g) is the unitary group $U(t) = e^{it\sqrt{\Delta}}$ generated by the square root of the positive Laplacian. Its Schwartz kernel is a distribution on $\mathbb{R} \times M \times M$ with the eigenfunction expansion

$$(62) \quad U(t, x, y) = \sum_{j=0}^{\infty} e^{it\lambda_j} \varphi_j(x) \varphi_j(y).$$

By the Poisson operator we mean the analytic continuation of $U(t)$ to positive imaginary time,

$$(63) \quad e^{-\tau\sqrt{\Delta}} = U(i\tau).$$

The eigenfunction expansion then converges absolutely to a real analytic function on $\mathbb{R}_+ \times M \times M$.

Let $A(\tau)$ denote the operator of analytic continuation of a function on M to the Grauert tube M_{τ} . Since

$$(64) \quad U_{\mathbb{C}}(i\tau)\varphi_{\lambda} = e^{-\tau\lambda} \varphi_{\lambda}^{\mathbb{C}},$$

it is simple to see that

$$(65) \quad A(\tau) = U_{\mathbb{C}}(i\tau)e^{\tau\sqrt{\Delta}}$$

where $U_{\mathbb{C}}(i\tau, \zeta, y)$ is the analytic continuation of the Poisson kernel in x to M_τ . In terms of the eigenfunction expansion, one has

$$(66) \quad U_{\mathbb{C}}(i\tau, \zeta, y) = \sum_{j=0}^{\infty} e^{-\tau\lambda_j} \varphi_j^{\mathbb{C}}(\zeta) \varphi_j(y), \quad (\zeta, y) \in M_\epsilon \times M.$$

This is a very useful observation because $U_{\mathbb{C}}(i\tau)$ (66) is a Fourier integral operator with complex phase and can be related to the geodesic flow. The analytic continuability of the Poisson operator to M_τ implies that every eigenfunction analytically continues to the same Grauert tube.

7.2. Analytic continuation of the Poisson wave group. The analytic continuation of the Poisson-wave kernel to M_τ in the x variable is discussed in detail in [Z8] and ultimately derives from the analysis by Hadamard of his parametrix construction. We only briefly discuss it here and refer to [Z8] for further details. In the case of Euclidean \mathbb{R}^n and its wave kernel $U(t, x, y) = \int_{\mathbb{R}^n} e^{it|\xi|} e^{i\langle \xi, x-y \rangle} d\xi$ which analytically continues to $t + i\tau, \zeta = x + ip \in \mathbb{C}_+ \times \mathbb{C}^n$ as the integral

$$U_{\mathbb{C}}(t + i\tau, x + ip, y) = \int_{\mathbb{R}^n} e^{i(t+i\tau)|\xi|} e^{i\langle \xi, x+ip-y \rangle} d\xi.$$

The integral clearly converges absolutely for $|p| < \tau$.

Exact formulae of this kind exist for S^m and \mathbf{H}^m . For a general real analytic Riemannian manifold, there exists an oscillatory integral expression for the wave kernel of the form,

$$(67) \quad U(t, x, y) = \int_{T_y^* M} e^{it|\xi|_{g_y}} e^{i\langle \xi, \exp_y^{-1}(x) \rangle} A(t, x, y, \xi) d\xi$$

where $A(t, x, y, \xi)$ is a polyhomogeneous amplitude of order 0. The holomorphic extension of (67) to the Grauert tube $|\zeta| < \tau$ in x at time $t = i\tau$ then has the form

$$(68) \quad U_{\mathbb{C}}(i\tau, \zeta, y) = \int_{T_y^*} e^{-\tau|\xi|_{g_y}} e^{i\langle \xi, \exp_y^{-1}(\zeta) \rangle} A(t, \zeta, y, \xi) d\xi \quad (\zeta = x + ip).$$

7.3. Analytic continuation of eigenfunctions. A function $f \in C^\infty(M)$ has a holomorphic extension to the closed tube $\sqrt{\rho}(\zeta) \leq \tau$ if and only if $f \in \text{Dom}(e^{\tau\sqrt{\Delta}})$, where $e^{\tau\sqrt{\Delta}}$ is the backwards ‘heat operator’ generated by $\sqrt{\Delta}$ (rather than Δ). That is, $f = \sum_{n=0}^{\infty} a_n \varphi_{\lambda_n}$ admits an analytic continuation to the open Grauert tube M_τ if and only if f is in the domain of $e^{\tau\sqrt{\Delta}}$, i.e. if $\sum_n |a_n|^2 e^{2\tau\lambda_n} < \infty$. Indeed, the analytic continuation is $U_{\mathbb{C}}(i\tau) e^{\tau\sqrt{\Delta}} f$. The subtlety is in the nature of the restriction to the boundary of the maximal Grauert tube.

This result generalizes one of the classical Paley-Wiener theorems to real analytic Riemannian manifolds [Bou, GS2]. In the simplest case of $M = S^1$, $f \sim \sum_{n \in \mathbb{Z}} a_n e^{in\theta} \in C^\omega(S^1)$ is the restriction of a holomorphic function $F \sim \sum_{n \in \mathbb{Z}} a_n z^n$ on the annulus $S_\tau^1 = \{|\log|z|| < \tau\}$ and with

$F \in L^2(\partial S_\tau^1)$ if and only if $\sum_n |\hat{f}(n)|^2 e^{2|n|\tau} < \infty$. The case of \mathbb{R}^m is more complicated since it is non-compact. We are mainly concerned with compact manifolds and so the complications are not very relevant here. But we recall that one of the classical Paley-Wiener theorems states that a real analytic function f on \mathbb{R}^n is the restriction of a holomorphic function on the closed tube $|\text{Im } \zeta| \leq \tau$ which satisfies $\int_{\mathbb{R}^m} |F(x + i\xi)|^2 dx \leq C$ for $\xi \leq \tau$ if and only if $\hat{f}e^{\tau|\text{Im } \zeta|} \in L^2(\mathbb{R}^n)$.

Let us consider examples of holomorphic continuations of eigenfunctions:

- On the flat torus $\mathbb{R}^m/\mathbb{Z}^m$, the real eigenfunctions are $\cos\langle k, x \rangle$, $\sin\langle k, x \rangle$ with $k \in 2\pi\mathbb{Z}^m$. The complexified torus is $\mathbb{C}^m/\mathbb{Z}^m$ and the complexified eigenfunctions are $\cos\langle k, \zeta \rangle$, $\sin\langle k, \zeta \rangle$ with $\zeta = x + i\xi$.
- On the unit sphere S^m , eigenfunctions are restrictions of homogeneous harmonic functions on \mathbb{R}^{m+1} . The latter extend holomorphically to holomorphic harmonic polynomials on \mathbb{C}^{m+1} and restrict to holomorphic function on S_C^m .
- On \mathbf{H}^m , one may use the hyperbolic plane waves $e^{(i\lambda+1)\langle z, b \rangle}$, where $\langle z, b \rangle$ is the (signed) hyperbolic distance of the horocycle passing through z and b to 0. They may be holomorphically extended to the maximal tube of radius $\pi/2$.
- On compact hyperbolic quotients \mathbf{H}^m/Γ , eigenfunctions can be then represented by Helgason's generalized Poisson integral formula [H],

$$\varphi_\lambda(z) = \int_B e^{(i\lambda+1)\langle z, b \rangle} dT_\lambda(b).$$

Here, $z \in D$ (the unit disc), $B = \partial D$, and $dT_\lambda \in \mathcal{D}'(B)$ is the boundary value of φ_λ , taken in a weak sense along circles centered at the origin 0. To analytically continue φ_λ it suffices to analytically continue $\langle z, b \rangle$. Writing the latter as $\langle \zeta, b \rangle$, we have:

$$(69) \quad \varphi_\lambda^C(\zeta) = \int_B e^{(i\lambda+1)\langle \zeta, b \rangle} dT_\lambda(b).$$

7.4. Complexified spectral projections. The next step is to holomorphically extend the spectral projectors $d\Pi_{[0,\lambda]}(x, y) = \sum_j \delta(\lambda - \lambda_j) \varphi_j(x) \varphi_j(y)$ of $\sqrt{\Delta}$. The complexified diagonal spectral projections measure is defined by

$$(70) \quad d_\lambda \Pi_{[0,\lambda]}^C(\zeta, \bar{\zeta}) = \sum_j \delta(\lambda - \lambda_j) |\varphi_j^C(\zeta)|^2.$$

Henceforth, we generally omit the superscript and write the kernel as $\Pi_{[0,\lambda]}^C(\zeta, \bar{\zeta})$. This kernel is not a tempered distribution due to the exponential growth of $|\varphi_j^C(\zeta)|^2$. Since many asymptotic techniques assume spectral functions are of polynomial growth, we simultaneously consider the damped

spectral projections measure

$$(71) \quad d_\lambda P_{[0,\lambda]}^\tau(\zeta, \bar{\zeta}) = \sum_j \delta(\lambda - \lambda_j) e^{-2\tau\lambda_j} |\varphi_j^C(\zeta)|^2,$$

which is a temperate distribution as long as $\sqrt{\rho}(\zeta) \leq \tau$. When we set $\tau = \sqrt{\rho}(\zeta)$ we omit the τ and put

$$(72) \quad d_\lambda P_{[0,\lambda]}(\zeta, \bar{\zeta}) = \sum_j \delta(\lambda - \lambda_j) e^{-2\sqrt{\rho}(\zeta)\lambda_j} |\varphi_j^C(\zeta)|^2.$$

The integral of the spectral measure over an interval I gives

$$\Pi_I(x, y) = \sum_{j: \lambda_j \in I} \varphi_j(x) \varphi_j(y).$$

Its complexification gives the kernel (121) along the diagonal,

$$(73) \quad \Pi_I(\zeta, \bar{\zeta}) = \sum_{j: \lambda_j \in I} |\varphi_j^C(\zeta)|^2,$$

and the integral of (71) gives its temperate version

$$(74) \quad P_I^\tau(\zeta, \bar{\zeta}) = \sum_{j: \lambda_j \in I} e^{-2\tau\lambda_j} |\varphi_j^C(\zeta)|^2,$$

or in the crucial case of $\tau = \sqrt{\rho}(\zeta)$,

$$(75) \quad P_I(\zeta, \bar{\zeta}) = \sum_{j: \lambda_j \in I} e^{-2\sqrt{\rho}(\zeta)\lambda_j} |\varphi_j^C(\zeta)|^2,$$

7.5. Poisson operator as a complex Fourier integral operator.

The damped spectral projection measure $d_\lambda P_{[0,\lambda]}^\tau(\zeta, \bar{\zeta})$ (71) is dual under the real Fourier transform in the t variable to the restriction

$$(76) \quad U(t + 2i\tau, \zeta, \bar{\zeta}) = \sum_j e^{(-2\tau+it)\lambda_j} |\varphi_j^C(\zeta)|^2$$

to the anti-diagonal of the mixed Poisson-wave group. The adjoint of the Poisson kernel $U(i\tau, x, y)$ also admits an anti-holomorphic extension in the y variable. The sum (76) are the diagonal values of the complexified wave kernel

$$(77) \quad \begin{aligned} U(t + 2i\tau, \zeta, \bar{\zeta}') &= \int_M U(t + i\tau, \zeta, y) E(i\tau, y, \bar{\zeta}') dV_g(y) \\ &= \sum_j e^{(-2\tau+it)\lambda_j} \varphi_j^C(\zeta) \overline{\varphi_j^C(\zeta')} \end{aligned}$$

We obtain (77) by orthogonality of the real eigenfunctions on M .

Since $U(t + 2i\tau, \zeta, y)$ takes its values in the CR holomorphic functions on ∂M_τ , we consider the Sobolev spaces $\mathcal{O}^{s+\frac{n-1}{4}}(\partial M_\tau)$ of CR holomorphic functions on the boundaries of the strictly pseudoconvex domains M_ϵ , i.e.

$$\mathcal{O}^{s+\frac{m-1}{4}}(\partial M_\tau) = W^{s+\frac{m-1}{4}}(\partial M_\tau) \cap \mathcal{O}(\partial M_\tau),$$

where W_s is the s th Sobolev space and where $\mathcal{O}(\partial M_\epsilon)$ is the space of boundary values of holomorphic functions. The inner product on $\mathcal{O}^0(\partial M_\tau)$ is with respect to the Liouville measure

$$(78) \quad d\mu_\tau = (i\partial\bar{\partial}\sqrt{\rho})^{m-1} \wedge d^c\sqrt{\rho}.$$

We then regard $U(t+i\tau, \zeta, y)$ as the kernel of an operator from $L^2(M) \rightarrow \mathcal{O}^0(\partial M_\tau)$. It equals its composition $\Pi_\tau \circ U(t+i\tau)$ with the Szegöprojector

$$\Pi_\tau : L^2(\partial M_\tau) \rightarrow \mathcal{O}^0(\partial M_\tau)$$

for the tube M_τ , i.e. the orthogonal projection onto boundary values of holomorphic functions in the tube.

This is a useful expression for the complexified wave kernel, because $\tilde{\Pi}_\tau$ is a complex Fourier integral operator with a small wave front relation. More precisely, the real points of its canonical relation form the graph Δ_Σ of the identity map on the symplectic one $\Sigma_\tau \subset T^*\partial M_\tau$ spanned by the real one-form $d^c\rho$, i.e.

$$(79) \quad \Sigma_\tau = \{(\zeta; rd^c\rho(\zeta)), \quad \zeta \in \partial M_\tau, \quad r > 0\} \subset T^*(\partial M_\tau).$$

We note that for each τ , there exists a symplectic equivalence $\Sigma_\tau \simeq T^*M$ by the map $(\zeta, rd^c\rho(\zeta)) \rightarrow (E_{\mathbb{C}}^{-1}(\zeta), r\alpha)$, where $\alpha = \xi \cdot dx$ is the action form (cf. [GS2]).

The following result was first stated by Boutet de Monvel (for more details, see also [GS2, Z8]).

THEOREM 7.1. [Bou, GS2] $\Pi_\epsilon \circ U(i\epsilon) : L^2(M) \rightarrow \mathcal{O}(\partial M_\epsilon)$ is a complex Fourier integral operator of order $-\frac{m-1}{4}$ associated to the canonical relation

$$\Gamma = \{(y, \eta, \iota_\epsilon(y, \eta))\} \subset T^*M \times \Sigma_\epsilon.$$

Moreover, for any s ,

$$\Pi_\epsilon \circ U(i\epsilon) : W^s(M) \rightarrow \mathcal{O}^{s+\frac{m-1}{4}}(\partial M_\epsilon)$$

is a continuous isomorphism.

In [Z8] we give the following sharpening of the sup norm estimates of [Bou, GLS]:

PROPOSITION 7.2. Suppose (M, g) is real analytic. Then

$$\sup_{\zeta \in M_\tau} |\varphi_\lambda^\mathbb{C}(\zeta)| \leq C\lambda^{\frac{m+1}{2}} e^{\tau\lambda}, \quad \sup_{\zeta \in M_\tau} \left| \frac{\partial \varphi_\lambda^\mathbb{C}(\zeta)}{\partial \zeta_j} \right| \leq C\lambda^{\frac{m+3}{2}} e^{\tau\lambda}$$

The proof follows easily from the fact that the complexified Poisson kernel is a complex Fourier integral operator of finite order. The estimates can be improved further.

7.6. Maximal plurisubharmonic functions and growth of $\varphi_\lambda^{\mathbb{C}}$.

In [Z8], we discussed analogues in the setting of Grauert tubes for the basic notions of pluripotential theory on domains in \mathbb{C}^n . Of relevance here is that the Grauert tube function $\sqrt{\rho}$ is the analogue of the pluri-complex Green's function. We recall that the maximal PSH function (or pluri-complex Green's function) relative to a subset $E \subset \Omega$ is defined by

$$V_E(\zeta) = \sup\{u(z) : u \in PSH(\Omega), u|_E \leq 0, u|_{\partial\Omega} \leq 1\}.$$

On a real analytic Riemannian manifold, the natural analogue of \mathcal{P}^N is the space

$$\mathcal{H}^\lambda = \{p = \sum_{j:\lambda_j \leq \lambda} a_j \varphi_{\lambda_j}, \quad a_1, \dots, a_{N(\lambda)} \in \mathbb{R}\}$$

spanned by eigenfunctions with frequencies $\leq \lambda$. Rather than using the sup norm, it is convenient to work with L^2 based norms than sup norms, and so we define

$$\mathcal{H}_M^\lambda = \{p = \sum_{j:\lambda_j \leq \lambda} a_j \varphi_{\lambda_j}, \quad \|p\|_{L^2(M)}^2 = \sum_{j=1}^{N(\lambda)} |a_j|^2 = 1\}.$$

We define the λ -Siciak extremal function by

$$\Phi_M^\lambda(z) = \sup\{|\psi(z)|^{1/\lambda} : \psi \in \mathcal{H}_\lambda; \|\psi\|_M \leq 1\},$$

and the extremal function by

$$\Phi_M(z) = \sup_\lambda \Phi_M^\lambda(z).$$

The extremal PSH function is defined by

$$V_g(\zeta; \tau) = \sup\{u(z) : u \in PSH(M_\tau), u|_M \leq 0, u|_{\partial M_\tau} \leq \tau\}.$$

In [Z8] we proved that $V_g = \sqrt{\rho}$ and that

$$(80) \quad \Phi_M = V_g.$$

The proof is based on the properties of (73). By using a Bernstein-Walsh inequality

$$\frac{1}{N(\lambda)} \leq \frac{\Pi_{[0,\lambda]}(\zeta, \bar{\zeta})}{\Phi_M^\lambda(\zeta)^2} \leq CN(\lambda) e^{\epsilon N(\lambda)},$$

it is not hard to show that

$$(81) \quad \Phi_M(z) = \lim_{\lambda \rightarrow \infty} \frac{1}{\lambda} \log \Pi_{[0,\lambda]}(\zeta, \bar{\zeta}).$$

To evaluate the logarithm, one can show that the kernel is essentially $e^{\lambda\sqrt{\rho}}$ times the temperate projection defined by the Poisson operator,

$$(82) \quad P_{[0,\lambda]}(\zeta, \bar{\zeta}) = \sum_{j:\lambda_j \in [0,\lambda]} e^{-2\sqrt{\rho}(\zeta)\lambda_j} |\varphi_j^{\mathbb{C}}(\zeta)|^2.$$

The equality (80) follows from the fact that $\lim_{\lambda \rightarrow \infty} \frac{1}{\lambda} \log P_{[0,\lambda]}(\zeta, \bar{\zeta}) = 0$.

We now return to nodal sets, where we will see the same extremal functions arise.

8. Counting nodal lines which touch the boundary in analytic plane domains

It is often possible to obtain more refined results on nodal sets by studying their intersections with some fixed (and often special) hypersurface. This has been most successful in dimension two. In this section, we review the results of [TZ] giving upper bounds on the number of intersections of the nodal set with the boundary of an analytic (or more generally piecewise analytic) plane domain. One may expect that the results of this section can also be generalized to higher dimensions by measuring codimension two nodal hypersurface volumes within the boundary.

Thus we would like to count the number of nodal lines (i.e. components of the nodal set) which touch the boundary. Here we assume that 0 is a regular value so that components of the nodal set are either loops in the interior (closed nodal loops) or curves which touch the boundary in two points (open nodal lines). It is known that for generic piecewise analytic plane domains, zero is a regular value of all the eigenfunctions φ_{λ_j} , i.e. $\nabla \varphi_{\lambda_j} \neq 0$ on $Z_{\varphi_{\lambda_j}} \setminus \partial\Omega$; we then call the nodal set regular. Since the boundary lies in the nodal set for Dirichlet boundary conditions, we remove it from the nodal set before counting components. Henceforth, the number of components of the nodal set in the Dirichlet case means the number of components of $Z_{\varphi_{\lambda_j}} \setminus \partial\Omega$.

In the following, and henceforth, $C_\Omega > 0$ denotes a positive constant depending only on the domain Ω .

THEOREM 8.1. *Let Ω be a piecewise analytic domain and let $n_{\partial\Omega}(\lambda_j)$ be the number of components of the nodal set of the j th Neumann or Dirichlet eigenfunction which intersect $\partial\Omega$. Then there exists C_Ω such that $n_{\partial\Omega}(\lambda_j) \leq C_\Omega \lambda_j$.*

By a piecewise analytic domain $\Omega^2 \subset \mathbb{R}^2$, we mean a compact domain with piecewise analytic boundary, i.e. $\partial\Omega$ is a union of a finite number of piecewise analytic curves which intersect only at their common endpoints. Such domains are often studied as archetypes of domains with ergodic billiards and quantum chaotic eigenfunctions, in particular the Bunimovich stadium or Sinai billiard. Their nodal sets have been the subject of a number of numerical studies (e.g. [BGS, FGS]).

In general, there does not exist a non-trivial lower bound for the number of components touching the boundary. E.g. in a disc, the zero sets of the eigenfunctions are unions of circles concentric with the origin and spokes emanating from the center. Only the spokes intersect the boundary and their number reflects the angular momentum rather than the eigenvalue of the eigenfunction. But we conjecture that for piecewise analytic domains

with ergodic billiards, the the number of complex zeros of $\varphi_{\lambda_j}^{\mathbb{C}}|_{\partial\Omega_{\mathbb{C}}}$ is bounded below by $C_{\Omega}\lambda_j$. We discuss work in progress on this conjecture in §10.

In comparison to the order $O(\lambda_j)$ of the number of boundary nodal points, the total number of connected components of $Z_{\varphi_{\lambda_j}}$ has the upper bound $O(\lambda_j^2)$ by the Courant nodal domain theorem. It is not known in general whether the Courant upper bound is achieved, but we expect that it is often achieved in order of magnitude. In [NS] it is proved that the average number of nodal components of a random spherical harmonic is of order of magnitude λ_j^2 . Thus, the number of components touching the boundary is one order of magnitude below the total number of components.

8.1. Boundary critical points. The article [TZ] also contains a similar estimate on the number of critical points of φ_{λ_j} which occur on the boundary. We denote the boundary critical set by

$$\mathcal{C}_{\varphi_{\lambda_j}} = \{q \in \partial\Omega : (d\varphi_{\lambda_j})(q) = 0\}.$$

In the case of Neumann eigenfunctions, $q \in \mathcal{C}_{\varphi_{\lambda_j}} \iff d(\varphi_{\lambda_j}|_{\partial\Omega}(q)) = 0$ since the normal derivative automatically equals zero on the boundary, while in the Dirichlet case $q \in \mathcal{C}_{\varphi_{\lambda_j}} \iff \partial_{\nu}\varphi_{\lambda_j}(q) = 0$ since the boundary is a level set.

We observe that radial eigenfunctions on the disc are constant on the boundary; thus, boundary critical point sets need not be isolated. We therefore impose a non-degeneracy condition on the tangential derivative $\partial_t(\varphi_{\lambda_j}|_{\partial\Omega})$ to ensure that its zeros are isolated and can be counted. We say that the Neumann problem for a bounded domain has the *asymptotic Schiffer property* if there exists $C > 0$ such that, for all Neumann eigenfunctions φ_{λ_j} with sufficiently large λ_j ,

$$(83) \quad \frac{\|\partial_t\varphi_{\lambda_j}\|_{L^2(\partial\Omega)}}{\|\varphi_{\lambda_j}\|_{L^2(\partial\Omega)}} \geq e^{-C\lambda_j}.$$

Here, ∂_t is the unit tangential derivative, and the L^2 norms refer to the restrictions of the eigenfunction to $\partial\Omega$.

THEOREM 8.2. *Let $\Omega \subset \mathbb{R}^2$ be piecewise real analytic. Suppose that $\varphi_{\lambda_j}|_{\partial\Omega}$ satisfies the asymptotic Schiffer condition (83) in the Neumann case. Then the number of $n_{\text{crit}}(\lambda_j) = \#\mathcal{C}_{\varphi_{\lambda_j}}$ of critical points of a Neumann or Dirichlet eigenfunction φ_{λ_j} which lie on $\partial\Omega$ satisfies $n_{\text{crit}}(\lambda_j) \leq C_{\Omega}\lambda_j$ for some $C_{\Omega} > 0$*

In the case of Dirichlet eigenfunctions, endpoints of open nodal lines are always boundary critical points, since they must be singular points of φ_{λ_j} . Hence, an upper bound for $n_{\text{crit}}(\lambda_j)$ also gives an upper bound for the number of open nodal lines.

COROLLARY 8.3. *Suppose that $\Omega \subset \mathbb{R}^2$ is a piecewise real analytic plane domain. Let $n_{\partial\Omega}(\lambda_j)$ be the number of open nodal lines of the j th Dirichlet*

eigenfunction, i.e. connected components of $\{\varphi_{\lambda_j} = 0\} \subset \Omega^o$ whose closure intersects $\partial\Omega$. Then there exists $C_\Omega > 0$ such that $n_{\partial\Omega}(\lambda_j) \leq C_\Omega \lambda_j$.

There does not exist a non-trivial lower bound on the number of interior critical points [JN].

8.2. Proof by analytic continuation. For the Neumann problem, the boundary nodal points are the same as the zeros of the boundary values $\varphi_{\lambda_j}|_{\partial\Omega}$ of the eigenfunctions. The number of boundary nodal points is thus twice the number of open nodal lines. Hence in the Neumann case, Theorem 8.1 follows from:

THEOREM 8.4. *Suppose that $\Omega \subset \mathbb{R}^2$ is a piecewise real analytic plane domain. Then the number $n(\lambda_j) = \#Z_{\varphi_{\lambda_j}} \cap \partial\Omega$ of zeros of the boundary values $\varphi_{\lambda_j}|_{\partial\Omega}$ of the j th Neumann eigenfunction satisfies $n(\lambda_j) \leq C_\Omega \lambda_j$, for some $C_\Omega > 0$.*

This is a more precise version of Theorem 8.1 since it does not assume that 0 is a regular value. In keeping with the theme of this survey, we prove Theorem 8.4 by analytically continuing the boundary values of the eigenfunctions and counting *complex zeros and critical points* of analytic continuations of Cauchy data of eigenfunctions. When $\partial\Omega \in C^\omega$, the eigenfunctions can be holomorphically continued to an open tube domain in \mathbb{C}^2 projecting over an open neighborhood W in \mathbb{R}^2 of Ω which is independent of the eigenvalue. We denote by $\Omega_{\mathbb{C}} \subset \mathbb{C}^2$ the points $\zeta = x + i\xi \in \mathbb{C}^2$ with $x \in \Omega$. Then $\varphi_{\lambda_j}(x)$ extends to a holomorphic function $\varphi_{\lambda_j}^{\mathbb{C}}(\zeta)$ where $x \in W$ and where $|\xi| \leq \epsilon_0$ for some $\epsilon_0 > 0$.

Assuming $\partial\Omega$ real analytic, we define the (interior) complex nodal set by

$$Z_{\varphi_{\lambda_j}}^{\mathbb{C}} = \{\zeta \in \Omega_{\mathbb{C}} : \varphi_{\lambda_j}^{\mathbb{C}}(\zeta) = 0\},$$

and the (interior) complex critical point set by

$$\mathcal{C}_{\varphi_{\lambda_j}}^{\mathbb{C}} = \{\zeta \in \Omega_{\mathbb{C}} : d\varphi_{\lambda_j}^{\mathbb{C}}(\zeta) = 0\}.$$

THEOREM 8.5. *Suppose that $\Omega \subset \mathbb{R}^2$ is a piecewise real analytic plane domain, and denote by $(\partial\Omega)_{\mathbb{C}}$ the union of the complexifications of its real analytic boundary components.*

- (1) *Let $n(\lambda_j, \partial\Omega_{\mathbb{C}}) = \#Z_{\varphi_{\lambda_j}}^{\partial\Omega_{\mathbb{C}}}$ be the number of complex zeros on the complex boundary. Then there exists a constant $C_\Omega > 0$ independent of the radius of $(\partial\Omega)_{\mathbb{C}}$ such that $n(\lambda_j, \partial\Omega_{\mathbb{C}}) \leq C_\Omega \lambda_j$.*
- (2) *Suppose that the Neumann eigenfunctions satisfy (83) and let $n_{crit}(\lambda_j, \partial\Omega_{\mathbb{C}}) = \#\mathcal{C}_{\varphi_{\lambda_j}}^{\partial\Omega_{\mathbb{C}}}$. Then there exists $C_\Omega > 0$ independent of the radius of $(\partial\Omega)_{\mathbb{C}}$ such that $n_{crit}(\lambda_j, \partial\Omega_{\mathbb{C}}) \leq C_\Omega \lambda_j$.*

The theorems on real nodal lines and critical points follow from the fact that real zeros and critical points are also complex zeros and critical points,

hence

$$(84) \quad n(\lambda_j) \leq n(\lambda_j, \partial\Omega_{\mathbb{C}}); \quad n_{\text{crit}}(\lambda_j) \leq n_{\text{crit}}(\lambda_j, \partial\Omega_{\mathbb{C}}).$$

All of the results are sharp, and are already obtained for certain sequences of eigenfunctions on a disc (see §4.6). If the condition (83) is not satisfied, the boundary value of φ_{λ_j} must equal a constant C_j modulo an error of the form $o(e^{-C\lambda_j})$. We conjecture that this forces the boundary values to be constant.

The method of proof of Theorem 8.5 generalizes from $\partial\Omega$ to a rather large class of real analytic curves $C \subset \Omega$, even when $\partial\Omega$ is not real analytic. Let us call a real analytic curve C a *good* curve if there exists a constant $a > 0$ so that for all λ_j sufficiently large,

$$(85) \quad \frac{\|\varphi_{\lambda_j}\|_{L^2(\partial\Omega)}}{\|\varphi_{\lambda_j}\|_{L^2(C)}} \leq e^{a\lambda_j}.$$

Here, the L^2 norms refer to the restrictions of the eigenfunction to C and to $\partial\Omega$. The following result deals with the case where $C \subset \partial\Omega$ is an *interior* real-analytic curve. The real curve C may then be holomorphically continued to a complex curve $C_{\mathbb{C}} \subset \mathbb{C}^2$ obtained by analytically continuing a real analytic parametrization of C .

THEOREM 8.6. *Suppose that $\Omega \subset \mathbb{R}^2$ is a C^∞ plane domain, and let $C \subset \Omega$ be a good interior real analytic curve in the sense of (85). Let $n(\lambda_j, C) = \#Z_{\varphi_{\lambda_j}} \cap C$ be the number of intersection points of the nodal set of the j -th Neumann (or Dirichlet) eigenfunction with C . Then there exists $A_{C,\Omega} > 0$ depending only on C, Ω such that $n(\lambda_j, C) \leq A_{C,\Omega}\lambda_j$.*

A recent paper of J. Jung shows that many natural curves in the hyperbolic plane are ‘good’ [JJ].

8.3. Application to Pleijel’s conjecture. We also note an interesting application due to I. Polterovich [Po] of Theorem 8.1 to an old conjecture of A. Pleijel regarding Courant’s nodal domain theorem, which says that the number n_k of nodal domains (components of $\Omega \setminus Z_{\varphi_{\lambda_k}}$) of the k th eigenfunction satisfies $n_k \leq k$. Pleijel [P] improved this result for Dirichlet eigenfunctions of plane domains: For any plane domain with Dirichlet boundary conditions, $\limsup_{k \rightarrow \infty} \frac{n_k}{k} \leq \frac{4}{j_1^2} \simeq 0.691\dots$, where j_1 is the first zero of the J_0 Bessel function. He conjectured that the same result should be true for a free membrane, i.e. for Neumann boundary conditions. This was recently proved in the real analytic case by I. Polterovich [Po]. His argument is roughly the following: Pleijel’s original argument applies to all nodal domains which do not touch the boundary, since the eigenfunction is a Dirichlet eigenfunction in such a nodal domain. The argument does not apply to nodal domains which touch the boundary, but by Theorem 8.1 the number of such domains is negligible for the Pleijel bound.

9. Equidistribution of complex nodal sets of real ergodic eigenfunctions on analytic (M, g) with ergodic geodesic flow

We now consider global results when hypotheses are made on the dynamics of the geodesic flow. Use of the global wave operator brings into play the relation between the geodesic flow and the complexified eigenfunctions, and this allows one to prove global results on nodal hypersurfaces that reflect the dynamics of the geodesic flow. In some cases, one can determine not just the volume, but the limit distribution of complex nodal hypersurfaces. Since we have discussed this result elsewhere [Z6] we only briefly review it here.

The complex nodal hypersurface of an eigenfunction is defined by

$$(86) \quad Z_{\varphi_\lambda^\mathbb{C}} = \{\zeta \in B_{\epsilon_0}^* M : \varphi_\lambda^\mathbb{C}(\zeta) = 0\}.$$

There exists a natural current of integration over the nodal hypersurface in any ball bundle $B_\epsilon^* M$ with $\epsilon < \epsilon_0$, given by

$$(87) \quad \langle [Z_{\varphi_\lambda^\mathbb{C}}], \varphi \rangle = \frac{i}{2\pi} \int_{B_\epsilon^* M} \partial \bar{\partial} \log |\varphi_\lambda^\mathbb{C}|^2 \wedge \varphi = \int_{Z_{\varphi_\lambda^\mathbb{C}}} \varphi, \quad \varphi \in \mathcal{D}^{(m-1, m-1)}(B_\epsilon^* M).$$

In the second equality we used the Poincaré-Lelong formula. The notation $\mathcal{D}^{(m-1, m-1)}(B_\epsilon^* M)$ stands for smooth test $(m-1, m-1)$ -forms with support in $B_\epsilon^* M$.

The nodal hypersurface $Z_{\varphi_\lambda^\mathbb{C}}$ also carries a natural volume form $|Z_{\varphi_\lambda^\mathbb{C}}|$ as a complex hypersurface in a Kähler manifold. By Wirtinger's formula, it equals the restriction of $\frac{\omega_g^{m-1}}{(m-1)!}$ to $Z_{\varphi_\lambda^\mathbb{C}}$. Hence, one can regard $Z_{\varphi_\lambda^\mathbb{C}}$ as defining the measure

$$(88) \quad \langle |Z_{\varphi_\lambda^\mathbb{C}}|, \varphi \rangle = \int_{Z_{\varphi_\lambda^\mathbb{C}}} \varphi \frac{\omega_g^{m-1}}{(m-1)!}, \quad \varphi \in C(B_\epsilon^* M).$$

We prefer to state results in terms of the current $[Z_{\varphi_\lambda^\mathbb{C}}]$ since it carries more information.

THEOREM 9.1. *Let (M, g) be real analytic, and let $\{\varphi_{j_k}\}$ denote a quantum ergodic sequence of eigenfunctions of its Laplacian Δ . Let $(B_{\epsilon_0}^* M, J)$ be the maximal Grauert tube around M with complex structure J_g adapted to g . Let $\epsilon < \epsilon_0$. Then:*

$$\frac{1}{\lambda_{j_k}} [Z_{\varphi_{j_k}^\mathbb{C}}] \rightarrow \frac{i}{\pi} \partial \bar{\partial} \sqrt{\rho} \text{ weakly in } \mathcal{D}'^{(1,1)}(B_\epsilon^* M),$$

in the sense that, for any continuous test form $\psi \in \mathcal{D}^{(m-1, m-1)}(B_\epsilon^* M)$, we have

$$\frac{1}{\lambda_{j_k}} \int_{Z_{\varphi_{j_k}^\mathbb{C}}} \psi \rightarrow \frac{i}{\pi} \int_{B_\epsilon^* M} \psi \wedge \partial \bar{\partial} \sqrt{\rho}.$$

Equivalently, for any $\varphi \in C(B_\epsilon^* M)$,

$$\frac{1}{\lambda_{j_k}} \int_{Z_{\varphi_{j_k}^C}} \varphi \frac{\omega_g^{m-1}}{(m-1)!} \rightarrow \frac{i}{\pi} \int_{B_\epsilon^* M} \varphi \partial \bar{\partial} \sqrt{\rho} \wedge \frac{\omega_g^{m-1}}{(m-1)!}.$$

COROLLARY 9.2. *Let (M, g) be a real analytic with ergodic geodesic flow. Let $\{\varphi_{j_k}\}$ denote a full density ergodic sequence. Then for all $\epsilon < \epsilon_0$,*

$$\frac{1}{\lambda_{j_k}} [Z_{\varphi_{j_k}^C}] \rightarrow \frac{i}{\pi} \partial \bar{\partial} \sqrt{\rho}, \text{ weakly in } \mathcal{D}'^{(1,1)}(B_\epsilon^* M).$$

The proof consists of three ingredients:

- (1) By the Poincaré-Lelong formula, $[Z_{\varphi_\lambda^C}] = i \partial \bar{\partial} \log |\varphi_\lambda^C|$. This reduces the theorem to determining the limit of $\frac{1}{\lambda} \log |\varphi_\lambda^C|$.
- (2) $\frac{1}{\lambda} \log |\varphi_\lambda^C|$ is a sequence of PSH functions which are uniformly bounded above by $\sqrt{\rho}$. By a standard compactness theorem, the sequence is pre-compact in L^1 : every sequence from the family has an L^1 convergent subsequence.
- (3) $|\varphi_\lambda^C|^2$, when properly L^2 normalized on each ∂M_τ is a quantum ergodic sequence on ∂M_τ . This property implies that the L^2 norm of $|\varphi_\lambda^C|^2$ on ∂M_τ is asymptotically $e^{\lambda_j \tau}$.
- (4) Ergodicity and the calculation of the L^2 norm imply that the only possible L^1 limit of $\frac{1}{\lambda} \log |\varphi_\lambda^C|$ is $\sqrt{\rho}$. This concludes the proof.

We note that the first two steps are valid on any real analytic (M, g) . The difference is that the L^2 norms of φ_λ^C may depend on the subsequence and can often not equal $\sqrt{\rho}$. That is, $\frac{1}{\lambda} |\varphi_\lambda^C|$ behaves like the maximal PSH function in the ergodic case, but not in general. For instance, on a flat torus, the complex zero sets of ladders of eigenfunctions concentrate on a real hypersurface in M_C . This may be seen from the complexified real eigenfunctions $\sin \langle k, x + i\xi \rangle$, which vanish if and only if $\langle k, x \rangle \in 2\pi\mathbb{Z}$ and $\langle k, \xi \rangle = 0$. Here, $k \in \mathbb{N}^m$ is a lattice point. The exact limit distribution depends on which ray or ladder of lattice points one takes in the limit. The result reflects the quantum integrability of the flat torus, and a similar (but more complicated) description of the zeros exists in all quantum integrable cases. The fact that $\frac{1}{\lambda} \log |\varphi_\lambda^C|$ is pre-compact on a Grauert tube of any real analytic Riemannian manifold confirms the upper bound on complex nodal hypersurface volumes.

10. Intersections of nodal sets and geodesics on real analytic surfaces

In §8 we discussed upper bounds on the number of intersection points of the nodal set with the boundary of a real analytic plane domain and more general ‘good’ analytic curves. In this section, we discuss work in progress on intersections of nodal sets and geodesics on surfaces with ergodic geodesic flow. Of course, the results are only tentative but it seems worthwhile at

this point in time to explain the role of ergodicity in obtaining lower bounds and asymptotics. We restrict to geodesic curves because they have rather special properties that makes the analysis somewhat different than for more general curves such as distance circles. The dimensional restriction is due to the fact that the results are partly based on the quantum ergodic restriction theorems of [TZ2, TZ3], which concern restrictions of eigenfunctions to hypersurfaces. Nodal sets and geodesics have complementary dimensions and intersect in points, and therefore it makes sense to count the number of intersections.

We fix $(x, \xi) \in S^*M$ and let

$$(89) \quad \gamma_{x,\xi} : \mathbb{R} \rightarrow M, \quad \gamma_{x,\xi}(0) = x, \quad \gamma'_{x,\xi}(0) = \xi \in T_x M$$

denote the corresponding parametrized geodesic. Our goal is to determine the asymptotic distribution of intersection points of $\gamma_{x,\xi}$ with the nodal set of a highly eigenfunction. As usual, we cannot cope with this problem in the real domain and therefore analytically continue it to the complex domain. Thus, we consider the intersections

$$\mathcal{N}_{\lambda_j}^{\gamma_{x,\xi}^{\mathbb{C}}} = Z_{\varphi_j^{\mathbb{C}}} \cap \gamma_{x,\xi}^{\mathbb{C}}$$

of the complex nodal set with the (image of the) complexification of a generic geodesic. If

$$(90) \quad S_\epsilon = \{(t + i\tau \in \mathbb{C} : |\tau| \leq \epsilon\}$$

then $\gamma_{x,\xi}$ admits an analytic continuation

$$(91) \quad \gamma_{x,\xi}^{\mathbb{C}} : S_\epsilon \rightarrow M_\epsilon.$$

In other words, we consider the zeros of the pullback,

$$\{\gamma_{x,\xi}^* \varphi_{\lambda}^{\mathbb{C}} = 0\} \subset S_\epsilon.$$

We encode the discrete set by the measure

$$(92) \quad [\mathcal{N}_{\lambda_j}^{\gamma_{x,\xi}^{\mathbb{C}}}] = \sum_{(t+i\tau) : \varphi_j^{\mathbb{C}}(\gamma_{x,\xi}^{\mathbb{C}}(t+i\tau))=0} \delta_{t+i\tau}.$$

We would like to show that for generic geodesics, the complex zeros on the complexified geodesic condense on the real points and become uniformly distributed with respect to arc-length. This does not always occur: as in our discussion of QER theorems, if $\gamma_{x,\xi}$ is the fixed point set of an isometric involution, then “odd” eigenfunctions under the involution will vanish on the geodesic. The additional hypothesis is that QER holds for $\gamma_{x,\xi}$, i.e. that Theorem 5.6 is valid. The following is proved ([Z3]):

THEOREM 10.1. *Let (M^2, g) be a real analytic Riemannian surface with ergodic geodesic flow. Let $\gamma_{x,\xi}$ satisfy the QER hypothesis. Then there exists*

a subsequence of eigenvalues λ_{j_k} of density one such that for any $f \in C_c(S_\epsilon)$,

$$\lim_{k \rightarrow \infty} \sum_{(t+i\tau) : \varphi_j^C(\gamma_{x,\xi}^C(t+i\tau))=0} f(t+i\tau) = \int_{\mathbb{R}} f(t) dt.$$

In other words,

$$\text{weak}^* \lim_{k \rightarrow \infty} \frac{i}{\pi \lambda_{j_k}} [\mathcal{N}_{\lambda_j}^{\gamma_{x,\xi}^C}] = \delta_{\tau=0},$$

in the sense of weak* convergence on $C_c(S_\epsilon)$. Thus, the complex nodal set intersects the (parametrized) complexified geodesic in a discrete set which is asymptotically (as $\lambda \rightarrow \infty$) concentrated along the real geodesic with respect to its arclength.

This concentration- equidistribution result is a ‘restricted’ version of the result of §9. As noted there, the limit distribution of complex nodal sets in the ergodic case is a singular current $dd^c \sqrt{\rho}$. The motivation for restricting to geodesics is that restriction magnifies the singularity of this current. In the case of a geodesic, the singularity is magnified to a delta-function; for other curves there is additionally a smooth background measure.

The assumption of ergodicity is crucial. For instance, in the case of a flat torus, say \mathbb{R}^2/L where $L \subset \mathbb{R}^2$ is a generic lattice, the real eigenfunctions are $\cos\langle \lambda, x \rangle, \sin\langle \lambda, x \rangle$ where $\lambda \in L^*$, the dual lattice, with eigenvalue $-|\lambda|^2$. Consider a geodesic $\gamma_{x,\xi}(t) = x + t\xi$. Due to the flatness, the restriction $\sin\langle \lambda, x_0 + t\xi_0 \rangle$ of the eigenfunction to a geodesic is an eigenfunction of the Laplacian $-\frac{d^2}{dt^2}$ of submanifold metric along the geodesic with eigenvalue $-\langle \lambda, \xi_0 \rangle^2$. The complexification of the restricted eigenfunction is $\sin\langle \lambda, x_0 + (t+i\tau)\xi_0 \rangle$ and its exponent of its growth is $\tau|\langle \frac{\lambda}{|\lambda|}, \xi_0 \rangle|$, which can have a wide range of values as the eigenvalue moves along different rays in L^* . The limit current is $i\partial\bar{\partial}$ applied to the limit and thus also has many limits

The proof involves several new principles which played no role in the global result of §9 and which are specific to geodesics. However, the first steps in the proof are the same as in the global case. By the Poincaré-Lelong formula, we may express the current of summation over the intersection points in (92) in the form,

$$(93) \quad [\mathcal{N}_{\lambda_j}^{\gamma_{x,\xi}^C}] = i\partial\bar{\partial}_{t+i\tau} \log \left| \gamma_{x,\xi}^* \varphi_{\lambda_j}^C(t+i\tau) \right|^2.$$

Thus, the main point of the proof is to determine the asymptotics of $\frac{1}{\lambda_j} \log \left| \gamma_{x,\xi}^* \varphi_{\lambda_j}^C(t+i\tau) \right|^2$. When we freeze τ we put

$$(94) \quad \gamma_{x,\xi}^\tau(t) = \gamma_{x,\xi}^C(t+i\tau).$$

PROPOSITION 10.2. (Growth saturation) *If $\{\varphi_{j_k}\}$ satisfies QER along any arcs of $\gamma_{x,\xi}$, then in $L^1_{loc}(S_\tau)$, we have*

$$\lim_{k \rightarrow \infty} \frac{1}{\lambda_{j_k}} \log \left| \gamma_{x,\xi}^* \varphi_{\lambda_{j_k}}^C(t+i\tau) \right|^2 = |\tau|.$$

Proposition 10.2 immediately implies Theorem 10.1 since we can apply $\partial\bar{\partial}$ to the L^1 convergent sequence $\frac{1}{\lambda_{j_k}} \log |\gamma_{x,\xi}^* \varphi_{\lambda_{j_k}}^{\mathbb{C}}(t + i\tau)|^2$ to obtain $\partial\bar{\partial}|\tau|$.

The upper bound in Proposition 10.2 follows immediately from the known global estimate

$$\lim_{k \rightarrow \infty} \frac{1}{\lambda_j} \log |\varphi_{j_k}(\gamma_{x,\xi}^{\mathbb{C}}(\zeta))| \leq |\tau|$$

on all of ∂M_τ . Hence the difficult point is to prove that this growth rate is actually obtained upon restriction to $\gamma_{x,\xi}^{\mathbb{C}}$. This requires new kinds of arguments related to the QER theorem.

- Complexifications of restrictions of eigenfunctions to geodesics have incommensurate Fourier modes, i.e. higher modes are exponentially larger than lower modes.
- The quantum ergodic restriction theorem in the real domain shows that the Fourier coefficients of the top allowed modes are ‘large’ (i.e. as large as the lower modes). Consequently, the L^2 norms of the complexified eigenfunctions along arcs of $\gamma_{x,\xi}^{\mathbb{C}}$ achieve the lower bound of Proposition 10.2.
- Invariance of Wigner measures along the geodesic flow implies that the Wigner measures of restrictions of complexified eigenfunctions to complexified geodesics should tend to constant multiples of Lebesgue measures dt for each $\tau > 0$. Hence the eigenfunctions everywhere on $\gamma_{x,\xi}^{\mathbb{C}}$ achieve the growth rate of the L^2 norms.

These principles are most easily understood in the case of periodic geodesics. We let $\gamma_{x,\xi} : S^1 \rightarrow M$ parametrize the geodesic with arc-length (where $S^1 = \mathbb{R}/L\mathbb{Z}$ where L is the length of $\gamma_{x,\xi}$).

First, we use Theorem 5.6 to prove

LEMMA 10.3. *Assume that $\{\varphi_j\}$ satsfies QER along the periodic geodesic $\gamma_{x,\xi}$. Let $\|\gamma_{x,\xi}^{\tau*} \varphi_j^{\mathbb{C}}\|_{L^2(S^1)}^2$ be the L^2 -norm of the complexified restriction of φ_j along $\gamma_{x,\xi}^\tau$. Then,*

$$\lim_{\lambda_j \rightarrow \infty} \frac{1}{\lambda_j} \log \|\gamma_{x,\xi}^{\tau*} \varphi_j^{\mathbb{C}}\|_{L^2(S^1)}^2 = |\tau|.$$

To prove Lemma 10.3, we study the orbital Fourier series of $\gamma_{x,\xi}^{\tau*} \varphi_j$ and of its complexification. The orbital Fourier coefficients are

$$\nu_{\lambda_j}^{x,\xi}(n) = \frac{1}{L_\gamma} \int_0^{L_\gamma} \varphi_{\lambda_j}(\gamma_{x,\xi}(t)) e^{-\frac{2\pi int}{L_\gamma}} dt,$$

and the orbital Fourier series is

$$(95) \quad \varphi_{\lambda_j}(\gamma_{x,\xi}(t)) = \sum_{n \in \mathbb{Z}} \nu_{\lambda_j}^{x,\xi}(n) e^{\frac{2\pi int}{L_\gamma}}.$$

Hence the analytic continuation of $\gamma_{x,\xi}^{\tau*}\varphi_j$ is given by

$$(96) \quad \varphi_{\lambda_j}^{\mathbb{C}}(\gamma_{x,\xi}(t + i\tau)) = \sum_{n \in \mathbb{Z}} \nu_{\lambda_j}^{x,\xi}(n) e^{\frac{2\pi i n(t+i\tau)}{L_\gamma}}.$$

By the Paley-Wiener theorem for Fourier series, the series converges absolutely and uniformly for $|\tau| \leq \epsilon_0$. By “energy localization” only the modes with $|n| \leq \lambda_j$ contribute substantially to the L^2 norm. We then observe that the Fourier modes decouple, since they have different exponential growth rates. We use the QER hypothesis in the following way:

LEMMA 10.4. *Suppose that $\{\varphi_{\lambda_j}\}$ is QER along the periodic geodesic $\gamma_{x,\xi}$. Then for all $\epsilon > 0$, there exists $C_\epsilon > 0$ so that*

$$\sum_{n:|n| \geq (1-\epsilon)\lambda_j} |\nu_{\lambda_j}^{x,\xi}(n)|^2 \geq C_\epsilon.$$

Lemma 10.4 implies Lemma 10.3 since it implies that for any $\epsilon > 0$,

$$\sum_{n:|n| \geq (1-\epsilon)\lambda_j} |\nu_{\lambda_j}^{x,\xi}(n)|^2 e^{-2n\tau} \geq C_\epsilon e^{2\tau(1-\epsilon)\lambda_j}.$$

To go from asymptotics of L^2 norms of restrictions to Proposition 10.2 we then use the third principle:

PROPOSITION 10.5. *(Lebesgue limits) If $\gamma_{x,\xi}^*\varphi_j \neq 0$ (identically), then for all $\tau > 0$ the sequence*

$$U_j^{x,\xi,\tau} = \frac{\gamma_{x,\xi}^{\tau*}\varphi_j^{\mathbb{C}}}{\|\gamma_{x,\xi}^{\tau*}\varphi_j^{\mathbb{C}}\|_{L^2(S^1)}}$$

is QUE with limit measure given by normalized Lebesgue measure on S^1 .

The proof of Proposition 10.2 is completed by combining Lemma 10.3 and Proposition 10.5. Theorem 10.1 follows easily from Proposition 6.1.

The proof for non-periodic geodesics is considerably more involved, since one cannot use Fourier analysis in quite the same way.

11. Nodal and critical sets of Riemannian random waves

We mentioned above that Riemannian random waves provide a probabilistic model that is conjectured to predict the behavior of eigenfunctions when the geodesic flow of (M, g) is ergodic. In this section, we define the model precisely as in [Z4] (see also [Nic] for a similar model) and survey some of the current results and conjectures. We should emphasize that some of the rigorous results on zeros or critical points of Riemannian random waves, both in the real and complex domain, are much simpler than for individual eigenfunctions, and therefore do not provide much guidance on how to prove results for an orthonormal basis of eigenfunctions. But the relative

simplicity of random waves and their value as predictors provide the motivation for studying random waves. And there are many hopelessly difficult problems on random waves as well, which we will survey in this section.

For expository simplicity we assume that the geodesic flow G^t of (M, g) is of one of the following two types:

- (1) *aperiodic*: The Liouville measure of the closed orbits of G^t , i.e. the set of vectors lying on closed geodesics, is zero; or
- (2) *periodic = Zoll*: $G^T = id$ for some $T > 0$; henceforth T denotes the minimal period. The common Morse index of the T -periodic geodesics will be denoted by β .

In the real analytic case, (M, g) is automatically one of these two types, since a positive measure of closed geodesics implies that all geodesics are closed. The two-term Weyl laws counting eigenvalues of $\sqrt{\Delta}$ are very different in these two cases.

- (1) In the *aperiodic* case, Ivrii's two term Weyl law states

$$N(\lambda) = \#\{j : \lambda_j \leq \lambda\} = c_m \operatorname{Vol}(M, g) \lambda^m + o(\lambda^{m-1})$$

where $m = \dim M$ and where c_m is a universal constant.

- (2) In the *periodic* case, the spectrum of $\sqrt{\Delta}$ is a union of eigenvalue clusters C_N of the form

$$C_N = \left\{ \left(\frac{2\pi}{T} \right) \left(N + \frac{\beta}{4} \right) + \mu_{Ni}, \quad i = 1 \dots d_N \right\}$$

with $\mu_{Ni} = O(N^{-1})$. The number d_N of eigenvalues in C_N is a polynomial of degree $m - 1$.

We refer to [HoI-IV, Z4] for background and further discussion.

To define Riemannian random waves, we partition the spectrum of $\sqrt{\Delta_g}$ into certain intervals I_N of width one and denote by Π_{I_N} the spectral projections for $\sqrt{\Delta_g}$ corresponding to the interval I_N . The choice of the intervals I_N is rather arbitrary for aperiodic (M, g) and as mentioned above we assume $I_N = [N, N + 1]$. In the Zoll case, we center the intervals around the center points $\frac{2\pi}{T}N + \frac{\beta}{4}$ of the N th cluster C_N . We call such a choice of intervals a cluster decomposition. We denote by d_N the number of eigenvalues in I_N and put $\mathcal{H}_N = \operatorname{ran} \Pi_{I_N}$ (the range of Π_{I_N}).

We choose an orthonormal basis $\{\varphi_{Nj}\}_{j=1}^{d_N}$ for \mathcal{H}_N . For instance, on S^2 one can choose the real and imaginary parts of the standard Y_m^N 's. We endow the real vector space \mathcal{H}_N with the Gaussian probability measure γ_N defined by

$$(97) \quad \gamma_N(f) = \left(\frac{d_N}{\pi} \right)^{d_N/2} e^{-d_N|c|^2} dc, \quad f = \sum_{j=1}^{d_N} c_j \varphi_{Nj}, \quad d_N = \dim \mathcal{H}_N.$$

Here, dc is d_N -dimensional real Lebesgue measure. The normalization is chosen so that $\mathbb{E}_{\gamma_N} \langle f, f \rangle = 1$, where \mathbb{E}_{γ_N} is the expected value with respect to γ_N . Equivalently, the d_N real variables c_j ($j = 1, \dots, d_N$) are independent

identically distributed (i.i.d.) random variables with mean 0 and variance $\frac{1}{2d_N}$; i.e.,

$$\mathbb{E}_{\gamma_N} c_j = 0, \quad \mathbb{E}_{\gamma_N} c_j c_k = \frac{1}{2d_N} \delta_{jk}.$$

We note that the Gaussian ensemble is equivalent to picking $f_N \in \mathcal{H}_N$ at random from the unit sphere in \mathcal{H}_N with respect to the L^2 inner product.

Depending on the choice of intervals, we obtain the following special ensembles:

- The asymptotically *fixed frequency* ensemble \mathcal{H}_{I_λ} , where $I_\lambda = [\lambda, \lambda + 1]$ and where \mathcal{H}_{I_λ} is the vector space of linear combinations

$$(98) \quad f_\lambda = \sum_{j: \lambda_j \in [\lambda, \lambda+1]} c_j \varphi_{\lambda_j},$$

of eigenfunctions with λ_j (the frequency) in an interval $[\lambda, \lambda + 1]$ of fixed width. (Note that it is the square root of the eigenvalue of Δ , not the eigenvalue, which is asymptotically fixed).

- The *high frequency cut-off* ensembles $\mathcal{H}_{[0, \lambda]}$ where the frequency is cut-off at λ :

$$(99) \quad f_\lambda = \sum_{j: \lambda_j \leq \lambda} c_j \varphi_{\lambda_j}.$$

- The *cut-off Gaussian free field*,

$$(100) \quad f_\lambda = \sum_{j: \lambda_j \leq \lambda} c_j \frac{\varphi_{\lambda_j}}{\lambda_j}.$$

One could use more general weights $w(\lambda_j)$ on a Sobolev space of functions or distributions on M . In the physics terminology, $w(\lambda_j)$ (or its square) is referred to as the power spectrum.

The key reason why we can study the limit distribution of nodal sets in this ensemble is that the covariance kernel

$$(101) \quad \Pi_{I_N}(x, y) = \mathbb{E}_{\gamma_N}(f_N(x)f_N(y)) = \sum_{j: \lambda_j \in I_N} \varphi_{\lambda_j}(x)\varphi_{\lambda_j}(y),$$

is the spectral projections kernel for $\sqrt{\Delta}$.

11.1. Equidistribution of nodal sets for almost all sequences of random waves. The real zeros are straightforward to define. For each $f_\lambda \in \mathcal{H}_{[0, \lambda]}$ or \mathcal{H}_{I_λ} we associate to the zero set $Z_{f_\lambda} = \{x \in M : f_\lambda(x) = 0\}$ the positive measure

$$(102) \quad \langle |Z_{f_\lambda}|, \psi \rangle = \int_{Z_{f_\lambda}} \psi d\mathcal{H}^{n-1},$$

where $d\mathcal{H}^{n-1}$ is the induced (Hausdorff) hypersurface measure.

The main result we review is the limit law for random sequences of random real Riemannian waves. By a random sequence, we mean an element of the product probability space

$$(103) \quad \mathcal{H}_\infty = \prod_{N=1}^{\infty} \mathcal{H}_N, \quad \gamma_\infty = \prod_{N=1}^{\infty} \gamma_N.$$

THEOREM 11.1. [Z4] *Let (M, g) be a compact Riemannian manifold, and let $\{f_N\}$ be a random sequence in (103). Then*

$$\frac{1}{N} \sum_{n=1}^N \frac{1}{\lambda_n} |Z_{f_n}| \rightarrow dV_g \quad \text{almost surely w.r.t. } (\mathcal{H}_\infty, \gamma_\infty).$$

11.2. Mean and variance. We first show that the normalized expected limit distribution $\frac{1}{\lambda} \mathbb{E} |Z_{f_\lambda}|$ of zeros of random Riemannian waves tends to the volume form dV_g as $\lambda \rightarrow \infty$. That is, we define the ‘linear statistic’,

$$(104) \quad X_\psi^N(f_N) = \langle \psi, |Z_{f_N}| \rangle, \quad \psi \in C(M)$$

and then define

$$(105) \quad \langle \mathbb{E}_{\gamma_N} |Z_{f_N}|, \psi \rangle = \mathbb{E}_{\gamma_N} X_\psi^N,$$

THEOREM 11.2. *Let (M, g) be a compact Riemannian manifold, let $\mathcal{H}_{[0, \lambda]}$ be the cutoff ensemble and let $(\mathcal{H}_N, \gamma_N)$ be the ensemble of Riemannian waves of asymptotically fixed frequency. Then in either ensemble:*

- (1) *For any $C^\infty(M, g)$, $\lim_{N \rightarrow \infty} \frac{1}{N} \mathbf{E}_{\gamma_N} \langle |Z_{f_N}|, \psi \rangle = \int_M \psi dV_g$.*
- (2) *For a real analytic (M, g) , $\text{Var}(\frac{1}{N} X_\psi^N) \leq C$.*

We restrict to real analytic metrics in (2) for the sake of brevity. In that case, the variance estimate follows easily from Theorem 2.1.

11.3. Density of real zeros. The formula for the density of zeros of random elements of \mathcal{H}_N can be derived from the general Kac-Rice formula [BSZ1, BSZ2, Nic]:

$$(106) \quad \mathbb{E} |Z_{f_N}| = K_1^N(z) dV_g, \quad K_1^N(x) = \int D(0, \xi, x) \|\xi\| d\xi.$$

Here, $D(q, \xi, x) dq d\xi$ is the joint probability distribution of the Gaussian random variables $(\psi(x), \nabla \psi(x))$, i.e. the pushforward of the Gaussian measure on \mathcal{H}_λ under the map $\psi \rightarrow (\psi(x), \nabla \psi(x))$. Note that the factor $\det(\xi \xi^*)$ in [BSZ1, BSZ2] equals $\|\xi\|^2$ in the codimension one case. Indeed, let df_x^* be the adjoint map with respect to the inner product g on $T_x M$. Let $df_x \circ df_x^* : \mathbb{R} \rightarrow \mathbb{R}$ be the composition. By $\det df_x \circ df_x^*$ is meant the determinant with respect to the inner product on $T_x M$; it clearly equals $|df|^2$ in the codimension one case.

The formulae of [**BSZ1**, **BSZ2**] (the ‘Kac-Rice’ formulae) give that

$$(107) \quad D(0, \xi; z) = Z_n(z) D_\Lambda(\xi; z),$$

where

$$(108) \quad D_\Lambda(\xi; z) = \frac{1}{\pi^m \sqrt{\det \Lambda}} \exp(-\langle \Lambda^{-1} \xi, \xi \rangle)$$

is the Gaussian density with covariance matrix

$$(109) \quad \Lambda = C - B^* A^{-1} B = \left(C_{q'}^q - B_q A^{-1} B_{q'} \right), \quad (q = 1, \dots, m)$$

and

$$(110) \quad Z(x) = \frac{\sqrt{\det \Lambda}}{\pi \sqrt{\det \Delta}} = \frac{1}{\pi \sqrt{A}}.$$

Here,

$$\begin{aligned} \Delta = \Delta^N(x) &= \begin{pmatrix} A^N & B^N \\ B^{N*} & C^N \end{pmatrix}, \\ (A^N) &= \mathbb{E}(X^2) = \frac{1}{d_N} \Pi_{I_N}(x, x), \\ (B^N)_q &= \mathbb{E}(X \Xi_q) = \frac{1}{d_N} \frac{\partial}{\partial y_q} \Pi_{I_N}(x, y)|_{x=y}, \\ (C^\lambda)_{q'}^q &= \mathbb{E}(\Xi_q \Xi_{q'}) = \frac{1}{d_N} \frac{\partial^2}{\partial x_q \partial y_{q'}} \Pi_{I_N}(x, y)|_{x=y}, \\ &\quad q, q' = 1, \dots, m. \end{aligned}$$

Making a simple change of variables in the integral (106), we have

PROPOSITION 11.3. [BSZ1] *On a real Riemannian manifold of dimension m , the density of zeros of a random Riemannian wave is*

$$(111) \quad K_1^N(x) = \frac{1}{\pi^m (\sqrt{d_N^{-1} \Pi_{I_N}(x, x)})} \int_{\mathbb{R}^m} \|\Lambda^N(x)^{1/2} \xi\| \exp(-\langle \xi, \xi \rangle) d\xi,$$

where $\Lambda^N(x)$ is a symmetric form on $T_x M$. For the asymptotically fixed frequency ensembles, it is given by

$$\begin{aligned} \Lambda^N(x) &= \frac{1}{d_N} \left(d_x \otimes d_y \Pi_{I_N}(x, y)|_{x=y} \right. \\ &\quad \left. - \frac{1}{\Pi_{I_N}(x, y)} d_x \Pi_{I_N}(x, y)|_{x=y} \otimes d_y \Pi_{I_N}(x, y)|_{x=y} \right). \end{aligned}$$

In the cutoff ensemble the formula is the same except that Π_{I_N} is replaced by $\Pi_{[0, N]}$.

We then need the asymptotics of the matrix elements of $\Delta^N(x)$. They are simplest for the round sphere, so we state them first in that case:

PROPOSITION 11.4. *Let $\Pi_N : L^2(S^m) \rightarrow \mathcal{H}_N$ be the orthogonal projection. Then:*

- (A) $\Pi_N(x, x) = \frac{1}{Vol(S^m)} d_N;$
- (B) $d_x \Pi_N(x, y)|_{x=y} = d_y \Pi_N(x, y)|_{x=y} = 0;$
- (C) $d_x \otimes d_y \Pi_N(x, y)|_{x=y} = \frac{1}{mVol(S^m)} \lambda_N^2 d_N g_x.$

We refer to [Z4] for the calculation, which is quite simple because of the invariance under rotations. The expected density of random nodal hypersurfaces is given as follows

PROPOSITION 11.5. *In the case of S^m ,*

$$(112) \quad K_1^N(x) = C_m \lambda_N \sim C_m N,$$

where $C_m = \frac{1}{\pi^m} \int_{\mathbb{R}^m} |\xi| \exp(-\langle \xi, \xi \rangle) d\xi.$

PROOF. By Propositiosn 11.3 and 11.4, we have

$$(113) \quad K_1^N(x) = \frac{\sqrt{Vol(S^m)}}{\pi^m} \int_{\mathbb{R}^m} ||\Lambda^N(x)^{1/2} \xi|| \exp(-\langle \xi, \xi \rangle) d\xi,$$

where

$$\Lambda^N(x) = \frac{1}{d_N} \left(\frac{1}{mVol(S^m)} \lambda_N^2 d_N g_x \right).$$

□

11.4. Random Riemannian waves: proof of Theorem 11.2. We now generalize the result to any compact C^∞ Riemannian manifold (M, g) which is either aperiodic or Zoll. As in the case of S^m , the key issue is the asymptotic behavior of derivatives of the spectral projections

$$(114) \quad \Pi_{I_N}(x, y) = \sum_{j: \lambda_j \in I_N} \varphi_{\lambda_j}(x) \varphi_{\lambda_j}(y).$$

PROPOSITION 11.6. *Assume (M, g) is either aperiodic and $I_N = [N, N+1]$ or Zoll and I_N is a cluster decomposition. Let $\Pi_{I_N} : L^2(M) \rightarrow \mathcal{H}_N$ be the orthogonal projection. Then:*

- (A) $\Pi_{I_N}(x, x) = \frac{1}{Vol(M, g)} d_N (1 + o(1));$
- (B) $d_x \Pi_{I_N}(x, y)|_{x=y} = d_y \Pi_{I_N}(x, y)|_{x=y} = o(N^m);$
- (C) $d_x \otimes d_y \Pi_{I_N}(x, y)|_{x=y} = \frac{1}{Vol(M, g)} \lambda_N^2 d_N g_x (1 + o(1)).$

In the aperiodic case,

- (1) $\Pi_{[0, \lambda]}(x, x) = C_m \lambda^m + o(\lambda^{m-1});$
- (2) $d_x \otimes d_y \Pi_{[0, \lambda]}(x, y)|_{x=y} = C_m \lambda^{m+2} g_x + o(\lambda^{m+1}).$

In the Zoll case, one adds the complete asymptotic expansions for Π_{I_N} over the N clusters to obtain expansions for Π_N .

We then have:

PROPOSITION 11.7. *For the asymptotically fixed frequency ensemble, and for any $C^\infty(M, g)$ which is either Zoll or aperiodic (and with I_N as in Proposition 11.6), we have*

$$(115) \quad \begin{aligned} K_1^N(x) &= \frac{1}{\pi^m(\lambda_N)^{m/2}} \int_{\mathbb{R}^m} \|\xi\| \exp\left(-\frac{1}{\lambda_N} \langle \xi, \xi \rangle\right) d\xi + o(1) \\ &\sim C_m N, \end{aligned}$$

where $C_m = \frac{1}{\pi^m} \int_{\mathbb{R}^m} \|\xi\| \exp(-\langle \xi, \xi \rangle) d\xi$. The same formula holds for the cutoff ensemble.

PROOF. Both on a sphere S^m or on a more general (M, g) which is either Zoll or aperiodic, we have by Propositions 11.4 resp. 11.6 and the general formula for Δ^N in §11.3 that

$$(116) \quad \Delta^N(z) = \frac{1}{Vol(M, g)} \begin{pmatrix} (1 + o(1)) & o(1) \\ o(1) & N^2 g_x(1 + o(1)) \end{pmatrix},$$

It follows that

$$(117) \quad \Lambda^N = C^N - B^{N*}(A^N)^{-1}B^N = \frac{1}{Vol(M, g)} N^2 g_x + o(N).$$

Thus, we have

$$(118) \quad \begin{aligned} K_1^N(x) &\sim \frac{\sqrt{Vol(M, g)}}{\pi^m} \int_{\mathbb{R}^m} \|\Lambda^N(x)^{1/2}\xi\| \exp(-\langle \xi, \xi \rangle) d\xi \\ &= \frac{N}{\pi^m} \int_{\mathbb{R}^m} \|(I + o(1))(x)^{1/2}\xi\| \exp(-\langle \xi, \xi \rangle) d\xi, \end{aligned}$$

where $o(1)$ denotes a matrix whose norm is $o(1)$, as as $N \rightarrow \infty$ we obtain the stated asymptotics. \square

So far, we have only determined the expected values of the nodal hypersurface measures. To complete the proof of Theorem 11.2, we need to prove:

PROPOSITION 11.8. *If (M, g) is real analytic, then the variance of $\frac{1}{\lambda_N} X_\psi^N$ is bounded.*

PROOF. By Theorem 2.1, for $f_N \in \mathcal{H}_{I_N}$, $\frac{1}{\lambda_N} Z_{f_N}$ has bounded mass. Hence, the random variable $\frac{1}{\lambda_N} X_\psi^N$ is bounded, and therefore so is its variance. \square

Remark: The variance of $\frac{1}{\lambda_N} X_\psi^N$ is given by

$$(119) \quad \begin{aligned} \text{Var}\left(\frac{1}{\lambda_N} X_\psi^N\right) \\ = \frac{1}{\lambda_N^2} \int_M \int_M (K_2^N(x, y) - K_1^N(x)K_1^N(y)) \psi(x)\psi(y) dV_g(x)dV_g(y), \end{aligned}$$

where $K_2^N(x, y) = \mathbb{E}_{\gamma_N}(Z_{f_N}(x) \otimes Z_{f_N}(y))$ is the pair correlation function for zeros. Hence, boundedness would follow from

$$(120) \quad \frac{1}{\lambda_N^2} \int_M \int_M K_2^N(x, y) dV_g(x)dV_g(y) \leq C.$$

There is a formula similar to that for the density in Proposition 11.3 for $K_2^N(x, y)$ and it is likely that it could be used to prove boundedness of the variance for any C^∞ Riemannian manifold.

11.5. Random sequences and proof of Theorem 11.1. We recall that the set of random sequences of Riemannian waves of increasing frequency is the probability space $\mathcal{H}_\infty = \prod_{N=1}^\infty \mathcal{H}_{I_N}$ with the measure $\gamma_\infty = \prod_{N=1}^\infty \gamma_N$. An element in \mathcal{H}_∞ will be denoted $\mathbf{f} = \{f_N\}$. We have,

$$\left| \left(\frac{1}{\lambda_N} Z_{f_N}, \psi \right) \right| \leq \frac{1}{\lambda_N} \mathcal{H}^{n-1}(Z_{f_N}) \|\psi\|_{C^0}.$$

By a density argument it suffices to prove that the linear statistics $\frac{1}{\lambda_N} (Z_{f_N}, \psi) - \frac{1}{Vol(M, g)} \int_M \psi dV_g \rightarrow 0$ almost surely in \mathcal{H}_∞ . We know that

- (i) $\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k \leq N} \mathbb{E}(\frac{1}{\lambda_k} X_\psi^k) = \frac{1}{Vol(M, g)} \int_M \psi dV_g;$
- (ii) $\text{Var}(\frac{1}{\lambda_N} X_\psi^N)$ is bounded on \mathcal{H}_∞ .

Since $\frac{1}{\lambda_N} X_\psi^N$ for $\{\cdot, N = 1, 2, \dots\}$ is a sequence of independent random variables in \mathcal{H}_∞ with bounded variances, the Kolmogorov strong law of large numbers implies that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k \leq N} \left(\frac{1}{\lambda_k} X_\psi^k \right) = \frac{1}{Vol(M, g)} \int_M \psi dV_g$$

almost surely.

11.6. Complex zeros of random waves. We now state a complex analogue of the equidistribution of real nodal sets and show that it agrees with the limit formula of Theorem 9.1.

We complexify Riemannian random waves as

$$f_N^\mathbb{C} = \sum_{j=1}^{d_N} c_{Nj} \varphi_{Nj}^\mathbb{C}.$$

We note that the coefficients c_{Nj} are real and that the Gaussian measure on the coefficients remains the real Gaussian measure γ_N . The two point

function is the analytic extensions to the totally real anti-diagonal in $M_{\mathbb{C}} \times M_{\mathbb{C}}$ is therefore

$$(121) \quad \mathbb{E}(|f_N(\zeta)|^2) = \Pi_{I_N}(\zeta, \bar{\zeta}) = \sum_{j: \lambda_j \in I_k} |\varphi_j^{\mathbb{C}}(\zeta)|^2.$$

As in the proof of Theorem 9.1, the current of integration over the complex zero set

$$Z_{f_N^{\mathbb{C}}} = \{\zeta \in M_{\mathbb{C}} : f_N^{\mathbb{C}} = 0\}$$

is the $(1, 1)$ current defined by

$$\langle [Z_{f_N^{\mathbb{C}}}], \psi \rangle = \int_{Z_{f_N^{\mathbb{C}}}} \psi, \quad \psi \in \mathcal{D}^{m-1, m-1}(M_{\mathbb{C}}),$$

for smooth test forms of bi-degree $(m-1, m-1)$. In terms of scalar functions ψ we may define $Z_{f_N^{\mathbb{C}}}$ as the measure,

$$\langle [Z_{f_N^{\mathbb{C}}}], \psi \rangle = \int_{Z_{f_N^{\mathbb{C}}}} \psi \omega_g^{m-1} / (m-1)!,$$

where $\omega_g = i\partial\bar{\partial}\rho$ is the Kählermetric adapted to g .

The proof of the next result is close to the proof of Theorem 9.1 and we therefore refer to [Z4] for the details:

THEOREM 11.9. [Z4] *Let (M, g) be a real analytic compact Riemannian manifold. Then for either of the Riemannian random wave ensembles*

$$\mathbb{E}_{\gamma_N} \left(\frac{1}{N} [Z_{f_N^{\mathbb{C}}}] \right) \rightarrow \frac{i}{\pi} \partial\bar{\partial}|\xi|_g, \quad \text{weakly in } \mathcal{D}'^{(1,1)}(B_{\epsilon}^* M).$$

As mentioned above, this result shows that the complex zeros of the random waves have the same expected limit distribution found in [Z3] for real analytic compact Riemannian manifolds with ergodic geodesic flow.

12. Percolation heuristics

In this final section, we review some of the more speculative conjectures relating nodal sets of both eigenfunctions and random waves to percolation theory. The conjectures are often quoted and it therefore seems worthwhile to try to state them precisely. The only rigorous result to date regarding eigenfunctions is the theorem of Nazarov-Sodin on the expected number of nodal domains for random spherical harmonics [NS] (see [Z5] for a brief over-view).

The percolation conjectures concern the statistics of sizes of nodal domains or nodal components. They are based on the idea that the nodal domains resemble percolation clusters. One might measure the ‘size’ of a nodal component A_{λ_j} by its hypersurface area $\mathcal{H}^{n-1}(A_{\lambda,j})$, and a nodal domain $D_{\lambda,j}$ by its volume $\mathcal{H}^n(D_{\lambda,j})$. Let us restrict to the case of surfaces.

For the purposes of this article, we introduce the term *length spectrum* of the nodal set as the set

$$(122) \quad Lsp(\varphi_\lambda) = \{(\mathcal{H}^1(C_{\lambda,j}) : Z_{\varphi_\lambda} = \bigcup C_{\lambda;j}\}$$

of lengths of its components, counted with multiplicity. It is encoded by the empirical measure of surface areas

$$(123) \quad d\mu_L = \frac{1}{\mathcal{H}^1(Z_{\varphi_\lambda})} \sum_{C_{\lambda,j}} \delta_{\mathcal{H}^1(C_{\lambda,j})} \in \mathcal{P}_1(\mathbb{R}),$$

(where $\mathcal{P}(\Omega)$ is the set of probability measures on Ω), or equivalently by the length distribution function,

$$(124) \quad \mathcal{L}_\lambda(t) = \sum_{j : \mathcal{H}^1(C_{\lambda,j}) \leq t} \mathcal{H}^1(C_{\lambda,j}).$$

We also consider the *area spectrum*,

$$(125) \quad Asp(\varphi_\lambda) = \{(\mathcal{H}^2(A_{\lambda,j}) : M \setminus Z_{\varphi_\lambda} = \bigcup A_{\lambda;j}\},$$

encoded by its empirical measure It is encoded by the empirical measure of surface areas

$$(126) \quad d\mu_A = \frac{1}{\text{Area}(M)} \sum_{A_{\lambda,j}} \delta_{\mathcal{H}^2(A_{\lambda,j})} \in \mathcal{P}_1(\mathbb{R}),$$

or by the area distribution function,

$$(127) \quad \mathcal{A}_\lambda(t) = \sum_{j : \mathcal{H}^2(A_{\lambda,j}) \leq t} \mathcal{H}^2(A_{\lambda,j}).$$

Of course, there are some obvious constraints on such spectra; e.g. in the analytic case, there could only exist $O(\lambda)$ components with \mathcal{H}^1 -length of order 1, and only a bounded number of order λ .

In computer graphics of eigenfunctions on plane domains or surfaces, one sees many ‘small’ components $C_{\lambda,j}$ of the nodal set whose length appears to be of order $\frac{1}{\lambda}$. But one also sees long snaky nodal lines. How long are they? Do they persist as $\lambda \rightarrow \infty$? Roughly speaking, one may ask what proportion of the components come in sizes with different orders of magnitude. Of course, this depends on how many components there are, so it could be simpler to work with $\mathcal{L}(\varphi_\lambda), \mathcal{A}(\varphi_\lambda)$.

- How many components have \mathcal{H}^{n-1} -surface measure which is $\geq C\lambda^\gamma$ for some given $0 < \gamma \leq 1$. It is possible that some individual nodal component has \mathcal{H}^{n-1} -surface area commensurate with that of the entire nodal set, as in the Lewy spherical harmonics with just two or three nodal components [**Lew**].
- How many components have \mathcal{H}^{n-1} -surface measure (i.e. length in dimension two) which is bounded below by a constant $C > 0$ independent of λ ? Such components are sometimes termed “percolating

“nodal lines” since their hypersurface volume is commensurate with the size of the macroscopic object (i.e. M).

- How many components have \mathcal{H}^{n-1} -surface measure of the minimal order $\frac{1}{\lambda}$?

The percolation conjectures relate the asymptotic distribution of lengths of nodal components and areas of nodal domains of eigenfunctions as defined in (123)-(126) to lengths of boundaries and areas of percolation clusters at criticality. There are different types of conjectures for the *fixed frequency* ensemble and the *high frequency cutoff* ensemble (see §11 for the definitions). According to the random wave hypothesis, the conjectures concerning the fixed frequency ensemble (e.g. random spherical harmonics of fixed degree) should also apply to nodal sets of eigenfunctions of quantum chaotic systems.

Percolation theory is concerned with connectivity and transport in a complex system. In particular, it studies connected clusters of objects in a random graph. In bond percolation the edges of the graph are independently open or closed with some probability p . The open edges form a subgraph whose connected components form the clusters. In site percolation the vertices are open or closed and an open path is a path through open vertices. The open cluster $C(v)$ of a vertex v is the set of all open vertices which are connected to v by an open path.

There also exists an analogous continuum percolation theory for level sets of random functions. We will assume the random functions are Gaussian Riemannian random waves on a surface. The main problem is to study the connectivity properties of level sets $\{f = t\}$. One imagines a random landscape of lakes and islands depending on the variable height t of the water, the islands being the super-level sets $\{f > t\}$ of the random functions. For high water levels, the islands are disconnected, but as the water level is lowered the islands become more connected. At a critical level t_c they ‘percolate’, i.e. it is possible to traverse the landscape while remaining on the land. A review with many illustrations is given by Isichenko [Isi] (see Section E (c), pages 980-984). As explained in [Isi] page 984, the contour lines of a random potential are associated to hulls of percolation clusters. Hence the area spectrum (125) is similar to the set of sizes of connected clusters in a percolation model.

In the physics literature, the random functions are usually functions on \mathbb{R}^2 (or possibly higher dimensional \mathbb{R}^n) and the Gaussian measure on the space of functions corresponds to a Hilbert space inner product. The Hilbert space is usually taken to be a Sobolev space, so that the inner product has the form $\int w(\xi)|\hat{f}(\xi)|^2d\xi$ (where \hat{f} is the Fourier transform of f) and $w(\xi) = |\xi|^{2(1+\zeta)}$. The case $\zeta = 0$ is known as the Gaussian free field (or massless scalar field) and is quite special in two dimensions since then the inner product $\int_{\mathbb{R}^2} |\nabla f|^2 dx$ is conformally invariant. There are rigorous results on level sets of discretizations of the Gaussian free field and their

continuum limits in [SS, Mi], with authoritative comments on the physics literature.

For purposes of this exposition, we assume the Riemannian random waves fall are of the types discussed in §11. In all cases, we truncate the frequency above a spectral parameter λ and consider asymptotics as $\lambda \rightarrow \infty$. In this high frequency limit, the random waves oscillate more rapidly on the length scale $\frac{1}{\lambda}$. Since the conjectures and results depend strongly on the chosen weight w , we break up the discussion into two cases as in §11: the high frequency cutoff ensemble and the fixed frequency ensemble. For each ensemble we let \mathbb{E}_λ denote the expectation with respect to the Gaussian measure on the relevant space of linear combinations. Then we may ask for the asymptotic behavior of the expected distribution of lengths of nodal lines, resp. area of nodal domains

$$(128) \quad \mathbb{E}_\lambda d\mu_L, \quad \mathbb{E}_\lambda d\mu_A,$$

where $d\mu_L$, resp. $d\mu_A$ are the empirical measures of lengths (123) of nodal lines, resp. areas (126) of nodal domains.

12.1. High frequency cutoff ensembles. The distribution of contour lengths of certain Gaussian random surfaces over \mathbb{R}^2 was studied at the physics level of rigor in [KH]. They define the Gaussian measure as $e^{-f_\zeta(h)}dh$ where the ‘free energy’ is defined by

$$f_\zeta(h) = \frac{K}{2} \int_{\mathbb{R}^2} \chi\left(\frac{|\xi|}{\lambda}\right) |\hat{h}(\xi)|^2 |\xi|^{2(1+\zeta)} d\xi,$$

where χ is a cutoff function to $[0, 1]$ (they use the notation a for $\frac{1}{\lambda}$ in our notation). When $\zeta = 0$, this is a truncated Gaussian free field (truncated at frequencies $\leq \lambda$) and its analogue on a surface (M, g) is the Riemannian random wave model with spectral interval $[0, \lambda]$ and weight $w(\lambda) = \frac{1}{\lambda}$. The parameter ζ is referred to as the ‘roughness exponent’ in the physics literature. In the case of the Gaussian free field $\zeta = 0$ the inner product is the Dirichlet inner product $\int_{\mathbb{R}^2} |\nabla f|^2 dx$.

An important feature of the ensembles is scale-invariance. In the special case $\zeta = 0$ (and dimension two), the Dirichlet inner product $\int_M |\nabla f|_g^2 dA_g$ is conformally invariant, i.e. invariant under conformal changes $g \rightarrow e^u g$ of the Riemannian metric. When $\zeta \neq 0$ this is not the case, but it is assumed in [KH] that the fluctuations of the random Gaussian surface with height function h are invariant under the rescaling $h(r) \rightarrow c^{-\zeta} h(cr)$ for any $c > 1$. The authors of [KH] then make a number of conjectures concerning the distribution of contour lengths, which we interpret as conjectures concerning $\mathbb{E}d\mu_L$. First, they consider contours (i.e. level sets) through a fixed point x_0 and measure its length with the re-scaled arc-length measure λds , i.e. with arclength s in units of $\frac{1}{\lambda}$. They define the fractal dimension of a nodal line component as the dimension D so that $s \sim R^D$ where R is the radius of the nodal component (i.e. half the diameter). They define $P(s)$ as the probability density that the contour through x_0 has length s . The principal claim is that

$P(s) \sim s^{-\tau-1}$ satisfies a power law for some exponent τ ([KH] (4)). They also defines the distribution of loop lengths $\tilde{P}(s) \sim P(s)/s$ as the probability density that a random component has length s . We interpret their $\tilde{P}(s)$ as the density of $\lim_{\lambda \rightarrow \infty} \mathbb{E} d\mu_L$ with respect to ds on \mathbb{R} . We thus interpret their conjecture as saying that a unique weak* limit of this family of measures exists and has a density relative to ds with a power law decay as above.

The claims are based in part on scaling properties of the contour ensemble. They also are based in part on the expectation that, at ‘criticality’, the key percolation ‘exponents’ of power laws are universal and therefore should be the same for the discrete and continuum percolation theories (see e.g. [IsiK]). In [KH], the authors suggest that when a certain roughness exponent ζ vanishes (the critical models), the continuum problem is related to the four-state Potts model. The q -state Potts model is an Ising type spin model on a lattice where each spin can take one of q values. It is known to be related to connectivity and percolation problems on a graph [Bax, Wu].

They compute D, τ by relating both to another exponent x_1 defined by a “contour correlation function” $\mathcal{G}_1(r)$, which measures the probability that points at $x, x+r$ lie on the same contour loop. They claim that $\mathcal{G}_1(r) \sim |r|^{-2x_1}$. They claim that $D(3-\tau) = 2 - 2x_1$ and $D(\tau-1) = 2 - \zeta$. As a result, $D = 2 - x_1 - \zeta/2, \tau - 1 = \frac{2-\zeta}{2-x_1-\zeta/2}$. From the mapping to the four-state Potts model, they conclude that $x_1 = \frac{1}{2}$.

There exist rigorous results in [SS, Mi] relating discretizations of the Gaussian free field (rather than high frequency truncations) to the percolation models. They prove that in various senses, the zero set of the discrete Gaussian free field tends to an SLE_4 curve. It does not seem to be known at present if zero sets of the high frequency truncation of the Gaussian free field also tends in the same sense to an SLE_4 curve. Note that the SLE curves are interfaces and that one must select one component of the zero set that should tend to an SLE curve. There might exist modified conjectures regarding CLE curves.

To determine the ‘critical exponents’ in continuum percolation, it is tempting to find a way to ‘map’ the continuum problem to a discrete percolation model. A geometric ‘map’ from a random wave to a graph is to associate to the random function its Morse-Smale decomposition, known in the physics literature as the “Morse skeleton” (see §2.6 or [Web] for an extensive exposition). As discussed in [Wei], and as illustrated in Figure 10 of [Isi], the Morse complex of the random function plays the role of the lattice in lattice percolation theory.

12.2. Fixed frequency ensembles. We now consider Riemannian random waves of asymptotically fixed frequency λ , such as random spherical harmonics of fixed degree or Euclidean random plane waves of fixed eigenvalue. In this case the weight is a delta function at the frequency. One would expect different behavior in the level sets since only one frequency is involved rather than the superposition of waves of all frequencies $\leq \lambda$.

A recent exposition in the specific setting of random Euclidean eigenfunctions of fixed frequency is given by [EGJS]. The level sets play the role of open paths. Super-level sets are compared to clusters of sites in a critical 2D percolation model, such as bond percolation on a lattice. Each site may of the percolation model may be visualized as a disc of area $\frac{2\pi^2}{\lambda^2}$, i.e. as a small component. The nodal domains may be thought of as connected clusters of a number n such discs. Since nodal domains are connected components in which the eigenfunction is either positive + or negative -, they are analogous to clusters of ‘open’ or ‘closed’ vertices.

The main conjectures in this fixed frequency ensemble are due to E. Bogolomny and C. Schmidt [BS]. They conjecture that the continuum percolation problem should belong to the same universality class as the Potts model at a certain critical point (where q is related to a certain temperature) for a large rectangular lattice and that the nodal lines in the $\lambda \rightarrow \infty$ limit tend to SLE_6 curves. This is similar to the predictions of [KH] but for a very different ensemble where there is little apriori reason to expect conformal invariance in the limit. There are parallel conjectures in [BBCF] for zero-vorticity isolines in 2D turbulence, which are also conjectured to tend to SLE_6 curves. They remark (page 127) that this limit is surprising since continuous percolation models assume short-correlations in the height functions whereas the vorticity field correlations decay only like $r^{-4/3}$. They write, “When the pair correlation function falls off slower than $r^{-3/2}$, the system is not expected generally to belong to the universality class of uncorrelated percolation and to be conformally invariant”. The same remarks apply to the fixed frequency ensemble, where the correlation function is the spectral projection $\Pi_{[\lambda, \lambda+1]}(x, y)$ for a fixed frequency. In this case, the correlations decay quite slowly as $r^{-\frac{1}{2}}$; we refer to [BS2] for this background and also for an argument why the nodal sets should nevertheless resemble conformally invariant SLE_g curves.

If the nodal lines in the fixed frequency model are equivalent to the critical percolation model, then the ‘probability’ of finding a nodal domain of area s should decay like $s^{-\tau}$ where $\tau = \frac{187}{91} > 2$ (see [SA], p. 52 for the percolation theory result). Under some shape assumptions adopted in [EGJS], it is equivalent that the probability of finding clusters consisting of n discs is of order $n^{-\tau}$. For random spherical harmonics, one may ask for the probability that a spherical harmonic of degree N has size n . For a fixed (M, g) with simple eigenvalues, this notion of probability from percolation theory does not make sense, but we might assume that the number of nodal components is of order λ^2 and ask what proportion of the nodal components has size 1. To obtain a percolating nodal line, one would need a cluster with $n = \lambda$ sites, and thus the proportion of such nodal components to the total number would be of order $\lambda^{-\tau}$. Thus, if there are $C\lambda^2$ total components, the number of such components would be around $\lambda^{2-\tau} = \lambda^{-\frac{5}{91}} < 1$, so the model seems to predict that such macroscopic

nodal lines are quite rare. It also predicts that the ‘vast majority’ of nodal components are close to the minimal size, which does not seem so evident from the computer graphics.

References

- [AP] J. C. Alvarez Paiva and E. Fernandes, Gelfand transforms and Crofton formulas. *Selecta Math. (N.S.)* 13 (2007), no. 3, 369–390.
- [AP2] J. C. Alvarez Paiva and G. Berck, What is wrong with the Hausdorff measure in Finsler spaces. *Adv. Math.* 204 (2006), no. 2, 647–663.
- [Ar] S. Ariturk, Lower bounds for nodal sets of Dirichlet and Neumann eigenfunctions, to appear in *Comm. Math. Phys.* (arXiv:1110.6885).
- [Ba] L. Bakri, Critical set of eigenfunctions of the Laplacian, arXiv:1008.1699.
- [Bae] C. Bär, On nodal sets for Dirac and Laplace operators. *Comm. Math. Phys.* 188 (1997), no. 3, 709–721.
- [Bax] R. J. Baxter, Potts model at the critical temperature, *Journal of Physics C: Solid State Physics* 6 (1973), L445.
- [BBCF] D. Bernard, G. Boffetta, A. Celani, and G. Falkovich, Conformal invariance in two-dimensional turbulence, *nature. physics* Vol. 2 (2002), p. 134.
- [Ber] M. V. Berry, Regular and irregular semiclassical wavefunctions. *J. Phys. A* 10 (1977), no. 12, 2083–2091.
- [Bers] L. Bers, Local behavior of solutions of general linear elliptic equations. *Comm. Pure Appl. Math.* 8 (1955), 473–496.
- [BGS] G. Blum, S. Gnutzmann and U. Smilansky, Nodal domain statistics: A Criterion for quantum chaos, *Phys. Rev. Lett.* 88, 114101 (2002).
- [BDS] E. Bogomolny, R. Dubertrand, and C. Schmit, SLE description of the nodal lines of random wavefunctions. *J. Phys. A* 40 (2007), no. 3, 381–395.
- [BS] E. Bogomolny and C. Schmit, Percolation model for nodal domains of chaotic wave functions, *Phys. Rev. Letters* 88 (18) (2002), 114102–114102–4.
- [BS2] E. Bogomolny and C. Schmit, Random wavefunctions and percolation. *J. Phys. A* 40 (2007), no. 47, 14033–14043.
- [Br] J. Brüning, Über Knoten von Eigenfunktionen des Laplace-Beltrami Operators”, *Math. Z.* 158 (1978), 15–21.
- [BSZ1] P. Bleher, B. Shiffman, and S. Zelditch, Universality and scaling of zeros on symplectic manifolds. *Random matrix models and their applications*, 31–69, Math. Sci. Res. Inst. Publ., 40, Cambridge Univ. Press, Cambridge, 2001.
- [BSZ2] P. Bleher, B. Shiffman and S. Zelditch, Universality and scaling of correlations between zeros on complex manifolds, *Invent. Math.* 142 (2000), no. 2, 351–395. <http://xxx.lanl.gov/abs/math-ph/9904020>.
- [Bourg] J. Bourgain, *Geodesic restrictions and L^p -estimates for eigenfunctions of Riemannian surfaces*, Linear and complex analysis, 27–35, Amer. Math. Soc. Transl. Ser. 2, 226, Amer. Math. Soc., Providence, RI, 2009.
- [BZ] J. Bourgain and Z. Rudnick, On the nodal sets of toral eigenfunctions. *Invent. Math.* 185 (2011), no. 1, 199–23.
- [Bou] L. Boutet de Monvel, Convergence dans le domaine complexe des séries de fonctions propres. *C. R. Acad. Sci. Paris Sér. A-B* 287 (1978), no. 13, A855–A856.
- [BGT] N. Burq, P. Gérard, and N. Tzvetkov, Restrictions of the Laplace-Beltrami eigenfunctions to submanifolds. *Duke Math. J.* 138 (2007), no. 3, 445–486
- [Bu] N. Burq, Quantum ergodicity of boundary values of eigenfunctions: A control theory approach, to appear in *Canadian Math. Bull.* (math.AP/0301349).

- [Ch1] S. Y. Cheng, Eigenfunctions and eigenvalues of Laplacian. Differential geometry (Proc. Sympos. Pure Math., Vol. XXVII, Stanford Univ., Stanford, Calif., 1973), Part 2, pp. 185–193. Amer. Math. Soc., Providence, R.I., 1975.
- [Ch2] S. Y. Cheng, Eigenfunctions and nodal sets. Comment. Math. Helv. 51 (1976), no. 1, 43–55.
- [CTZ] H. Christianson, J. A. Toth and S. Zelditch, Quantum ergodic restriction for Cauchy Data: Interior QUE and restricted QUE (arXiv:1205.0286).
- [CM] T. H. Colding and W. P. Minicozzi II, Lower bounds for nodal sets of eigenfunctions. Comm. Math. Phys. 306 (2011), no. 3, 777 – 784.
- [CV] Y.Colin de Verdière, Ergodicité et fonctions propres du Laplacien, Comm.Math.Phys. 102 (1985), 497-502.
- [C] R. Cooper, The extremal values of Legendre polynomials and of certain related functions. Proc. Cambridge Philos. Soc. 46, (1950). 549–55.
- [Dong] R-T. Dong, Nodal sets of eigenfunctions on Riemann surfaces. J. Differential Geom. 36 (1992), no. 2, 493–506.
- [DF] H. Donnelly and C. Fefferman, Nodal sets of eigenfunctions on Riemannian manifolds, Invent. Math. 93 (1988), 161–183.
- [DF2] H. Donnelly and C. Fefferman, Nodal sets of eigenfunctions: Riemannian manifolds with boundary. Analysis, et cetera, 251–262, Academic Press, Boston, MA, 1990.
- [DF3] H. Donnelly and C. Fefferman, Growth and geometry of eigenfunctions of the Laplacian. Analysis and partial differential equations, 635–655, Lecture Notes in Pure and Appl. Math., 122, Dekker, New York, 1990.
- [DF4] H. Donnelly and C. Fefferman, Nodal sets for eigenfunctions of the Laplacian on surfaces. J. Amer. Math. Soc. 3 (1990), no. 2, 333–353.
- [DSZ] M. R. Douglas, B. Shiffman, and S. Zelditch, Critical points and supersymmetric vacua. II. Asymptotics and extremal metrics. J. Differential Geom. 72 (2006), no. 3, 381—427.
- [DZ] S. Dyatlov, and M. Zworski, Quantum ergodicity for restrictions to hypersurfaces (arXiv:1204.0284).
- [EK] Y. Egorov and V. Kondratiev, *On spectral theory of elliptic operators*. Operator Theory: Advances and Applications, 89. Birkhäuser Verlag, Basel, 1996.
- [EGJS] Y. Elon, S. Gnutzmann, C. Joas, and U. Smilansky, Geometric characterization of nodal domains: the area-to-perimeter ratio. J. Phys. A 40 (2007), no. 11, 2689–2707.
- [EJN] A. Eremenko, D. Jakobson and N. Nadirashvili, On nodal sets and nodal domains on S^2 and R^2 . Festival Yves Colin de Verdire. Ann. Inst. Fourier (Grenoble) 57 (2007), no. 7, 2345–2360.
- [Fed] H. Federer, *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153 Springer-Verlag New York Inc., New York 1969.
- [FGS] G. Foltin, S. Gnutzmann, and U. Smilansky, The morphology of nodal lines—random waves versus percolation. J. Phys. A 37 (2004), no. 47, 11363–11371.
- [GaL] N. Garofalo and F. H. Lin, Monotonicity properties of variational integrals, A_p weights and unique continuation. Indiana Univ. Math. J. 35 (1986), no. 2, 245–268.
- [GaL2] ———, Unique continuation for elliptic operators: a geometric-variational approach. Comm. Pure Appl. Math. 40 (1987), no. 3, 347–366
- [GS] I. M. Gelfand and M. Smirnov, Lagrangians satisfying Crofton formulas, Radon transforms, and nonlocal differentials. Adv. Math. 109 (1994), 188–227.
- [GL] P.Gérard and E.Leichtnam, Ergodic properties of eigenfunctions for the Dirichlet problem, Duke Math J. 71 (1993), 559-607.
- [Gi] V. M. Gichev, A Note on the Common Zeros of Laplace Beltrami Eigenfunctions. Ann. Global Anal. Geome. 26, 201–208 (2004).

- [GLS] F. Golse, E. Leichtnam, and M. Stenzel, Intrinsic microlocal analysis and inversion formulae for the heat equation on compact real-analytic Riemannian manifolds. *Ann. Sci. École Norm. Sup.* (4) 29 (1996), no. 6, 669–736.
- [GSj] A. Grigis and J. Sjöstrand, *Microlocal analysis for differential operators*, London Math. Soc. Lecture Notes 196 (1994).
- [GS1] V. Guillemin and M. Stenzel, Grauert tubes and the homogeneous Monge-Ampère equation. *J. Differential Geom.* 34 (1991), no. 2, 561–570.
- [GS2] ———, Grauert tubes and the homogeneous Monge-Ampère equation. II. *J. Differential Geom.* 35 (1992), no. 3, 627–641.
- [H2] Q. Han, Nodal sets of harmonic functions, *Pure and Applied Mathematics Quarterly* 3 (3) (2007), 647–688.
- [HHL] Q. Han, R. Hardt, and F. H. Lin, Geometric measure of singular sets of elliptic equations. *Comm. Pure Appl. Math.* 51 (1998), no. 11–12, 1425–1443.
- [H] Q. Han and F.H. Lin *Nodal sets of solutions of elliptic differential equations*, book in preparation (online at <http://www.nd.edu/~qhan/>).
- [HL] X. Han and G. Lu, A geometric covering lemma and nodal sets of eigenfunctions. (English summary) *Math. Res. Lett.* 18 (2011), no. 2, 337–352.
- [HHON] R. Hardt, M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof and N. Nadirashvili, Critical sets of solutions to elliptic equations. *J. Differential Geom.* 51 (1999), no. 2, 359–373.
- [HaS] R. Hardt and L. Simon, Nodal sets for solutions of elliptic equations. *J. Differential Geom.* 30 (1989), no. 2, 505–522.
- [HZ] A. Hassell and S. Zelditch, Quantum ergodicity of boundary values of eigenfunctions. *Comm. Math. Phys.* 248 (2004), no. 1, 119–168.
- [Hel] S. Helgason, *Topics in harmonic analysis on homogeneous spaces*. Progress in Mathematics, 13. Birkhäuser, Boston, Mass., 1981.
- [HEJ] E. J. Heller, Gallery (Quantum random waves), <http://www.ericjhellergallery.com/>.
- [He] H. Hezari, Complex zeros of eigenfunctions of 1D Schrödinger operators. *Int. Math. Res. Not. IMRN* 2008, no. 3, Art. ID rnm148.
- [HS] H. Hezari and C. D. Sogge, A natural lower bound for the size of nodal sets, *Analysis and PDE* vol. 5 (2012), 1133–1137 (arXiv:1107.3440).
- [HW] H. Hezari and Z. Wang, Lower bounds for volumes of nodal sets: an improvement of a result of Sogge-Zelditch, to appear arXiv:1107.0092.
- [HC] D. Hilbert and R. Courant, *Methods of mathematical physics*, Vol. I and Vol. II: Interscience Publishers (John Wiley & Sons), New York-London 1962.
- [HoI-IV] L. Hörmander, *Theory of Linear Partial Differential Operators I-IV*, Springer-Verlag, New York (1985).
- [Hu] R. Hu, L^p norm estimates of eigenfunctions restricted to submanifolds. *Forum Math.* 21 (2009), no. 6, 1021 – 1052.
- [Isi] M. B. Isichenko, Percolation, statistical topography, and transport in random media. *Rev. Modern Phys.* 64 (1992), no. 4, 961–1043.
- [IsiK] M. B. Isichenko and J. Kalda, Statistical topography. I. Fractal dimension of coastlines and number-area rule for islands. *J. Nonlinear Sci.* 1 (1991), no. 3, 255–277
- [JN] D. Jakobson and N. Nadirashvili, Eigenfunctions with few critical points. *J. Differential Geom.* 53 (1999), no. 1, 177–182.
- [JN2] ———, Quasi-symmetry of L^p norms of eigenfunctions. *Comm. Anal. Geom.* 10 (2002), no. 2, 397–408.
- [JM] D. Jakobson and D. Mangoubi, Tubular Neighborhoods of Nodal Sets and Diophantine Approximation, *Amer. J. Math.* 131 (2009), no. 4, 1109–1135 (arXiv:0707.4045).

- [JL] D. Jerison and G. Lebeau, Nodal sets of sums of eigenfunctions. Harmonic analysis and partial differential equations (Chicago, IL, 1996), 223–239, Chicago Lectures in Math., Univ. Chicago Press, Chicago, IL, 1999.
- [JJ] J. Jung, Zeros of eigenfunctions on hyperbolic surfaces lying on a curve, to appear in JEMS (arXiv: 1108.2335).
- [KH] J. Kondev and C. L. Henley, Geometrical exponents of contour loops on random Gaussian surfaces, Phys. Rev. Lett. 74 (1995), 4580 - 4583.
- [KHS] J. Kondev, C. L. Henley, and D.G. Salinas, Nonlinear measures for characterizing rough surface morphologies. Phys. Rev. E, 61 (2000), 104-125.
- [Kua] I. Kukavica, Nodal volumes for eigenfunctions of analytic regular elliptic problems. J. Anal. Math. 67 (1995), 269–280.
- [Ku] ———, Quantitative uniqueness for second-order elliptic operators. Duke Math. J. 91 (1998), no. 2, 225–240.
- [LS1] L. Lempert and R. Szöke, Global solutions of the homogeneous complex Monge-Ampère equation and complex structures on the tangent bundle of Riemannian manifolds. Math. Ann. 290 (1991), no. 4, 689–712.
- [LS2] ———, The tangent bundle of an almost complex manifold, Canad. Math. Bull. 44 (2001), no. 1, 70–79.
- [Lew] H. Lewy, On the minimum number of domains in which the nodal lines of spherical harmonics divide the sphere. Comm. Partial Differential Equations 2 (1977), no. 12, 1233-1244.
- [Ley] J. Leydold, On the number of nodal domains of spherical harmonics. Topology 35 (1996), no. 2, 301–321.
- [Lin] F.H. Lin, Nodal sets of solutions of elliptic and parabolic equations. Comm. Pure Appl. Math. 44 (1991), no. 3, 287–308.
- [Man] D. Mangoubi, A Remark on Recent Lower Bounds for Nodal Sets, Comm. Partial Differential Equations 36 (2011), no. 12, 2208–2212 (arXiv:1010.4579.)
- [Man2] D. Mangoubi, The Volume of a Local Nodal Domain, J. Topol. Anal. 2 (2010), no. 2, 259–275 (arXiv:0806.3327).
- [Man3] D. Mangoubi, On the inner radius of a nodal domain. Canad. Math. Bull. 51 (2008), no. 2, 249-260.
- [Me] A. D. Melas, On the nodal line of the second eigenfunction of the Laplacian in $\mathbf{R}^2\mathbf{R}^2$. J. Differential Geom. 35 (1992), no. 1, 255–263.
- [Mi] J. Miller, Universality for SLE(4), arXiv:1010.1356.
- [NJT] N. Nadirashvili, D. Jakobson, and J.A. Toth, Geometric properties of eigenfunctions. (Russian) Uspekhi Mat. Nauk 56 (2001), no. 6(342), 67–88; translation in Russian Math. Surveys 56 (2001), no. 6, 1085–1105
- [NPS] F. Nazarov, L. Polterovich and M. Sodin, Sign and area in nodal geometry of Laplace eigenfunctions. Amer. J. Math. 127 (2005), no. 4, 879–910.
- [NS] F. Nazarov and M. Sodin, On the Number of Nodal Domains of Random Spherical Harmonics. Amer. J. Math. 131 (2009), no. 5, 1337-1357 (arXiv:0706.2409).
- [Neu] J. Neuheisel, “Asymptotic distribution of nodal sets on spheres,” PhD Thesis, Johns Hopkins University, Baltimore, MD 2000, 1994, <http://mathnt.mat.jhu.edu/mathnew/Thesis/joshuaneuheisel.pdf>.
- [Nic] L. I. Nicolaescu, Critical sets of random smooth functions on compact manifolds (arXiv:1008.5085).
- [P] A. Pleijel, Remarks on Courant’s nodal line theorem, Comm. Pure Appl. Math., 9, 543-550 (1956).
- [Po] I. Polterovich, Pleijel’s nodal domain theorem for free membranes, Proc. Amer. Math. Soc. 137 (2009), no. 3, 1021–1024 (arXiv:0805.1553).
- [PS] L. Polterovich and M. Sodin, Nodal inequalities on surfaces. Math. Proc. Cambridge Philos. Soc. 143 (2007), no. 2, 459–467 (arXiv:math/0604493).

- [R] J. Ralston, Gaussian beams and the propagation of singularities. *Studies in partial differential equations*, 206–248, MAA Stud. Math., 23, Math. Assoc. America, Washington, DC, 1982.
- [Reu] M. Reuter. Hierarchical Shape Segmentation and Registration via Topological Features of Laplace-Beltrami Eigenfunctions. International Journal of Computer Vision 89 (2), pp. 287-308, 2010.
- [Reu2] M. Reuter, *Laplace Spectra for Shape Recognition*, Books on Demand (2006).
- [Ri] G. Rivière, Letter to the author (2012).
- [SY] R. Schoen and S. T. Yau, Lectures on differential geometry. . Conference Proceedings and Lecture Notes in Geometry and Topology, I. International Press, Cambridge, MA, 1994.
- [Sh.1] A.I.Schnirelman, Ergodic properties of eigenfunctions, Usp.Math.Nauk. 29/6 (1974), 181-182.
- [SS] U. Smilansky and H.-J. Stöckmann, Nodal Patterns in Physics and Mathematics, The European Physical Journal Special Topics Vol. 145 (June 2007).
- [Sog] C. D. Sogge, Concerning the L^p norm of spectral clusters for second-order elliptic operators on compact manifolds, J. Funct. Anal. 77 (1988), 123–138.
- [Sog2] C. D. Sogge, *Kakeya-Nikodym averages and L^p -norms of eigenfunctions*, Tohoku Math. J. (2) 63 (2011), no. 4, 519-538 (arXiv:0907.4827).
- [Sog3] C. D. Sogge: *Fourier integrals in classical analysis*, Cambridge Tracts in Mathematics, 105, Cambridge University Press, Cambridge, 1993.
- [STZ] C.D. Sogge, J. A. Toth and S. Zelditch, About the blowup of quasimodes on Riemannian manifolds. J. Geom. Anal. 21 (2011), no. 1, 150-173.
- [SoZ] C. D. Sogge and S. Zelditch, Lower bounds on the hypersurface measure of nodal sets, Math. Research Letters 18 (2011), 27-39 (arXiv:1009.3573).
- [SoZa] C. D. Sogge and S. Zelditch, Lower bounds on the Hausdorff measure of nodal sets II, to appear in Math. Res. Lett. (arXiv:1208.2045).
- [SoZ2] C.D. Sogge and S. Zelditch, On eigenfunction restriction estimates and L^4 -bounds for compact surfaces with nonpositive curvature (arXiv:1108.2726).
- [SoZ3] C.D. Sogge and S. Zelditch, Concerning the L^4 norms of typical eigenfunctions on compact surfaces, Recent Developments in Geometry and Analysis, 23 (2013), 407–423, International Press of Boston, Boston (arXiv:1011.0215).
- [SA] D. Stauffer and A. Aharony, *Introduction to Percolation theory*, Taylor and Francis, London (1994).
- [Sz] G. Szegő, Inequalities for the zeros of Legendre polynomials and related functions. Trans. Amer. Math. Soc. 39 (1936), no. 1, 1–17.
- [Sz2] G. Szegő. On the relative extrema of Legendre polynomials. Boll. Un. Mat. Ital. (3) 5, (1950). 120–121.
- [Taa] D. Tataru, On the regularity of boundary traces for the wave equation, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) 26 (1998), 185 – 206.
- [TZ] J. A. Toth and S. Zelditch, Counting Nodal Lines Which Touch the Boundary of an Analytic Domain, Jour. Diff. Geom. 81 (2009), 649- 686 (arXiv:0710.0101).
- [TZ2] J. A. Toth and S. Zelditch, Quantum ergodic restriction theorems, I: interior hypersurfaces in analytic domains,, Ann. H. Poincaré 13, Issue 4 (2012), Page 599-670 (arXiv:1005.1636).
- [TZ3] J. A. Toth and S. Zelditch, Quantum ergodic restriction theorems, II: manifolds without boundary, to appear in GAFA (arXiv:1104.4531).
- [U] K. Uhlenbeck, Generic properties of eigenfunctions. Amer. J. Math. 98 (1976), no. 4, 1059–1078.
- [Web] J. Weber, The Morse-Witten complex via dynamical systems. Expo. Math. 24 (2006), no. 2, 127-159.
- [Wei] A. Weinrib, Percolation threshold of a two-dimensional continuum system. Phys. Rev. B (3) 26 (1982), no. 3, 1352-1361.

- [Wig] I. Wigman, On the distribution of the nodal sets of random spherical harmonics. *J. Math. Phys.* 50 (2009), no. 1, 013521.
- [Wu] F. Y. Wu, Percolation and the Potts model. *J. Statist. Phys.* 18 (1978), no. 2, 115–123.
- [Y1] S.T. Yau, Survey on partial differential equations in differential geometry. *Seminar on Differential Geometry*, pp. 3–71, Ann. of Math. Stud., 102, Princeton Univ. Press, Princeton, N.J., 1982.
- [Y2] ———, Open problems in geometry. *Differential geometry: partial differential equations on manifolds* (Los Angeles, CA, 1990), 1–28, Proc. Sympos. Pure Math., 54, Part 1, Amer. Math. Soc., Providence, RI, 1993.
- [Y3] ———, A note on the distribution of critical points of eigenfunctions, *Tsing Hua Lectures in Geometry and Analysis* 315–317, Internat. Press, 1997.
- [Z1] S. Zelditch, Uniform distribution of eigenfunctions on compact hyperbolic surfaces. *Duke Math. J.* 55 (1987), no. 4, 919–941.
- [Z2] S. Zelditch,, Complex zeros of real ergodic eigenfunctions. *Invent. Math.* 167 (2007), no. 2, 419–443.
- [Z3] S. Zelditch, Ergodicity and intersections of nodal sets and eigenfunctions on real analytic surfaces (arXiv:1210.0834).
- [Z4] S. Zelditch, Real and complex zeros of Riemannian random waves. *Spectral analysis in geometry and number theory*, 321–342, Contemp. Math., 484, Amer. Math. Soc., Providence, RI, 2009.
- [Z5] S. Zelditch, Local and global analysis of eigenfunctions on Riemannian manifolds. *Handbook of geometric analysis. No. 1*, 545–658, Adv. Lect. Math. (ALM), 7, Int. Press, Somerville, MA, 2008.
- [Z6] S. Zelditch, New Results in Mathematics of Quantum Chaos, Current Developments in Mathematics 2009, p. 115–202 (arXiv:0911.4312).
- [Z7] S. Zelditch, Kuznecov sum formulae and Szegő limit formulae on manifolds, *Comm. PDE* **17** (1&2) (1992), 221–260.
- [Z8] S. Zelditch, Pluri-potential theory on Grauert tubes of real analytic Riemannian manifolds, I. Spectral geometry, 299 – 339, *Proc. Sympos. Pure Math.*, 84, Amer. Math. Soc., Providence, RI, 2012.
- [Z9] S. Zelditch, Complex zeros of quantum integrable eigenfunctions, (in preparation).
- [ZZw] S.Zelditch and M.Zworski, Ergodicity of eigenfunctions for ergodic billiards, *Comm.Math. Phys.* 175 (1996), 673-682.
- [Zw] M. Zworski, *Semiclassical analysis*, Graduate Studies in Mathematics, 138. American Mathematical Society, Providence, RI, 2012.

DEPARTMENT OF MATHEMATICS, NORTHWESTERN UNIVERSITY, EVANSTON, IL 60208, USA

E-mail address: zelditch@math.northwestern.edu