CHAPTER 3: MINIMAL SURFACES

DANNY CALEGARI

ABSTRACT. These are notes on minimal surfaces, which are being transformed into Chapter 3 of a book on 3-Manifolds. The emphasis is on the classical theory and its connection to complex analysis, and the topological applications to 3-manifold geometry/topology. These notes follow a course given at the University of Chicago in Spring 2014.

CONTENTS

1. MINIMAL SURFACES IN EUCLIDEAN SPACE

In this section we describe the classical theory of minimal surfaces in Euclidean spaces, especially in dimension 3. We emphasize throughout the connections to complex analysis. The local theory of minimal surfaces in Riemannian manifolds is well approximated by the theory of minimal surfaces in Euclidean space, so the theory we develop in this section has applications more broadly.

1.1. **Graphs.** A smooth surface in $\mathbb{R}^3$ can be expressed locally as the graph of a real-valued function defined on a domain in $\mathbb{R}^2$. For historical reasons, such graphs are sometimes referred to as *nonparametric* surfaces.

Fix a compact domain $\Omega \subset \mathbb{R}^2$ with boundary, and a smooth function $f : \Omega \to \mathbb{R}$. Denote by $\Gamma(f)$ the *graph* of f in $\mathbb{R}^3$; i.e.

$$\Gamma(f) = \{(x, y, f(x, y)) \in \mathbb{R}^3 \text{ such that } (x, y) \in \Omega\}$$

Ignoring the last coordinate defines a projection from $\Gamma(f)$ to Ω, which is a diffeomorphism.

An infinitesimal square in Ω with edges of length dx, dy is in the image of an infinitesimal parallelogram in $\Gamma(u)$ with edges $(dx, 0, f_x dx)$ and $(0, dy, f_y dy)$ so

$$\text{area}(\Gamma(f)) = \int_\Omega |(1, 0, f_x) \times (0, 1, f_y)| dxdy = \int_\Omega \sqrt{1 + |\text{grad}(f)|^2} dxdy$$

Date: February 10, 2019.

Note that in this formula we are thinking of $\mathrm{grad}(f)$ as a vector field on Ω in the usual way.

If $g : \Omega \to \mathbb{R}$ is smooth and vanishes on $\partial\Omega$ we obtain a 1-parameter family of functions $f(t) := f + tg$ and a 1-parameter family of graphs $\Gamma(t) := \Gamma(f(t))$. The hypothesis that g vanishes on $\partial\Omega$ ensures that these graphs all have the same boundary. We are interested in how the area of $\Gamma(t)$ varies as a function of t.

Since $\mathrm{grad}(f + tg) = \mathrm{grad}(f) + t\,\mathrm{grad}(g)$ we can compute

$$\begin{aligned}\frac{d}{dt}\Big|_{t=0}\mathrm{area}(\Gamma(t)) &= \int_\Omega \frac{d}{dt}\Big|_{t=0}\sqrt{1+|\mathrm{grad}(f)+t\,\mathrm{grad}(g)|^2}dxdy \\ &= \int_\Omega \frac{\langle \mathrm{grad}(f), \mathrm{grad}(g)\rangle}{\sqrt{1+|\mathrm{grad}(f)|^2}}dxdy\end{aligned}$$

Integrating by parts, and using the vanishing of g on $\partial\Omega$, this equals

$$\int_\Omega -g\,\mathrm{div}\left(\frac{\mathrm{grad}(f)}{\sqrt{1+|\mathrm{grad}(f)|^2}}\right)dxdy$$

(formally, this is the observation that $-\mathrm{div}$ is an *adjoint* for grad).

Thus: f is a *critical point* for area (among all smooth 1-parameter variations with support in the interior of Ω) if and only if f satisfies the *minimal surface equation* in divergence form:

$$\mathrm{div}\left(\frac{\mathrm{grad}(f)}{\sqrt{1+|\mathrm{grad}(f)|^2}}\right) = 0$$

In this case we also say that Γ is a *minimal surface*.

Expanding the minimal surface equation, and multiplying through by the factor $(1 + |\mathrm{grad}(f)|^2)^{3/2}$ we obtain the equation

$$(1+f_y^2)f_{xx} + (1+f_x^2)f_{yy} - 2f_xf_yf_{xy} = 0$$

This is a *second order quasi-linear elliptic PDE*. We explain these terms:

(1) The order of a PDE is the degree of the highest derivatives appearing in the equation. In this case the order is 2.
(2) A PDE is quasi-linear if it is linear in the derivatives of highest order, with coefficients that depend on the independent variables and derivatives of strictly smaller order. In this case the coefficients of the highest derivatives are $(1 + f_y^2)$, $(1 + f_x^2)$ and $-2f_xf_y$ which depend only on the independent variables (the domain variables x and y) and derivatives of order at most 1.
(3) A PDE is elliptic if the discriminant is negative; here the discriminant is the discriminant of the homogeneous polynomial obtained by replacing the highest order derivatives by monomials. In this case since the PDE is second order, the discriminant is the polynomial $B^2 - 4AC$, which is equal to

$$4f_x^2f_y^2 - 4(1+f_y^2)(1+f_x^2) = -4(1+f_x^2+f_y^2) < 0$$

Solutions of elliptic PDE are as smooth as the coefficients allow, within the interior of the domain. Thus minimal surfaces in $\mathbb{R}^3$ are real analytic in the interior.

If f is constant to first order (i.e. if $f_x, f_y \sim \epsilon$) then this equation approximates $f_{xx}+f_{yy} = 0$; i.e. $-\Delta f = 0$, the Dirichlet equation, whose solutions are harmonic functions. Thus, the minimal surface equation is a nonlinear generalization of the Dirichlet equation, and functions with minimal graphs are generalizations of harmonic functions.

In particular, the *qualitative* structure of such functions — their singularities, how they behave in families, etc. — very closely resembles the theory of harmonic functions.

1.2. **Second fundamental form and mean curvature.** A *parametric* surface Σ in $\mathbb{R}^n$ is just a smooth map from some domain Ω in $\mathbb{R}^2$ to $\mathbb{R}^n$. Let's denote the map $x : \Omega \to \mathbb{R}^n$, so $\Sigma = x(\Omega)$.

The *second fundamental form* II is the perpendicular part of the Hessian of x. In terms of co-ordinates, let e_1, e_2 be vector fields on Ω linearly independent at p. Then write

$$\mathrm{II}_p(e_i, e_j) := e_i(e_j(x))(p)^\perp$$

i.e the component of the vector $e_i(e_j(x))(p) \in \mathbb{R}^n$ perpendicular to Σ at $x(p)$.

It turns out that II_p is a symmetric quadratic form on $T_p\Sigma$. To see this, observe first that II_p depends only on the value of e_i at p. Furthermore, since $[e_i, e_j](x) = dx([e_i, e_j])$ is tangent to Σ, it follows that II_p is symmetric in its arguments, and therefore it depends only on the value of e_j at p.

The *mean curvature* H is the trace of the second fundamental form; i.e.

$$H(p) = \sum \mathrm{II}_p(e_i, e_i)$$

where e_i runs over an orthonormal basis for T_p.

Now let N be a unit normal vector field on Σ. By abuse of notation we write e_i for $dx(e_i)$ and think of it as a vector field on Σ. Then pointwise,

$$\langle H, N\rangle = \sum \langle e_i(e_i), N\rangle = -\sum \langle e_i, e_i(N)\rangle$$

where we use the fact that N is everywhere perpendicular to each e_i. An infinitesimal unit square in T_p spanned by e_1, e_2 flows under the normal vector field N to an infinitesimal parallelogram, and the derivative of its area under the flow is exactly $\sum\langle e_i, e_i(N)\rangle$. Thus near any point p where H is nonzero we can change the area to first order by a normal variation supported near p, and we see that Σ is critical for area if and only if the mean curvature H vanishes identically. This observation is due to Meusnier.

For Σ a surface in $\mathbb{R}^3$ the unit normal field N is determined by Σ (up to sign). We extend N smoothly to a unit length vector field in a neighborhood of Σ. Along Σ the vectors e_1, e_2, N form an orthonormal basis, and the expression $\sum\langle e_i, e_i(N)\rangle$ is equal to $\mathrm{div}(N)$ which in particular does not depend on the choice of extension.

Specializing to the case of a graph, the vector field

$$N := \frac{(-f_x, -f_y, 1)}{\sqrt{1 + |\mathrm{grad}(f)|^2}}$$

is nothing but the unit normal field to $\Gamma(f)$. We can extend N to a vector field on $\Omega \times \mathbb{R}$ by translating it parallel to the z axis. We obtain the identity:

$$-\mathrm{div}_{\mathbb{R}^3}(N) = \mathrm{div}_{\mathbb{R}^2}\left(\frac{\mathrm{grad}(f)}{\sqrt{1+|\mathrm{grad}(f)|^2}}\right)$$

where the subscript gives the domain of the vector field where each of the two divergences are computed. Thus Γ is minimal if and only if the divergence of N vanishes.

1.3. **Conformal parameterization.** We consider a parametric surface $x : \Omega \to \mathbb{R}^n$. Let's denote the coordinates on $\mathbb{R}^2$ by u and v, and the coordinates on $\mathbb{R}^n$ by $x_1, \cdots, x_n$. The Jacobian is the matrix with column vectors $\frac{\partial x}{\partial u}$ and $\frac{\partial x}{\partial v}$, and where this matrix has rank 2 the image is smooth, and the parameterization is locally a diffeomorphism to its image. The metric on $\mathbb{R}^n$ makes the image into a Riemannian surface, and every Riemannian metric on a surface is locally conformally equivalent to a flat metric. Thus after precomposing x with a diffeomorphism, we may assume the parameterization is *conformal* (one also says that we have chosen *isothermal* coordinates). This means exactly that there is a smooth nowhere vanishing function λ on Ω so that

$$\left|\frac{\partial x}{\partial u}\right|^2 = \left|\frac{\partial x}{\partial v}\right|^2 = \lambda^2 \quad \text{and} \quad \frac{\partial x}{\partial u} \cdot \frac{\partial x}{\partial v} = 0$$

We can identify Ω with a domain in $\mathbb{C}$ with complex coordinate $\zeta := u + iv$, and for each coordinate x_j define

$$\phi_j := 2\frac{\partial x_j}{\partial \zeta} = \frac{\partial x_j}{\partial u} - i\frac{\partial x_j}{\partial v}$$

Then we have the following lemma:

Lemma 1.1 (Conformal parameterization)**.** *Let $\Omega \subset \mathbb{C}$ be a domain with coordinate ζ, and $x : \Omega \to \mathbb{R}^n$ a smooth immersion. Then x is a conformal parameterization of its image (with conformal structure inherited from $\mathbb{R}^n$) if and only if $\sum \phi_j^2 = 0$, where $\phi_j = 2\partial x_j/\partial\zeta$. Furthermore, functions ϕ_j as above with $\sum \phi_j^2 = 0$ define an immersion if and only if $\sum |\phi_j|^2 > 0$.*

Proof. By definition,

$$\sum \phi_j^2 = \left(\sum_j \left(\frac{\partial x_j}{\partial u}\right)^2 - \left(\frac{\partial x_j}{\partial v}\right)^2\right) - i\left(\sum_j \frac{\partial x_j}{\partial u}\frac{\partial x_j}{\partial v}\right)$$

whose real and imaginary parts vanish identically if and only if the parameterization is conformal, where it is an immersion. Furthermore,

$$\sum |\phi_j|^2 = \lambda^2$$

so the map is an immersion everywhere if and only if $\sum |\phi_j|^2 > 0$. □

If the parameterization $x : \Omega \to \mathbb{R}^n$ is conformal, we can consider the orthonormal basis

$$e_1 = \lambda^{-1}\frac{\partial}{\partial u} \text{ and } e_2 = \lambda^{-1}\frac{\partial}{\partial v}$$

Differentiating the defining equations for x to be conformal, we obtain

$$\frac{\partial^2 x}{\partial u^2} \cdot \frac{\partial x}{\partial u} = \frac{\partial^2 x}{\partial u \partial v} \cdot \frac{\partial x}{\partial v} = -\frac{\partial^2 x}{\partial v^2} \cdot \frac{\partial x}{\partial u}$$

and therefore Δx is perpendicular to Σ. Thus the mean curvature H — i.e. the trace of the second fundamental form — is the vector

$$H = \lambda^{-2} \left(\frac{\partial^2 x}{\partial u^2} + \frac{\partial^2 x}{\partial v^2} \right) = -\lambda^2 \Delta x$$

Thus we obtain the following elegant characterization of minimal surfaces in terms of conformal parameterizations:

Lemma 1.2 (Harmonic coordinates)**.** *Let $\Omega \subset \mathbb{C}$ be a domain with coordinate ζ, and $x : \Omega \to \mathbb{R}^n$ a conformal parameterization of a smooth surface. Then the image is minimal if and only if the coordinate functions x_j are harmonic on Ω; equivalently, if and only if the functions $\phi_j := 2\frac{\partial x_j}{\partial \zeta}$ are holomorphic functions of ζ.*

Proof. All that must be checked is the fact that the equation $-\Delta x_j = 0$ is equivalent to the Cauchy–Riemann equations for ϕ_j:

$$-\Delta x_j = \frac{\partial}{\partial u}\left(\frac{\partial x_j}{\partial u} - i\frac{\partial x_j}{\partial v} \right) + i\frac{\partial}{\partial v}\left(\frac{\partial x_j}{\partial u} - i\frac{\partial x_j}{\partial v} \right) = 2\frac{\partial \phi_j}{\partial \bar{\zeta}}$$

□

Corollary 1.3 (Convex hull)**.** *Every compact minimal surface Σ in $\mathbb{R}^n$ lies in the convex hull of its boundary.*

Proof. Every linear function on $\mathbb{R}^n$ is harmonic on Σ. Therefore by the maximum principle, if the boundary lies in a given half-space of $\mathbb{R}^n$, so does Σ. □

A holomorphic reparameterization of Ω transforms the coordinate ζ and the functions ϕ_j, but keeps fixed the 1-form $\phi_j d\zeta$. Combining this observation with the two lemmas, we obtain the following proposition, characterizing minimal surfaces in $\mathbb{R}^n$ parameterized by arbitrary Riemann surfaces:

Proposition 1.4. *Every minimal surface in $\mathbb{R}^n$ is obtained from some Riemann surface Ω together with a family of n complex valued 1-forms ϕ_j satisfying the following conditions:*

(1) **(conformal):** $\sum \phi_j^2 = 0$*;*
(2) **(minimal):** *the ϕ_j are holomorphic;*
(3) **(regular):** $\sum |\phi_j|^2 > 0$*; and*
(4) **(period):** *the integral of ϕ_j over any closed loop on Ω is purely imaginary.*

The map $x : \Omega \to \mathbb{R}^n$ may then be obtained uniquely up to a translation by integration:

$$x_j = \operatorname{Re}\left(\int_0^\zeta \phi_j \right) + c_j$$

Proof. All that remains is to observe that the period condition is both necessary and sufficient to let us recover the coordinates x_j by integrating the real part of the ϕ_j. □

If the ϕ_j are holomorphic and not identically zero, then $\sum |\phi_j|^2$ can equal zero only at isolated points in Ω. Near such points the map from Ω to its image is *branched*. We say that a surface is a *generalized minimal surface* if it is parameterized by some Ω as in Proposition 1.4, omitting the condition of regularity.

1.4. **Conjugate families.** Let Ω be a Riemann surface, and ϕ_j a collection of n holomorphic 1-forms satisfying $\sum \phi_j^2 = 0$. Integrating the ϕ_j along loops in Ω determines a period map $H_1(\Omega;\mathbb{Z}) \to \mathbb{C}^n$, and an abelian cover $\bar{\Omega}$ corresponding to the kernel of the period map.

Then we obtain a further integration map $\Phi : \bar{\Omega} \to \mathbb{C}^n$ whose coordinates z_j are given by $z_j = \int_0^\zeta \phi_j$. The standard (complex) orthogonal quadratic form has the value $\sum z_j^2$ on a vector z with coordinates z_j; by Proposition 1.4 the image $\Phi(\bar{\Omega})$ is *isotropic* for this orthogonal form. Consequently we obtain a family of minimal surfaces in $\mathbb{R}^n$ parameterized by the action of the *complex affine group* $\mathbb{C}^n \rtimes (\mathbb{C}^* \times \mathrm{O}(n,\mathbb{C}))$ where the first factor acts on $\mathbb{C}^n$ by translation, and the second factor acts linearly.

The $\mathbb{C}^n$ action just projects to translation of the minimal surface in $\mathbb{R}^n$, and $\mathbb{R}^* \times \mathrm{O}(n,\mathbb{R})$ just acts by scaling and rotation; so this subgroup acts by ambient similarities of $\mathbb{R}^n$. The action of $S^1 \subset \mathbb{C}^*$ is more interesting; the family of minimal surfaces related by this action are said to be a *conjugate family*. At the level of 1-forms this action is a phase shift $\phi_j \to e^{i\theta}\phi_j$.

Lemma 1.5. *Let $x(\theta) : \Omega \to \Sigma_\theta$ be a conjugate family of generalized minimal surfaces in $\mathbb{R}^n$; i.e. their coordinates are given by integration*

$$x_j(\theta)(p) = \mathrm{Re} \int_0^p e^{i\theta}\phi_j$$

for some fixed family of 1-forms ϕ_j on Ω with $\sum \phi_j^2 = 0$. Then for any θ the composition $x(0) \circ x(\theta)^{-1} : \Sigma_\theta \to \Sigma_0$ is a local isometry, and each fixed point $p \in \Omega$ traces out an ellipse $x(\cdot)(p) : S^1 \to \mathbb{R}^n$.

Proof. If we write $\phi_j = a_j + ib_j$ then $\sum a_j^2 = \sum b_j^2$ and $\sum a_j b_j = 0$; i.e. the vectors a and b are perpendicular with the same length, and this length is the length of $dx(0)(\partial_u)$ in $Tx(0)(\Omega)$. But the length of $dx(\theta)(\partial_u)$ in $Tx(\theta)(\Omega)$ is just $|\cos(\theta)a + \sin(\theta)b| = |a| = |b|$ for all θ, so $x(0) \circ x(\theta)^{-1} : \Sigma_\theta \to \Sigma_0$ is an isometry as claimed.

The second claim is immediate:

$$x_j(\theta)(p) = \cos(\theta)\mathrm{Re}\left(\int_0^p \phi_j\right) + \sin(\theta)\mathrm{Re}\left(\int_0^p i\phi_j\right)$$

□

Example 1.6 (catenoid and helicoid). If we write $\zeta := u + iv$ then

$$\phi_1 = -\sin(\zeta), \quad \phi_2 = \cos(\zeta), \quad \phi_3 = -i$$

satisfies the conditions of Proposition 1.4. The surface obtained by integrating the real part of the ϕ_j is the *catenoid*, with the parameterization

$$x = \cosh(v)\cos(u); \quad y = \cosh(v)\sin(u); \quad z = v$$

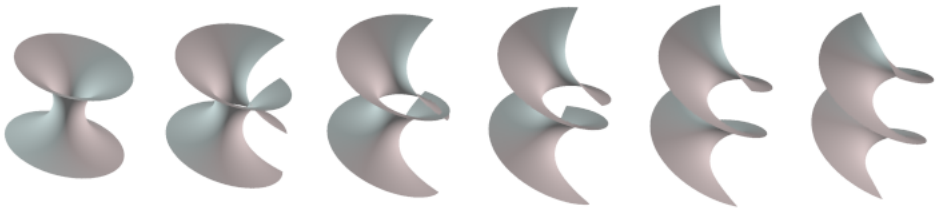

FIGURE 1. A conjugate family of minimal surfaces interpolating between the catenoid and the helicoid

Multipying the ϕ_j by i and integrating gives the *helicoid*, with the parameterization

$$x = \sinh(v)\sin(u); \quad y = -\sinh(v)\cos(u); \quad z = u$$

Interpolating between these two values gives an *isometric* deformation of one surface into the other.

1.5. **Weierstrass–Enneper parameterization.** Let ϕ_j be three holomorphic functions on a domain Ω satisfying the conditions of Proposition 1.4. Using the condition $\sum \phi_j^2 = 0$ we can eliminate one of the functions. If we write $f = \phi_1 - i\phi_2$ and $g = \phi_3/f$ then

$$\phi_1 = f(1-g^2)/2; \quad \phi_2 = if(1+g^2)/2; \quad \phi_3 = fg$$

Conversely, any pair of functions f, g where f is holomorphic, and g is meromorphic so that f has a zero of order at least $2k$ wherever g has a pole of order k, define a (generalized) minimal surface. This normalization is the so-called *Weierstrass–Enneper parameterization.*

This parameterization is particularly nice because of its connection to the Gauss map. For Σ an oriented surface in $\mathbb{R}^3$, the *Gauss map* $N : \Sigma \to S^2$ sends each point to its unit normal. If we take $\Sigma = x(\Omega)$ then the unit normal field is given by $N = (x_u \times x_v)/|x_u \times x_v|$. Now, by definition $x_u = \operatorname{Re}\phi$ and $x_v = -\operatorname{Im}\phi$ so $x_u \times x_v = (\operatorname{Im}\phi \times \overline{\phi})/2$. Likewise, since the parameterization is conformal, $|x_u \times x_v| = |x_u|\,|x_v| = |\phi|^2/2$. It follows that $N = (\operatorname{Im}\phi \times \overline{\phi})/|\phi|^2$.

Stereographic projection from the north pole sends S^2 to $\mathbb{C}$ by

$$\mathrm{St} : (a, b, c) \to (a + ib)/(1 - c)$$

Thus the composition $G := \mathrm{St} \circ N \circ x : \Omega \to \mathbb{C}$ is given by the formula

$$G(u,v) = \frac{2\operatorname{Im}\phi_2\overline{\phi}_3 + 2i\operatorname{Im}\phi_3\overline{\phi}_1}{|\phi|^2 - 2\operatorname{Im}\phi_1\overline{\phi}_2} = \frac{\phi_3}{\phi_1 - i\phi_2} = g$$

Corollary 1.7 (Gauss map conformal)**.** *If Σ is a minimal surface in $\mathbb{R}^3$, the Gauss map $N : \Sigma \to S^2$ is conformal.*

1.6. **Finite total curvature.** Now let us suppose $x : \Omega \to \mathbb{R}^3$ with image Σ is a *complete* minimal surface in $\mathbb{R}^3$. It is natural to impose finiteness conditions on x and Ω. Since Σ is minimal, we have $K \le 0$ everywhere.

We say Σ has *finite total curvature* if

$$\int_\Sigma |K| d\mathrm{area} = \int_\Sigma -K d\mathrm{area} < \infty$$

The condition of finite total curvature imposes very strong constraints on Ω and x.

Lemma 1.8. *Suppose Σ has finite total curvature. Then Ω is of finite type; i.e. it is homeomorphic to a surface of finite genus with finitely many points removed.*

Proof. Fix a point $p \in \Sigma$ and let $B(t) \subset \Sigma$ be the set of q with distance at most t from p as measured intrinsically in Σ. Let $\ell(t) = \text{length}(\partial B(t))$. Then by Gauss-Bonnet, $\ell'(t) = 2\pi\chi(B(t)) - \int_{B(t)} K d\text{area}$. If Ω is not of finite type, then $\chi(B(t)) \to -\infty$. But then $\ell(t) < 0$ for large t which is absurd. □

Lemma 1.9. *Suppose Σ has finite total curvature. Then Ω has parabolic ends; i.e. Ω is conformally equivalent to a closed surface of finite genus with finitely many points removed.*

Proof. We have already seen that Ω has finitely many cylindrical ends. For each of these ends $\ell''(t) \to 0$ so $\ell'(t) \to \text{constant} \geq 0$. We estimate the modulus using the method of extremal length. For simplicity suppose there is one annular component, namely $B(t) - B(s)$ for some fixed s.

Suppose $\lim \ell'(t) = \text{constant} > 0$. Let Γ be the system of curves from $\partial(B(s))$ to $\partial(B(t))$. Using $\rho(t) = 1/\ell(t)$ gives

$$L_\rho(\Gamma) := \inf_{\gamma \in \Gamma} \int_\gamma \rho \, d\text{length} \sim \log(t) \text{ and } A_\rho := \int_{B(t)-B(s)} \rho^2 d\text{area} \sim \log(t)$$

The modulus of the annulus $B(t) - B(s)$ is the supremum $\sup_\rho L_\rho(\Gamma)^2/A_\rho$ over all ρ; hence as $t \to \infty$ the modulus goes to infinity, proving that the end is parabolic.

The case $\lim \ell'(t) = 0$ is proved similarly using $\rho = 1$. □

Theorem 1.10 (Osserman)**.** *Suppose Σ has finite total curvature. Then Ω is conformally equivalent to a closed surface $\overline{\Omega}$ of finite genus, with finitely many points removed. Consequently the Gauss map extends to a holomorphic map $g : \overline{\Omega} \to \mathbb{CP}^1$, and the total curvature of Σ is an integer multiple of 4π.*

Proof. The curvature K is the pullback of the area form on the unit sphere under the Gauss map. Since Σ has finite total curvature, g does not have an essential singularity at each of the punctures of Ω, and therefore it extends holomorphically over each puncture, and $g : \overline{\Omega} \to \mathbb{CP}^1$ is a branched cover. Thus the total curvature is the degree of g (which is an integer) times the area of the unit sphere. □

2. First variation formula

In this section we derive the first variation formula, which characterizes those submanifolds that are critical for the volume functional among compactly supported variations: the critical submanifolds are precisely those with *vanishing mean curvature*. We also develop some elements of the theory of Riemannian geometry in coordinate-free language (as far as possible).

2.1. **Grad, div, curl.**

There are many natural differential operators on functions, vector fields, and differential forms on $\mathbb{R}^3$ which make use of many implicit isomorphisms between various spaces. This can make it confusing to figure out the analogs of these operators on Riemannian 3-manifolds (or Riemannian manifolds of other dimensions). In this section

we recall the co-ordinate free definitions of some of these operators, which generalize the familiar case of Euclidean $\mathbb{R}^3$.

2.1.1. *Gradient.* On a smooth manifold M, there is a natural differential operator d, which operates on functions and forms of all orders. By definition, if f is a smooth function, df is the 1-form such that for all vector fields X, there is an identity

$$df(X) = Xf$$

Where f is nondegenerate, the kernel of df is a hyperplane field, which is simply the tangent space to the level set of f through the given point.

If M is a Riemannian manifold with inner product denoted $\langle \cdot, \cdot \rangle$, there are natural isomorphisms between 1-forms and vector fields (called the *sharp* and the *flat* isomorphisms) defined by

$$\langle \alpha^\sharp, X \rangle = \alpha(X)$$

for a 1-form α and a vector field X, and

$$X^\flat(Y) = \langle X, Y \rangle$$

for vector fields X and Y. Using these isomorphisms, for any function f on a Riemannian manifold the *gradient*, denoted $\mathrm{grad}(f)$ or sometimes ∇f, is the vector field defined by the formula

$$\mathrm{grad}(f) := (df)^\sharp$$

In other words, $\mathrm{grad}(f)$ is the unique vector field such that, for any other vector field X, we have

$$\langle \mathrm{grad}(f), X \rangle = df(X)$$

For any 1-form α, the vector field $\alpha^\sharp$ is perpendicular to the hyperplane field $\ker(\alpha)$; thus $\mathrm{grad}(f)$ is perpendicular to the level sets of f, and points in the direction in which f is increasing, with size proportional to the rate at which f increases. The zeros of the gradient are the critical points of f; for instance, $\mathrm{grad}(f)$ vanishes at the minimum and the maximum of f.

2.1.2. *Divergence.* On an oriented Riemannian n-manifold there is a volume form $d\mathrm{vol}$, and a Hodge star $*$ taking k-forms to $(n-k)$-forms, satisfying

$$\alpha \wedge *\alpha = \|\alpha\|^2 d\mathrm{vol}$$

This does not define $*\alpha$ uniquely; we must further add that $*\alpha$ is orthogonal (with respect to the pointwise inner product on $(n-k)$-forms) to the subspace of forms β with $\alpha \wedge \beta = 0$. In other words, $*\alpha$ is the form of smallest (pointwise) norm subject to $\alpha \wedge *\alpha = \|\alpha\|^2 \omega$.

With this notation, $*d\mathrm{vol}$ is the constant function 1; conversely for any smooth function f, we have $*f = f d\mathrm{vol}$. If X is a smooth vector field, then (at least locally and for short time) flow along X determines a 1-parameter family of diffeomorphisms $\phi(X)_t$. The Lie derivative of a (contravariant) tensor field α, denoted $\mathcal{L}_X \alpha$, is by definition

$$\mathcal{L}_X \alpha = \frac{d}{dt}\Big|_{t=0} \phi(X)_t^* \alpha$$

For forms α it satisfies the *Cartan formula*

$$\mathcal{L}_X \alpha = \iota_X d\alpha + d\iota_X \alpha$$

where $\iota_X \beta$ is the interior product of X with a form β (i.e. the form obtained by contracting β with X). The *divergence* of a vector field X, denoted $\mathrm{div}(X)$ (or sometimes $-\nabla^* X$ or $\nabla \cdot X$), is the function defined by the formula

$$\mathrm{div}(X) = *(\mathcal{L}_X d\mathrm{vol})$$

By Cartan's formula, $\mathcal{L}_X d\mathrm{vol} = d\iota_X d\mathrm{vol}$, because $d\mathrm{vol}$ is closed. Furthermore, for any vector field X we have the identity

$$\iota_X d\mathrm{vol} = *(X^\flat)$$

which can be verified pointwise, since both sides depend only on the values at a point. Thus we obtain the equivalent formula

$$\mathrm{div}(X) = * d * (X^\flat)$$

The operator $- * d *$ on 1-forms is often denoted d^*; likewise we sometimes we denote the operator $-\mathrm{div}(X)$ by ∇^*, on the grounds that $\nabla^*(X) = d^*(X^\flat)$. If X is a vector field and f is a compactly supported function (which holds automatically for instance if M is closed) then

$$\int_M \langle X, \nabla f \rangle d\mathrm{vol} = \int_M df(X) d\mathrm{vol} = \int_M df \wedge \iota_X d\mathrm{vol}$$

Now,

$$d(f \iota_X d\mathrm{vol}) = df \wedge \iota_X d\mathrm{vol} + f d\iota_X d\mathrm{vol} = df \wedge \iota_X d\mathrm{vol} + f \mathrm{div}(X) d\mathrm{vol}$$

But if f is compactly supported, $\int_M d(f \iota_X d\mathrm{vol}) = 0$ and we deduce that

$$\int_M \langle X, \nabla f \rangle d\mathrm{vol} = \int_M -f \mathrm{div}(X) d\mathrm{vol}$$

So that $-\mathrm{div}$ (i.e. ∇^*) is a "formal" adjoint to grad (i.e. ∇), justifying the notation.

The divergence of a vector field vanishes where $\mathcal{L}_X d\mathrm{vol} = 0$; i.e. where the flow generated by X preserves the volume.

2.1.3. *Laplacian.* If f is a function, we can first apply the gradient and then the divergence to obtain another function; this composition (or rather its negative) is the *Laplacian*, and is denoted Δ. In other words,

$$\Delta f = -\mathrm{div}\ \mathrm{grad}(f) = d^* df$$

Thus formally, Δ is a non-negative self-adjoint operator, so that we expect to be able to decompose the functions on M into a direct sum of eigenspaces with non-negative eigenvalues. Indeed, if M is closed, then $L^2(M)$ decomposes into an (infinite) direct sum of the eigenspaces of Δ, which are finite dimensional, and whose eigenvalues are discrete and non-negative. A function f with $\Delta f = 0$ is *harmonic*; on a closed manifold, the only harmonic functions are constants.

On Euclidean space, harmonic functions satisfy the *mean value property*: the value of f at each point is equal to the average of f over any round ball (or sphere) centered at f. In general, the value of a harmonic function f at each point is a weighted average of the values on a ball centered at that point; in particular, a harmonic function on a compact subset of any Riemannian manifold attains its maximum (or minimum) only at points on the frontier.

2.1.4. *Curl.* Now we specialize to an oriented Riemannian 3-manifold. The operator $*d$ takes 1-forms to 1-forms. Using the sharp and flat operators, it induces a map from vector fields to vector fields. The *curl* of a vector field X, denoted $\operatorname{curl}(X)$ (or sometimes $\nabla \times X$), is the vector field defined by the formula

$$\operatorname{curl}(X) = (*d(X^\flat))^\sharp$$

Notice that this satisfies the identities

$$\operatorname{div}\operatorname{curl}(X) = *\,d\,*\,d(X^\flat) = 0$$

(because $*^2 = \pm 1$ and $d^2 = 0$) and

$$\operatorname{curl}\operatorname{grad}(f) = (*ddf)^\sharp = 0$$

On a Riemannian manifold of arbitrary dimension, it still makes sense to talk about the 2-form $d(X^\flat)$, which we can identify (using the metric) with a section of the bundle of skew-symmetric endomorphisms of the tangent space. Identifying skew-symmetric endomorphisms with elements of the Lie algebra of the orthogonal group, we can define $\operatorname{curl}(X)$ in general to be the field of infinitesimal rotations corresponding to $d(X^\flat)$. On a 3-manifold, the vector field $\operatorname{curl}(X)$ points in the direction of the axis of this infinitesimal rotation, and its magnitude is the size of the rotation.

2.1.5. *Flows and parallel transport.* On a Riemannian manifold, there is a unique torsion-free connection for which the metric tensor is parallel, namely the *Levi-Civita connection*, usually denoted ∇ (when we mix the connection with gradient in formulae, we will denote the gradient by grad). If X is a vector field on M, we can generate a 1-parameter family of automorphisms of the tangent space at each point by flowing by X, then parallel transporting back along the flowlines of X by the connection. The derivative of this family of automorphisms is a 1-parameter family of *endomorphisms* of the tangent space at each point, denoted A_X. In terms of differential operators, $A_X := \mathcal{L}_X - \nabla_X$, and one can verify that $A_X Y$ is tensorial in Y. Thus, X determines a section A_X of the bundle $\operatorname{End}(TM)$.

On an oriented Riemannian manifold, the vector space $\operatorname{End}(T_pM) = T_p^*M \otimes T_pM$ is an $\mathfrak{o}(n)$-module in an obvious way, and it makes sense to decompose an endomorphism into components, corresponding to the irreducible $\mathfrak{o}(n)$-factors. Each endomorphism decomposes into an antisymmetric and a symmetric part, and the symmetric part decomposes further into the trace, and the trace-free part.

In this language,

(1) the divergence of X is the negative of the trace of A_X. As a formula, this is given pointwise by

$$\operatorname{div}(X)(p) = \text{trace of } V \to \nabla_V X \text{ on } T_pM$$

(2) the curl of X is the skew-symmetric part of A_X; and
(3) the strain of X (a measure of the infinitesimal failure of flow by X to be conformal) is the trace-free symmetric part of A_X.

2.2. **First variation formula.** Let M be a Riemannian n-manifold, and let Ω be a smooth bounded domain in $\mathbb{R}^k$. Let $f : \Omega \to M$ be a smooth immersion with image Σ, and let $F : \Omega \times (-\epsilon, \epsilon) \to M$ be a one-parameter variation supported in the interior of Ω. Let t denote the coordinate on $(-\epsilon, \epsilon)$, and let $T = dF(\partial_t)$, which we think of (at least locally) as a vector field on M generating a family of diffeomorphisms $\phi(T)_t$ (really we should think of T as a vector field along F; i.e. a section of the pullback of TM to Ω by F^*). Under this flow, Σ evolves to $\Sigma(t)$, and at each time is equal to $F(\Omega, t)$.

The flow T determines an endomorphism field A_T along f. This endomorphism decomposes into a skew-symmetric part (which rotates the tangent space to Σ but preserves volume) and a symmetric part. The derivative at $t = 0$ of the area of an infinitesimal plane tangent to $T\Sigma(t)$ is the negative of the trace of A_T restricted to $T\Sigma$. As in § 2.1.5 this can be expressed as the trace of $V \to \nabla_V T$ *restricted to* $T\Sigma$. If e_i is an orthonormal frame field along $T\Sigma$, we obtain a formula

$$\frac{d}{dt}\Big|_{t=0} \text{volume}(\Sigma(t)) = \int_\Sigma \sum_i \langle \nabla_{e_i} T, e_i \rangle d\text{vol}$$

The integrand on the right hand side of this formula is sometimes abbreviated to $\text{div}_\Sigma(T)$.

If we decompose T into a normal and tangential part as $T = T^\perp + T^\top$, this can be expressed as

$$\frac{d}{dt}\Big|_{t=0} \text{volume}(\Sigma(t)) = \int_\Sigma \text{div}(T^\top) + \sum_i \langle \nabla_{e_i} T^\perp, e_i \rangle d\text{vol}$$

where $\text{div}(T^\top)$ means the divergence in the usual sense on Σ of the vector field $T^\top$, thought of as a vector field on Σ. But

$$\langle \nabla_{e_i} T^\perp, e_i \rangle = e_i \langle T^\perp, e_i \rangle - \langle T^\perp, \nabla_{e_i} e_i \rangle = -\langle T^\perp, \nabla_{e_i} e_i \rangle$$

by the metric property of the Levi-Civita connection, and the fact that $T^\perp$ is orthogonal to e_i. Similarly, $\int_\Sigma \text{div}(T^\top) d\text{vol} = 0$ by Stokes' formula, because $T^\top$ is compactly supported in the interior. The sum $H := \sum_i \nabla^\perp_{e_i} e_i$ (where $\nabla^\perp$ denotes the normal part of the covariant derivative) is the *mean curvature vector*, which is the trace of the second fundamental form, and is normal to Σ by definition; thus $\langle T^\perp, \sum_i \nabla_{e_i} e_i \rangle = \langle T, H \rangle$.

Putting this together, we obtain the *first variation formula*:

Proposition 2.1 (First Variation Formula)**.** *Let Σ be a compact immersed submanifold of a Riemannian manifold, and let T be a compactly supported vector field on M along Σ. If $\Sigma(t)$ is a 1-parameter family of immersed manifolds tangent at $t = 0$ to the variation T, then*

$$\frac{d}{dt}\Big|_{t=0} \text{volume}(\Sigma(t)) = \int_\Sigma -\langle T, H \rangle d\text{vol}$$

Consequently, Σ is a critical point for volume among compactly supported variations if and only if the mean curvature vector H vanishes identically.

This motivates the following definition:

Definition 2.2. A submanifold is said to be *minimal* if its mean curvature vector H vanishes identically.

The terminology "minimal" is widely established, but the reader should be warned that minimal submanifolds (in this sense) are not always even local minima for volume.

Example 2.3 (Totally geodesic submanifolds). If Σ is 1-dimensional, the mean curvature is just the geodesic curvature, so a 1-manifold is minimal if and only if it is a geodesic.

A totally geodesic manifold has vanishing second fundamental form, and therefore vanishing mean curvature, and is minimal. An equatorial sphere in S^n is an example which is minimal, but not a local minimum for volume.

Warning 2.4. It is more usual to define the mean curvature of a k-dimensional submanifold Σ to be equal to $\frac{1}{k}\sum_i \nabla^{\perp}_{e_i} e_i$. With this definition, the mean curvature is the *average* of the principal curvatures of Σ — i.e. the eigenvalues of the second fundamental form — rather than their sum. But the convention we adhere to seems to be common in the minimal surfaces literature; see e.g. [3] p. 5 or [22] p. 5.

2.3. **Calibrations.** Now suppose that $\mathcal{F}$ is a codimension 1 foliation of a manifold M. Locally we can coorient $\mathcal{F}$, and let X denote the unit normal vector field to $\mathcal{F}$. For each leaf λ of the foliation we can consider a compactly supported normal variation fX, and suppose that $\lambda(t)$ is a 1-parameter family tangent at $t = 0$ to fX. Then

$$\frac{d}{dt}\Big|_{t=0} \text{volume}(\lambda(t)) = \int_\lambda \text{div}_\lambda(fX) d\text{vol}$$

and because X is normal, this simplifies (by the Leibniz rule for covariant differentiation) to

$$\frac{d}{dt}\Big|_{t=0} \text{volume}(\lambda(t)) = \int_\lambda f\text{div}_\lambda(X) d\text{vol}$$

Because X is a normal vector field of *unit* length, it satisfies $\text{div}(X) = \text{div}_\lambda(X)$. Thus we obtain the lemma:

Lemma 2.5 (Normal field volume preserving)**.** *A cooriented codimension 1 foliation $\mathcal{F}$ has minimal leaves if and only if the unit normal vector field X is volume-preserving.*

Now, suppose $\mathcal{F}$ is a foliation with minimal leaves, and let X be the unit normal vector field. It follows that the $(n-1)$-form $\omega := \iota_X d\text{vol}$ is *closed*. On the other hand, it evidently satisfies the following two properties:

(1) the restriction of ω to $T\mathcal{F}$ is equal to the volume form on leaves; and
(2) the restriction of ω to any $(n-1)$ plane not tangent to $T\mathcal{F}$ has norm strictly less than the volume form on that plane.

Such a form ω is said to *calibrate* the foliation.

Lemma 2.6. *Let $\mathcal{F}$ be a foliation with minimal leaves. Then leaves of $\mathcal{F}$ are* globally *area minimizing, among all compactly supported variations in the same relative homology class.*

Proof. Let λ be a leaf, and let μ be obtained from λ by cutting out some submanifold and replacing it by another homologous submanifold. Then

$$\text{volume}(\mu) \geq \int_\mu \omega = \int_\lambda \omega = \text{volume}(\lambda)$$

where the middle equality follows because ω is closed, and the inequality is strict unless μ is tangent to $\mathcal{F}$. But μ agrees with λ outside a compact part; so in this case $\mu = \lambda$. □

Example 2.7. Let Σ be an immersed minimal $(n-1)$-manifold in $\mathbb{R}^n$. Let $p \in \Sigma$ be the center of a round disk D in the tangent space $T_p\Sigma$. Let C be the cylindrical region obtained by translating D normal to itself. Then $C \cap \Sigma$ is a graph over D, and we can foliate C by parallel copies of $C \cap \Sigma$, translated in the normal direction. Thus there exists a calibration ω defined on C, and we see that $C \cap \Sigma$ is least volume among all surfaces in C obtained by a compactly supported variation. But C is convex, so the nearest point projection to C is volume non-increasing. We deduce that any immersed minimal $(n-1)$-manifold in $\mathbb{R}^n$ is *locally* volume minimizing. This should be compared to the fact that geodesics in any Riemannian manifold are locally distance minimizing.

2.4. **Gauss equations.** Recall that the second fundamental form on a submanifold Σ of a Riemannian manifold M is the symmetric vector-valued bilinear form

$$\mathrm{II}(X, Y) := \nabla^{\perp}_X Y$$

and H is the trace of II; i.e. $H = \sum_i \mathrm{II}(e_i, e_i)$ where e_i is an orthonormal basis for $T\Sigma$.

The *Gauss equation* for X, Y vector fields on Σ is the equation

$$K_\Sigma(X,Y)|X \wedge Y|^2 = K_M(X,Y)|X \wedge Y|^2 + \langle \mathrm{II}(X,X), \mathrm{II}(Y,Y)\rangle - |\mathrm{II}(X,Y)|^2$$

where $|X \wedge Y|^2 := |X|^2|Y|^2 - \langle X, Y\rangle^2$ is the square of the area of the parallelogram spanned by X and Y, and

$$K(X,Y) := \frac{\langle R(X,Y)Y, X\rangle}{|X \wedge Y|^2}$$

is the sectional curvature in the plane spanned by X and Y, where $R(X,Y)Z := \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$ is the curvature tensor. The subscripts K_Σ and K_M denote sectional curvature as measured in Σ and in M respectively; the difference is that in the latter case curvature is measured using the Levi-Civita connection ∇ on M, whereas in the former it is measured using the Levi-Civita connection on Σ, which is $\nabla^\top := \nabla - \nabla^\perp$.

If Σ is a 2-dimensional surface in a 3-manifold M, then we can take X and Y to be the directions of principal curvature on Σ (i.e. the unit eigenvectors for the second fundamental form). If we coorient Σ, the unit normal field lets us express II as an $\mathbb{R}$-valued quadratic form, and the principal curvatures are real eigenvalues k_1 and k_2. Since $H = k_1 + k_2$, for Σ a minimal surface we have $k_1 = -k_2$ and

$$K_\Sigma = K_M - \frac{|\mathrm{II}|^2}{2} \tag{2.1}$$

Where $|\mathrm{II}|^2 := \sum_{i,j} |\mathrm{II}(e_i, e_j)|^2$. In other words, a minimal surface in a 3-manifold has curvature pointwise bounded above by the curvature of the ambient manifold. So for example, if M has non-positive curvature, the same is true of Σ.

3. Second variation formula

The first variation formula shows that minimal surfaces are critical for volume, among smooth variations, compactly supported in the interior. To determine the index of these critical surfaces requires the computation of the second variation. In this section, we derive the second variation formula and some of its consequences.

3.1. **Second variation formula.** We specialize to the case that Σ is a hypersurface in a Riemannian manifold. We further restrict attention to variations in the normal direction (this is reasonable, since a small variation of Σ supported in the interior will be transverse to the exponentiated normal bundle). Denote the unit normal vector field along Σ by N, and extend N into a neighborhood along normal geodesics, so that $\nabla_N N = 0$. Let $F : \Omega \times (-\epsilon, \epsilon) \to M$ satisfy $F(\cdot, 0) : \Omega \to \Sigma$, and if t parameterizes the interval $(-\epsilon, \epsilon)$, there is a smooth function f on $\Omega \times (-\epsilon, \epsilon)$ with compact support so that $T := dF(\partial_t) = fN$. Then define $\Sigma(t) := F(\Omega, t)$.

Let e_i be vector fields defined locally on Ω so that $dF(\cdot, 0)(e_i)$ are an orthonormal frame on Σ locally, and extend them to vector fields on $\Omega \times (-\epsilon, \epsilon)$ so that they are constant in the t direction; i.e. they are tangent to $\Omega \times t$ for each fixed t, and satisfy $[e_i, \partial_t] = 0$ for all i. By abuse of notation we denote the pushforward of the e_i by dF also as e_i, and think of them as vector fields on $\Sigma(t)$ for each t.

For each point $p \in \Sigma$ corresponding to a point $q \in \Omega$ the curve $F(q \times (-\epsilon, \epsilon))$ is contained in the normal geodesic to Σ through p, and is parameterized by t. Along this curve we define $g(t)$ to be the matrix whose ij-entry is the function $\langle e_i, e_j \rangle$ (where we take the inner product in M).

The infinitesimal parallelepiped spanned by the e_i at each point in $\Sigma(t)$ has volume $\sqrt{\det(g(t))}$. Projecting along the fibers of the variation, we can push this forward to a density on $\Sigma(0)$ which may be integrated against the volume form to give the volume of $\Sigma(t)$; thus

$$\text{volume}(\Sigma(t)) = \int_{\Sigma(0)} \sqrt{\det(g(t))} d\text{vol}$$

We now compute the second variation of volume. Taking second derivatives, we obtain

$$\frac{d^2}{dt^2}\Big|_{t=0} \text{volume}(\Sigma(t)) = \int_\Sigma \frac{d^2}{dt^2}\Big|_{t=0} \sqrt{\det(g(t))} d\text{vol}$$

Suppose further that Σ is a minimal surface, so that $\det'(g)(0) = \text{tr}(g'(0)) = 0$. Then since $\det(g(0)) = 1$, we have

$$\frac{d^2}{dt^2}\Big|_{t=0} \sqrt{\det(g(t))} = \frac{1}{2} \frac{d^2}{dt^2}\Big|_{t=0} \det(g(t))$$

By expanding $g(t)$ in a Taylor series, we can compute

$$\frac{1}{2} \frac{d^2}{dt^2}\Big|_{t=0} \det(g(t)) = \frac{1}{2} \text{tr}(g''(0)) + \sigma_2(g'(0))$$

where tr means trace, and for a matrix M with eigenvalues κ_i we have $\sigma_2(M) = \sum_{i<j} \kappa_i \kappa_j$.

First we compute $\sigma_2(g'(0))$. By definition, g' is the matrix with entries $T\langle e_i, e_j \rangle = \langle \nabla_T e_i, e_j \rangle + \langle e_i, \nabla_T e_j \rangle$. But

$$\langle \nabla_T e_i, e_j \rangle = \langle \nabla_{e_i} T, e_j \rangle = -\langle T, \nabla_{e_i} e_j \rangle \text{ at } t = 0$$

because T is perpendicular to e_j along $\Sigma(0)$, so $e_i \langle T, e_j \rangle = 0$ at $t = 0$. Note that $\langle N, \nabla_{e_i} e_j \rangle$ is the ij-entry of the second fundamental form II, which is symmetric, and therefore we obtain the formula $g'(0) = -2f\text{II}$. Moreover, $\text{tr}(g'(0)) = 0$ so

$$0 = \text{tr}(g'(0))^2 = |g'(0)|^2 + 2\sigma_2(g'(0))$$

and therefore $\sigma_2(g'(0)) = -2f^2|\mathrm{II}|^2$.

Next we compute $\mathrm{tr}(g''(0))/2$. By definition, $g''/2$ is the matrix with diagonal entries

$$\frac{1}{2}T(T\langle e_i, e_i\rangle) = \langle \nabla_T e_i, \nabla_T e_i\rangle + \langle \nabla_T \nabla_T e_i, e_i\rangle$$

Expanding the first term gives

$$\langle \nabla_T e_i, \nabla_T e_i\rangle = \langle \nabla_{e_i} T, \nabla_{e_i} T\rangle = |e_i(f)|^2\langle N, N\rangle + 2fe_i(f)\langle N, \nabla_{e_i} N\rangle + f^2\langle \nabla_{e_i} N, \nabla_{e_i} N\rangle$$

Since N has unit length, the first term is $|e_i(f)|^2$, and the second term vanishes, because the vector field $\nabla_{e_i} N$ is perpendicular to N. The operator $X \to \nabla_X N$ along $T\Sigma(0)$ is the Weingarten operator, a symmetric bilinear form on $T\Sigma(0)$ whose matrix is $-\mathrm{II}$; thus $\sum_i \langle \nabla_{e_i} N, \nabla_{e_i} N\rangle = |\mathrm{II}|^2$ at $t = 0$, and also at $t = 0$ we have $\sum_i |e_i(f)|^2 = |\mathrm{grad}_{\Sigma(0)}(f)|^2$. Thus

$$\sum_i \langle \nabla_T e_i, \nabla_T e_i\rangle\Big|_{t=0} = |\mathrm{grad}_\Sigma(f)|^2 + f^2|\mathrm{II}|^2$$

Expanding the second term gives

$$\langle \nabla_T \nabla_T e_i, e_i\rangle = \langle \nabla_T \nabla_{e_i} T, e_i\rangle = \langle \nabla_{e_i} \nabla_T T, e_i\rangle - \langle R(e_i, T)T, e_i\rangle$$

Now $\nabla_T T = f\nabla_N fN = fN(f)N + f^2\nabla_N N$. But the integral curves of N are geodesics, so $\nabla_N N = 0$. Thus

$$\langle \nabla_{e_i} \nabla_T T, e_i\rangle = e_i(fN(f))\langle N, e_i\rangle + fN(f)\langle \nabla_{e_i} N, e_i\rangle$$

But $\langle N, e_i\rangle = 0$ at $t = 0$, and $\sum_i \langle \nabla_{e_i} N, e_i\rangle = H = 0$ at $t = 0$, so

$$\sum_i \langle \nabla_T \nabla_T e_i, e_i\rangle\Big|_{t=0} = -f^2 \sum_i \langle R(e_i, N)N, e_i\rangle = -f^2\mathrm{Ric}(N)$$

And therefore $\frac{1}{2}\mathrm{tr}(g''(0)) = |\mathrm{grad}_\Sigma(f)|^2 + f^2|\mathrm{II}|^2 - f^2\mathrm{Ric}(N)$.

Putting this together gives

$$\frac{d^2}{dt^2}\Big|_{t=0} \mathrm{volume}(\Sigma(t)) = \int_\Sigma -f^2|\mathrm{II}|^2 + |\mathrm{grad}_\Sigma(f)|^2 - f^2\mathrm{Ric}(N)d\mathrm{vol}$$

Integrating by parts gives

$$\frac{d^2}{dt^2}\Big|_{t=0} \mathrm{volume}(\Sigma(t)) = \int_\Sigma \langle(\Delta_\Sigma - |\mathrm{II}|^2 - \mathrm{Ric}(N))f, f\rangle d\mathrm{vol}$$

(remember our convention that $\Delta = -\mathrm{div}\,\mathrm{grad}$). If we define the *stability operator* $L := \Delta_\Sigma - |\mathrm{II}|^2 - \mathrm{Ric}(N)$ (also called the *Jacobi operator*), we obtain the *second variation formula*:

Proposition 3.1 (Second Variation Formula)**.** *Let Σ be a compact immersed codimension one submanifold of a Riemannian manifold, and let $T = fN$ where N in the unit normal vector field along Σ, and f is smooth with compact support in the interior of Σ. Suppose that Σ is minimal (i.e. that $H = 0$ identically). If $\Sigma(t)$ is a 1-parameter family of immersed manifolds tangent at $t = 0$ to the variation T, then*

$$\frac{d^2}{dt^2}\Big|_{t=0} \mathrm{volume}(\Sigma(t)) = \int_\Sigma L(f)f dvol$$

where $L := \Delta_\Sigma - |\mathrm{II}|^2 - \mathrm{Ric}(N)$ and $\Delta_\Sigma := -\mathrm{div}_\Sigma\,\mathrm{grad}_\Sigma$ is the Laplacian on Σ.

A critical point for a smooth function on a finite dimensional manifold is usually called stable when the Hessian (i.e. the matrix of second partial derivatives) is positive definite. This ensures that the point is an isolated local minimum for the function. However, in minimal surface theory one says that minimal submanifolds are stable when the second variation is merely non-negative:

Definition 3.2. A minimal submanifold Σ is *stable* if no smooth compactly supported variation can decrease the volume to second order.

Example 3.3. A calibrated surface is locally least area and therefore stable. For example, minimal graphs.

Integrating by parts gives rise to the so-called stability inequality:

Proposition 3.4 (Stability inequality)**.** *If Σ is a stable codimension 1 minimal submanifold of M, then for every Lipschitz function f compactly supported in the interior of Σ, there is an inequality*

$$\int_\Sigma \left(\mathrm{Ric}(N) + |\mathrm{II}|^2\right) f^2 dvol \leq \int_\Sigma |\mathrm{grad}_\Sigma f|^2 dvol$$

3.1.1. *Spectral theory of L.* Stability can also be expressed in spectral terms. The operator L is morally the Hessian at Σ on the space of smooth compactly supported normal variations. Thus, as an operator on functions on Σ, it is linear, second order and self-adjoint on the L^2 completion of $C_0^\infty(\Sigma)$, which we denote $L^2(\Sigma)$.

It is obtained from the second order operator Δ_Σ by adding a 0th order perturbation $-|\mathrm{II}|^2 - \mathrm{Ric}(N)$. The spectrum of Δ_Σ is non-negative and discrete, with finite multiplicity, and $L^2(\Sigma)$ admits an orthogonal decomposition into eigenspaces. The eigenfunctions are as regular as Σ, and therefore as regular as M (since Σ is minimal), so for instance they are real analytic if M is.

When we obtain L from Δ_Σ by perturbation, finitely many eigenvalues might become negative, but the spectrum is still discrete and bounded below, so that the index (i.e. the number of negative eigenvalues, counted with multiplicity) is finite.

Proposition 3.5. *Let Σ be compact, possibly with boundary, with a trivial normal bundle, and let λ be the least eigenvalue for L, and V_λ the associated eigenspace. Then any nonzero $u \in V_\lambda$ cannot change sign. In particular, V_λ is one dimensional.*

Proof. By the Rayleigh-Ritz method, we have

$$\lambda = \inf_f \frac{\langle Lf, f\rangle}{\|f\|^2} = \inf_f \frac{\int_\Sigma |\mathrm{grad} f|^2 - (\mathrm{Ric}(N) + |\mathrm{II}|^2) f^2 d\mathrm{vol}}{\|f\|^2}$$

over all Lipschitz f. The infimum is achieved exactly by eigenfunctions u; but if u achieves the infimum, so does $|u|$ (note that this implies $|u|$ is smooth). Since $|u|$ is non-negative, the Harnack inequality says there is a uniform bound on the logarithmic derivative of $|u|$ away from $\partial\Sigma$, and therefore $|u|$ is never zero, and $|u| = \pm u$.

If $\dim V_\lambda > 1$ we can find eigenfunctions u_1, u_2 for which some nonconstant linear combination changes sign somewhere, contrary to the above. □

Eigenfunctions of L with different eigenvalues are orthogonal; consequently only the eigenfunction of lease eigenvalue does not change sign.

Corollary 3.6. *If Σ is stable with trivial normal bundle, so is any cover $\pi : \tilde{\Sigma} \to \Sigma$.*

Proof. If $\tilde{\Sigma}$ is unstable, there is some compactly supported f with $\langle Lf, f\rangle < 0$, and by Proposition 3.5 we can take $f \geq 0$ everywhere. Define g on Σ by $g(p) = \sum_{q\in\pi^{-1}(p)} f(q)$. Since f is compactly supported, g is finite and compactly supported, and $\langle Lg, g\rangle < 0$ so Σ is unstable. □

3.1.2. *Stability controls area growth.* Colding-Minicozzi [4] Thm. 2.1 derive a simple upper bound on the area growth for a stable minimal surface in a Riemannian 3-manifold.

They consider D an immersed disk in M of intrinsic radius r, and a non-negative operator of the form $\Delta_\Sigma - \nu + \kappa + c_1 K_\Sigma$ where $c_1 > (1 + \kappa r^2)/2$, and then obtain an estimate of the form

$$\frac{\text{area}(D)}{r^2} + \frac{c_2}{2\pi c_1}\int_D \nu\left(1 - \frac{s}{r}\right)^2 ds \leq c_2 \tag{3.1}$$

for $c_2 := 2\pi c_1/(2c_1 - 1 - \kappa r^2)$ (here s is the intrinsic radial coordinate on the disk D).

To apply this to a stable minimal surface, we can take $\nu = \text{Ric}(N)$, $\kappa \geq 2|K_M|$ and $c_1 = 2$, at least for r sufficiently small depending on κ. For $M = \mathbb{R}^3$ we have $\nu = \kappa = 0$ and $c_1 = 2$ works unconditionally. In this form the estimate of Colding-Minicozzi gives:

Proposition 3.7 (Colding-Minicozzi)**.** *Let D be a stable minimal disk in $\mathbb{R}^3$ of radius r in its intrinsic metric. Then*

$$\pi r^2 \leq \text{area}(D) \leq 4/3\,\pi r^2$$

Proof. For any minimal surface in $\mathbb{R}^3$, stable or not, we have $K \leq 0$ pointwise, so Gauss-Bonnet gives $\pi r^2 \leq \text{area}(D)$. The other bound needs stability.

For a stable minimal surface in $\mathbb{R}^3$ the operator $\Delta_\Sigma + 2K_\Sigma$ is non-negative; i.e. for any compactly supported test function f on Σ we have

$$0 \leq \int_\Sigma |\text{grad}_\Sigma f|^2 d\text{area} + 2\int_\Sigma K_\Sigma f^2 d\text{area}$$

Let $f(s, \theta) = \eta(s)$ be a test function on D depending only on the radius s. Since f is constant as a function of θ, we integrate out the θ direction to obtain the inequality

$$0 \leq \int_0^r (\eta'(s))^2 \ell(s) ds + 2\int_0^r K'(s)\eta^2(s) ds$$

where $K(s)$ denotes the integral of K over the disk $D(s)$ of radius s, and $\ell(s)$ is the length of its boundary $\partial D(s)$.

Now, Gauss-Bonnet says that $\ell'(s) = 2\pi - K(s)$. Thus integrating by parts we get

$$0 \leq \int_0^r (\eta'(s))^2 \ell(s) ds - 2\int_0^r (2\pi - \ell'(s))(\eta^2(s))' ds$$

Make the explicit choice $\eta(s) := 1 - s/r$ so that $\eta' = -1/r$ and $(\eta^2)' = -2/r(1 - s/r)$. Substituting gives

$$-\frac{1}{r^2}\int_0^r \ell(s) ds + \frac{4}{r}\int_0^r \ell'(s)(1 - s/r) ds \leq \frac{8\pi}{r}\int_0^r (1 - s/r) ds = 4\pi$$

Whereas integration by parts shows

$$LHS = \frac{3}{r^2}\int_0^r \ell(s)ds = \frac{3}{r^2}\text{area}(D) \leq 4\pi$$

□

3.2. **Stable complete minimal surfaces in $\mathbb{R}^3$ are planes.** Colding-Minicozzi's estimate can be used to give a short proof of a famous theorem of do Carmo and Peng [2]:

Theorem 3.8 (do Carmo-Peng)**.** *Let $x : \Omega \to \mathbb{R}^3$ be an oriented stable complete minimal immersion. Then $x(\Omega)$ is a plane.*

Proof. By Corollary 3.6, Stability passes to covering spaces, so we may assume Ω is topologically a plane. Proposition 3.7 and the coarea formula implies that $\ell'(t) \leq 8/3\pi$ for big t, so the total curvature is bounded by $2/3\pi$, and is therefore finite. But Theorem 1.10 says that the total curvature is an integer multiple of 4π, so the only possibility is that K is identically zero, so that g is constant and $x(\Omega)$ is a plane. □

Every graph is calibrated, and therefore stable. Thus this recovers in a special case a theorem of Bernstein that the only complete minimal graph in $\mathbb{R}^3$ is a plane.

3.3. **Schoen's curvature inequality.** All minimal surfaces satisfy a pointwise upper curvature bound, by equation 2.4. In general there can be no pointwise lower curvature bounds, even for stable minimal surfaces. For example, the graph of a complex polynomial $p : \mathbb{C} \to \mathbb{C}$ in $\mathbb{C}^2$ is calibrated, and therefore is always a stable minimal surface. By scaling the graph by a homothety we can find a sequence of such surfaces with $K \to -\infty$.

However in dimension three, Schoen [19] derived *lower* pointwise curvature bounds for a stable minimal surface at a point far from the boundary.

Theorem 3.9 (Schoen)**.** *Let M be a closed 3-manifold, and Σ a stable minimal surface. Given $r \in (0, 1]$ and $p \in \Sigma$ such that $B_r(p) \cap \Sigma$ has compact closure in Σ, there is a constant C depending only on the supremum of $|K_M|$ and $|\nabla K_M|$ on $B_r(p)$, and on r, such that*

$$|\text{II}|^2(p) \leq Cr^{-2}$$

Moreover, there is a constant $\epsilon > 0$ depending on the same data so that $\Sigma \cap B_{\epsilon r}(p)$ is a union of embedded disks.

For stable minimal surfaces in $\mathbb{R}^3$ this simplifies as follows:

Theorem 3.10 (Schoen)**.** *There exists a positive constant $C > 0$ such that for any stable minimal surface Σ in $\mathbb{R}^3$ and any point $p \in \Sigma$ with distance at least r to $\partial\Sigma$, there is an estimate*

$$K_\Sigma(p) \geq -Cr^{-2}$$

Theorem 3.9 can be derived from Colding-Minicozzi's Equation 3.1. Theorem 3.10 follows from Theorem 3.8 as follows: a sequence of stable minimal surfaces in $\mathbb{R}^3$ with $Kr^2 \to -\infty$ can be rescaled so that $K \to -1$ (say) and $r \to \infty$. The limit is a proper stable minimal surface which is not a plane; this is a contradiction.

Schoen's Theorem implies a kind of compactness property for stable minimal surfaces, and has many applications to the theory of 3-manifolds. It implies that in a fixed (closed)

3-manifold M, any sequence of stable minimal surfaces Σ_i with points $p_i \in \Sigma_i$ at which the Σ_i have injectivity radius $\geq r$, has a subsequence that converges on compact subsets to a stable minimal surface

4. Existence of minimal surfaces

In this section we discuss techniques to demonstrate the existence of minimal surfaces satisfying certain geometric and topological conditions.

4.1. Björling's Problem and the Schwarz surface.

The following is known historically as Björling's Problem:

Problem 4.1 (Björling)**.** *Given a real analytic curve γ in $\mathbb{R}^3$ and a real analytic unit normal vector field ν along γ, construct a minimal surface Σ containing γ, with normal field extending ν.*

This problem was solved by Hermann Schwarz, using techniques from complex function theory. His solution is as follows. Suppose $\gamma : I \subset \mathbb{R} \subset \mathbb{C} \to \mathbb{R}^3$ is parameterized to be unit speed. Then $\nu \times \gamma'$ is a unit normal vector field along γ that must be tangent to such a surface Σ and perpendicular to ν.

Now, since γ and ν are real analytic, they admit *holomorphic* extensions $\gamma_{\mathbb{C}}, \nu_{\mathbb{C}} : \Omega \to \mathbb{C}^3$ for some open neighborhood Ω of I in $\mathbb{C}$. If we let z denote the complex coordinate on Ω, and define

$$x(z) := \mathrm{Re}\left\{\gamma_{\mathbb{C}}(z) - i\int_0^z \nu_{\mathbb{C}}(\omega) \times \gamma'_{\mathbb{C}}(\omega)d\omega\right\}$$

then x has coordinates which are harmonic functions of z, and $x(t) = \gamma(t)$ along I.

If we define

$$x_{\mathbb{C}}(z) := \gamma_{\mathbb{C}}(z) - i\int_0^z \nu_{\mathbb{C}}(\omega) \times \gamma'_{\mathbb{C}}(\omega)d\omega$$

then $\gamma'(z) = \mathrm{Re}x'_{\mathbb{C}}(z) = x'$ and $\mathrm{Im}x'_{\mathbb{C}}(z) = -\nu(z) \times \gamma'(z)$ along $\gamma(I)$. Hence $\langle x'_{\mathbb{C}}, x'_{\mathbb{C}}\rangle$ vanishes identically on I, and therefore also on Ω (because it is holomorphic). But if we define $x'_{\mathbb{C}} = (\phi_1, \phi_2, \phi_3)$ this implies the ϕ_j satisfy the minimal surface equation, and x solves the Björling Problem.

In fact, if we run this argument in reverse, it is easy to see that x defines the most general form of a solution to the Björling Problem. From this one obtains the following corollaries:

Corollary 4.2 (Schwarz)**.** *(1) Every straight line contained in a minimal surface is an axis of rotational symmetry of the surface; and*

(2) if a minimal surface intersects a plane perpendicularly, then that plane is a plane of reflectional symmetry of the surface.

These corollaries follow immediately from the symmetry of the surface and the Schwarz reflection principle (which was invented historically in exactly this context).

This corollary motivates the search for real analytic minimal surfaces Σ bounded by a *Schwarz chain*; i.e. a cyclic sequence of straight line segments and flat planes, where we insist that $\partial\Sigma$ should contain each straight line segment, and be perpendicular to each flat plane. Such a surface can then be continued indefinitely by repeated reflection and rotation in the planes and segments.

Example 4.3 (Schwarz surface). We describe in detail a beautiful classical example discovered by Schwarz.

Conider the quadrilateral in $\mathbb{R}^3$ composed of four straight segments between the four vertices A, B, C, D in cyclic order of a regular tetrahedron. For example, we could take

$$A = (1, 1, 1), \quad B = (1, -1, -1), \quad C = (-1, -1, 1), \quad D = (-1, 1, -1)$$

We seek a piece of minimal surface Q that is topologically a disk, and is bounded by this quadrilateral. See Figure 2.

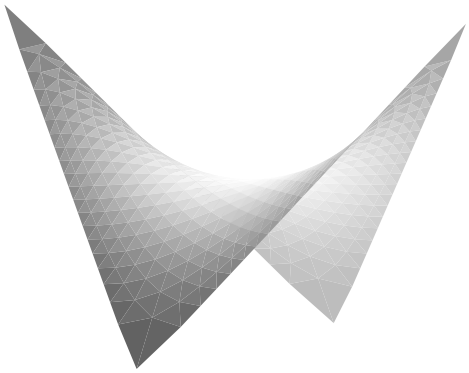

FIGURE 2. The quadrilateral Q

At each vertex the surface Q has a corner with an angle of $60°$. The surface may be continued by rotating copies of Q around each boundary edge successively. Thus, six copies of Q fit together around each vertex, three of them pointing 'up' and three pointing 'down'. By repeated applications of the reflection principle, Q may be continued to a *triply periodic* properly embedded surface $\tilde{\Sigma}$ in $\mathbb{R}^3$ called the *D surface*. To say that it is triply periodic means that there is a lattice Λ of translations in $\mathbb{R}^3$ leaving $\tilde{\Sigma}$ invariant. The quotient $\tilde{\Sigma}/\Lambda = \Sigma$ is a closed surface of genus 3, that we may think of as an embedded minimal surface in the 3-torus $\mathbb{R}^3/\Lambda$. Note that Λ is isomorphic to $\mathbb{Z}^3$ as an abstract group.

The symmetries of $\tilde{\Sigma}$ and Σ will let us determine the functions g, f in the Weierstrass–Enneper parameterization explicitly (see § 1.5).

First we consider g, which we think of as the Gauss map $g : Q \to S^2$ under the usual stereographic identification of S^2 with $\mathbb{CP}^1$. Because ∂Q is the union of four straight segments AB, BC, CD, DA, the normals to Q along ∂Q are segments of great circles in S^2. In fact, these great circles are precisely four of the six great circles one obtains by extending the (geodesic) edges of a spherical "cube". Thus, the images of A, B, C, D under g are the vertices of one face of a regular cube inscribed in S^2, and $g(Q)$ is a regular spherical quadrilateral with angles of $120°$ at each vertex. By extension we get $g : \Sigma \to S^2$, which is a double branched cover, with branch points the eight vertices of an inscribed cube. Four of these vertices are the images of A, B, C, D under g. They are the points

$$\pm\frac{\sqrt{3}-1}{\sqrt{2}} \text{ and } \pm\frac{\sqrt{3}-1}{\sqrt{2}}i$$

The other four vertices are

$$\pm\frac{\sqrt{3}+1}{\sqrt{2}} \text{ and } \pm\frac{\sqrt{3}+1}{\sqrt{2}}i$$

We denote these eight points by ω_j for $j = 1, \cdots, 8$.

To simplify matters we use the Gauss map g itself to define holomorphic coordinates on Σ, at least where it is nonsingular. By abuse of notation we write $f = F(g)$ where f is as in the Weierstrass–Enneper parameterization. Since ∞ is not a branch point for g, the 1-form $F(g)dg$ should have a zero of order 2 at $g = \infty$, and should therefore be of the form $F(\omega) \sim \omega^{-4}$ near infinity. Since it has order 2 branch points at the ω_j the only possibility is

$$F(\omega) := \frac{K}{\prod_j(\omega - \omega_j)^{1/2}} = \frac{K}{\sqrt{\omega^8 - 14\omega^4 + 1}}$$

for some as yet undetermined constant K.

To determine K we need to evaluate the periods of the one forms

$$\phi_1 = \frac{K(1-\omega^2)}{2\sqrt{\omega^8 - 14\omega^4 + 1}}d\omega, \quad \phi_2 = \frac{iK(1+\omega^2)}{2\sqrt{\omega^8 - 14\omega^4 + 1}}d\omega, \quad \phi_3 = \frac{K\omega}{\sqrt{\omega^8 - 14\omega^4 + 1}}d\omega$$

on the surface Σ.

Taking K real ensures that the periods of the ϕ_j are pure imaginary on three suitable generators for the homology of Σ; taking K imaginary makes the period pure imaginary on the other three generators, and produces a conjugate triply periodic surface, called the *P surface*. In 1968, Alan Schoen [18], an engineer at NASA, discovered that taking $K = e^{i\theta}$ where $\theta \sim 0.66348$ gives another triply-periodic embedded surface, known as the *gyroid*, or *G surface*, which contains neither straight lines nor planes of reflection symmetry.

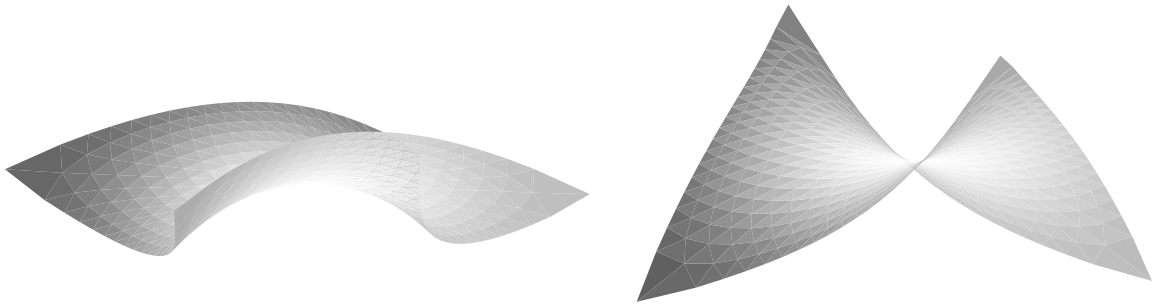

FIGURE 3. Associates of Q in the P and G surfaces

4.2. **Plateau problem.** The *Plateau problem* asks for the surface of least area spanned by a given closed Jordan curve Γ (in $\mathbb{R}^3$ if nothing more is specified). In this formulation the topology of the surface is unspecified; also left in the air is the question whether the surface should be embedded or not. Finally, unless Γ is rectifiable we cannot expect it to bound any surface of finite area, in which case it seems meaningless to ask for one of 'least' area.

Thus to simplify matters we insist first that Γ should be rectifiable, second that the surface we are seeking should be topologically a disk, and thirdly that we do not insist that the surface is embedded or even immersed. In other words we give the Plateau problem the following formulation:

Problem 4.4 (Plateau)**.** *Given a rectifiable Jordan curve Γ in $\mathbb{R}^3$, construct a map*

$$x : \overline{B} \to \mathbb{R}^3$$

from the closed unit disk to $\mathbb{R}^3$ so that $x : \partial B \to \Gamma$ is an orientation-preserving homeomorphism, and so that x is a generalized minimal surface on B.

Recall from Proposition 1.4 that $x|B$ is a generalized minimal surface providing it is conformal (i.e. $x_u \perp x_v$ and $|x_u|^2 = |x_v|^2$ everywhere) and the coordinate functions are harmonic (i.e. $\Delta x = 0$).

The solution, due to Douglas [5] and independently to Rado [15], is to find a suitable class $\mathcal{C}(\Gamma)$ of maps $x : \overline{B} \to \mathbb{R}^3$ with good compactness properties, and to show that the minimizer of a certain functional in $\mathcal{C}(\Gamma)$ solves Plateau's problem. The trick is to decide on the correct functional. The function $A(x)$ which assigns to a map the area of its image is no good for several reasons. First of all, there are too many symmetries: any reparameterization of the domain produces a map with the same area. Second, the area functional is too degenerate: a map of a disk can be modified dramatically by adding 'hair' with very small 2-dimensional measure but very large diameter. Understanding and controlling the geometric limits of such objects is subtle. Douglas's idea was to use a different functional $D(x)$, which measures *energy* instead of area; the minimizer of this functional at once finds the surface with the smallest area, and the most efficient parameterization. A similar philosophy is familiar in the theory of Riemannian geometry, where we seek to minimize energy rather than length to define geodesics.

4.2.1. *The Dirichlet integral and the class $\mathcal{C}^*(\Gamma)$.* Naively we could start by defining $\mathcal{C}(\Gamma)$ to be the class of maps from the closed disk to $\mathbb{R}^3$, smooth in the interior, which restrict to orientation-preserving homeomorphisms $\partial B \to \Gamma$. However, limits of sequences of smooth maps are rarely smooth, and limits of homeomorphisms are not always homeomorphisms.

Thus we extend the class $\mathcal{C}(\Gamma)$ in two ways: first we allow maps which on B are in L^2; this means the *Dirichlet integral*

$$D(x) := \frac{1}{2}\int_B |x_u|^2 + |x_v|^2 du dv$$

is finite (where the derivatives are taken in the sense of distribution). The quantity $D(x)$ is called the *energy* of the map x, and formalizes the physical intuition of elastic energy of a membrane. Second we allow maps for which $\partial B \to \Gamma$ is merely monotone; this means that it is surjective, and the point preimages are connected. A monotone map between circles might degenerate the cyclic order of triples of points, but it does not reverse them. It will turn out that a generalized minimal surface whose boundary parameterization is monotone is actually a homeomorphism; we return to this in the sequel.

A further source of noncompactness comes from the group of symmetries of D. The integrand of D is invariant under conformal reparameterizations of the domain; thus if $\varphi : B \to B$ is any (orientation-preserving) conformal automorphism, we have $D(x) = D(x \circ \varphi)$.

The group of conformal automorphisms of the unit disk is noncompact. Therefore it makes sense to break this symmetry by choosing three points $p_1, p_2, p_3 \in \partial B$ and three points $q_1, q_2, q_3 \in \Gamma$ in the same cyclic order, and restricting to the class $\mathcal{C}^*(\Gamma)$ of maps as above for which $x(p_j) = q_j$.

4.2.2. *Harmonic extension of boundary values.* Now, for a given monotone map $x : \partial B \to \Gamma$ there is a unique extension to $x : \overline{B} \to \mathbb{R}^3$ which minimizes $D(x)$, namely the harmonic extension; i.e. the map for which all the coordinate functions are harmonic. If we identify B with the hyperbolic plane, and ∂B with the circle at infinity, any measurable function f on ∂B extends to a unique harmonic function on the interior as follows. At each point $p \in B$ exponentiation of geodesics defines a bijection between the unit tangent vectors UT_pB to p and ∂B. Under this identification the value of the extension at p is equal to the average of f on ∂B with respect to the angular measure on UT_pB.

If $x_i \to x$ measurably on ∂B, their harmonic extensions converge also, and if by abuse of notation we use the same symbol to denote these extensions, then $D(x_i) \to D(x)$ when these quantities are finite.

Now there is one last source of potential non-compactness we must contend with. An orientation-preserving homeomorphism $y : \partial B \to \Gamma$ determines its graph $g(y) \subset \partial B \times \Gamma$. If we think of $\partial B \times \Gamma$ as a torus, the graph $g(y)$ is a $(1,1)$-curve that intersects every meridian and every longitude in exactly one point. Any sequence of homeomorphisms $y_j : \partial B \to \Gamma$ has a subsequence for which the graphs $g(y_j)$ converge to some $(1,1)$-curve g, but g might have horizontal or vertical segments; see Figure 4.

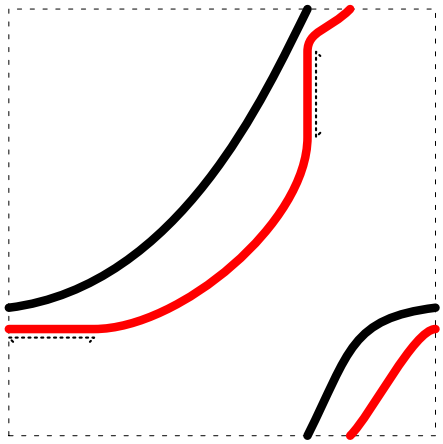

FIGURE 4. the black curve is the graph of a homeomorphism; the red curve is the limit of a sequence of graphs but has both horizontal and vertical segments

A horizontal segment is where the limiting map is monotone but not locally injective; a vertical segment indicates a discontinuity. Note that because we have normalized our boundary maps to take p_j to q_j, a limiting map cannot consist solely of the union of a

meridian and a longitude. In other words, a vertical segment in the limit graph, if it exists, has distinct endpoints.

Our concern is for the possible appearance of vertical segments in g, since such a g is not the graph of a continuous function. So suppose we have a sequence of orientation-preserving homeomorphisms $x_i : \partial B \to \Gamma$ for which the graphs $g(x_i)$ accumulate on a vertical segment. If we let $x : \partial B \to \Gamma$ denote the merely *measurable* map which is a pointwise limit of the x_i then the vertical segment indicates that there is some point $q \in \partial B$ for which the left and right limits $x^{\pm}(q)$ exist and are not equal.

Since x is measurable, it still has a harmonic extension. We claim that for such an x, the harmonic extension has $D(x) = \infty$, and therefore $D(x_i) \to \infty$. To see this, we use the conformal invariance of D. Let φ be a conformal automorphism of B which is a hyperbolic isometry with attracting fixed point q and repelling fixed point r. Consider the harmonic extensions of the sequence of maps $x^n := x \circ \varphi^n$. The boundary maps converge to a step function taking the two values $x^{\pm}$ on the intervals of $\partial B - \{q, r\}$. Let y be the harmonic extension of this step function. If we let $B(1/2)$ denote the ball of radius $1/2$, then

$$\frac{1}{2}\int_{B(1/2)} |y_u|^2 + |y_v|^2 du dv = K > 0$$

Choose φ as above so that the translates $\varphi^n(B(1/2))$ are all disjoint. We can estimate

$$D(x) \geq \sum_n \frac{1}{2}\int_{B(1/2)} |x^n_u|^2 + |x^n_v|^2 du dv$$

But when n is big, each of the integrals has value close to K. Hence $D(x) = \infty$, proving the claim.

Now let $x_j \in \mathcal{C}^*(\Gamma)$ be a sequence of maps with $D(x_j) \to \inf_{x \in \mathcal{C}^*(\Gamma)} D(x)$. By the discussion above, there is a subsequence whose restriction to ∂B converges to a map $x : \partial B \to \Gamma$ which is continuous and monotone. For each x_j, let y_j denote the harmonic extension of $x_j|\partial B$. Then $D(y_j) \leq D(x_j)$ and $y_j \to y$, the harmonic extension of x. Thus

$$D(y) = \lim D(y_j) = \inf_{x \in \mathcal{C}^*(\Gamma)} D(x)$$

4.2.3. *Energy versus Area.* What is the relation between $D(x)$ and $A(x)$, the area of the surface $x(B)$? We can write

$$A(x) := \frac{1}{2}\int_B |x_u \times x_v| du dv$$

Comparing integrands pointwise, we see that $A(x) \leq D(x)$ with equality iff $x_u \perp x_v$ and $|x_u| = |x_v|$ everywhere; i.e. precisely if x is conformal.

From this it follows that any y which minimizes the Dirichlet integral in $\mathcal{C}^*(\Gamma)$ is conformal and, since it is also harmonic, it is minimal. For, suppose y is not conformal. Then we have $A(y) < D(y)$. But we can pull back the conformal structure on the image $y(B)$, and using the fact that the resulting surface is conformally equivalent to the standard disk, we can find a (quasiconformal) homeomorphism $\varphi : B \to B$ so that $y \circ \varphi$ *is* conformal, and lies in $\mathcal{C}^*(\Gamma)$.

Now, reparameterizing a surface does not change its area, so $A(y \circ \varphi) = A(y)$. On the other hand, since $y \circ \varphi$ is conformal,

$$D(y \circ \varphi) = A(y \circ \varphi) < D(y)$$

contrary to the assumption that y minimized D.

We have almost solved Plateau's problem, in that we have constructed a generalized minimal surface y with boundary on Γ, except that a priori $y|\partial B$ might not be a homeomorphism but merely monotone. It turns out this possibility cannot occur. If y were constant on some interval $I \subset \partial B$ we could apply the reflection principle to continue y across I. But this would give a minimal surface which is constant on an interval in the interior, which is impossible.

Note by this discussion that the map y we obtain is also a global minimizer for A. Putting this together we deduce:

Theorem 4.5 (Douglas-Rado)**.** *Any rectifiable Jordan curve Γ in $\mathbb{R}^3$ is the boundary of a generalized minimal surface with the topology of the disk, which furthermore minimizes area among all disks with boundary Γ.*

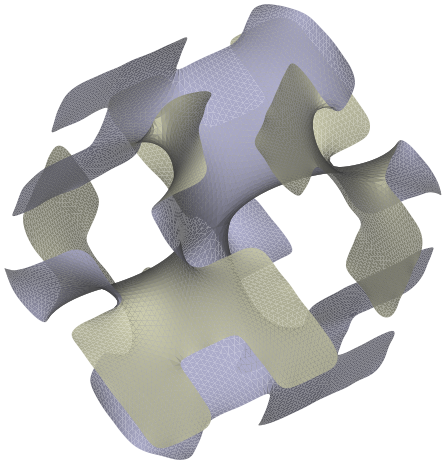

FIGURE 5. A minimal disk bounding an approximation of a Hilbert curve in the boundary of a cube

4.2.4. *Morrey's theorem.* The Douglas-Rado theorem was generalized in 1948 by Morrey [11]. A Riemannian manifold M is said to be *homogeneously regular* if there is a uniform upper bound on the sectional curvature K_M, and a uniform lower bound on the injectivity radius (note that this implies M is complete). For example: any closed Riemannian manifold, or any covering space of such a manifold, is homogeneously regular.

A proof of Morrey's theorem parallel to Douglas' argument can be carried out by introducing the *energy* $E(f)$ of a map $f : B \to M$ with $\partial f : \partial B \to \Gamma$, a quantity that generalizes the Dirichlet integral when M is Euclidean space. An f extending ∂f and minimizing $E(f)$ is said to be harmonic, and a map that is harmonic and conformal is minimal. Morrey's argument is more indirect than this sketch indicates, but a proof along these lines can be obtained by the (more general) methods outlined in § 4.4.

4.3. **Branch points and Gulliver-Osserman's Theorem.** A generalized minimal surface $x : \Omega \to \mathbb{R}^3$ is and immersion away from an isolated set of branch points where dx vanishes. These branch points come in two kinds:

(1) $p \in \Omega$ is a *false* branch point if there is a neighborhood U of p so that $x(U)$ is a smooth embedded surface; and
(2) $p \in \Omega$ is a *true* branch point otherwise.

A false branch point is really an artefact of the parameterization.

We say a minimal surface is *locally least area* if its area cannot be strictly decreased by an arbitrarily small compactly supported variation. A locally least area surface is necessarily stable, but not conversely. In high dimensions locally least area surfaces can have branch points, but for surfaces in dimension three, one has the following:

Theorem 4.6 (Gulliver-Osserman)**.** *A locally least area surface in a 3-manifold has no interior branch points.*

The history of this theorem is that Osserman [13] showed that a locally least area disk in $\mathbb{R}^3$ has no true branch points, and then Gulliver [8] extended this result to all branch points, and to all 3-manifolds.

The fundamental idea is as follows. Near a true branch point p, Osserman shows a minimal surface must have a transverse arc α of self-intersection. Let us leave the proof of this for the moment, and indicate how to modify the map along α near p.

The case of a branch point of 'order 2' is easiest to understand, and is indicated in Figure 6. On the left is an immersed disk with an arc α of intersection emanating from a single branch point p. On the right is a pair of embedded disks which osculate along a single point on the boundary. We may cut out a neighborhood as in the left, replace it with a neighborhood as on the right, and modify a map of a surface locally so as to reduce area. One way to think of this modification is that we have taken the branch point p and 'pushed' it along the arc α where the single branch sheet crosses itself, unzipping the arc of self-intersection and regluing to make two non-singular arcs on the top and bottom sheet.

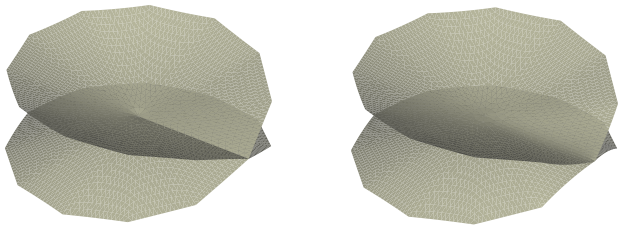

FIGURE 6. Push the branch point along an arc of self-intersection and smooth to reduce area

The modification of the domain is indicated in Figure 7. On the left we have the disk whose diameter maps to α by a two-to-one map branched over the midpoint, which maps to p. We cut open the disk along this diameter, and 'buckle' it outwards to form a diamond-shaped hole, then collapse the diamond along the other axis to form two disks.

Note that this modification can be performed whenever there is a branch point p that ends at an arc of intersection α with (at least) two disjoint preimages in the domain.

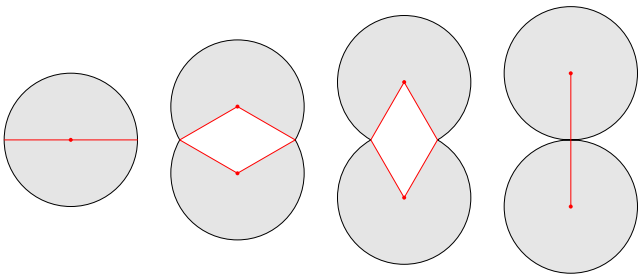

FIGURE 7. A disk is cut along a diameter, to form a diamond-shape hole which buckles out and then collapses to form two disks

To demonstrate the existence of a transverse arc of self-intersection, Osserman considers the Weierstrass parameterization near a branch point in terms of holomorphic functions f and g. Let the branch point be the origin (in the domain), and suppose that the order of vanishing of f and g at the origin are l and m respectively. Then Osserman shows ([13], eq. 3.13 and 3.14) that if we choose a local holomorphic co-ordinate w on the domain so that $g(w) = w^m$, the coordinates x_j of the map have the form

$$x_1 + ix_2 = a_l w^l + \cdots + a_{l+2m-1} w^{l+2m-1} + \left(a_{l+2m} w^{l+2m} - \frac{l}{l+2m} \overline{a}_l \overline{w}^{l+2m} \right) + \cdots \tag{4.1}$$

and

$$x_3 = \mathrm{Re} \left\{ \frac{l}{l+m} a_l w^{l+m} + \cdots \right\} \tag{4.2}$$

for certain coefficients a_j.

From the form of this equation one can show either that x multiply covers its image (i.e. we are at a fake branch point) or there is a $\delta > 0$ with the following property: whenever $0 < |w_1|, |w_2| < \delta$ are distinct points with the same image $x(w_1) = x(w_2)$, then $g(w_1) \neq g(w_2)$; i.e. the surface intersects itself transversely.

It follows that sufficiently near the branch point, the subset where the surface intersects itself consists of a discrete collection of pairs of real analytic arcs. For topological reasons there must be self-intersections arbitrarily near the branch point; thus if we analytically continue the (finitely many!) pairs of arcs of self-intersection within the disk $|w| < \delta$, at least one pair can be continued all the way to 0.

Thus the minimal surfaces constructed by Douglas have no interior branch points.

4.4. **Minimal surfaces in Riemannian manifolds.** Let M be a closed Riemannian manifold, and let $f : \Sigma_g \to M$ be a continuous map where Σ_g is a closed oriented surface of genus g. It is natural to ask: when is there a map $h : \Sigma_g \to M$, homotopic to f, for which $h(\Sigma_g)$ is a minimal surface?

A rather satisfying answer was obtained independently by Schoen-Yau [20] and Sacks-Uhlenbeck [17]:

Theorem 4.7 (Schoen-Yau, Sacks-Uhlenbeck)**.** *Let M be a closed Riemannian manifold. Suppose that $f : \Sigma_g \to M$ is a continuous map such that $f_* : \pi_1(\Sigma_g) \to \pi_1(M)$ has no conjugacy class corresponding to an essential simple closed curve in the kernel. Then there*

is a map $h : \Sigma_g \to M$ for which $h(\Sigma_g)$ is a branched minimal surface, and such that h induces the same map as f on π_1.

If $\pi_2(M) = 1$ we may take h to be homotopic to f. Finally, if M has dimension 3 then h can be chosen to have no branch points.

The basic strategy is very close to that of Douglas. If we fix a marked conformal structure ϕ on Σ_g, we can look for a *harmonic* map $f_\phi : \Sigma_g \to M$ in the homotopy class of f. A harmonic map is one which minimizes energy; a harmonic map which is conformal is minimal.

To construct f_ϕ we must understand convergence properties of sequences of maps with energy bounds; a limit might 'bubble off' spheres; cutting off these spheres changes the homotopy class of the map by an element of $\pi_2(M)$.

Varying the marked conformal structure leads us to confront the non-compactness of the Teichmüller space $\mathcal{T}$ of marked conformal structure on Σ_g. It is here that we make use of the hypothesis that f_* should be injective on essential simple curves.

Finally we observe that the surface we obtain is locally least area; thus if M is 3-dimensional, Theorem 4.6 of Gulliver-Osserman implies that h has no branch points.

4.4.1. *Harmonic maps.* We discuss briefly the setup for the theory of harmonic maps between Riemannian manifolds, as developed by Eells-Sampson [6].

For a smooth map $f : N \to M$ between Riemannian manifolds, the *energy* of f is the functional

$$E(f) := \frac{1}{2}\int_N |df|^2 d\text{vol}$$

The integrand $|df|^2/2$ is the *energy density* of f. Here we are thinking of df as a section of the bundle $T^*N \otimes f^*(TM)$, so that if e_i is an orthonormal basis for TN at a point p, then $|df_p|^2 = \sum |df_p(e_i)|^2$. Scaling the metric on N at p by λ scales $|df|^2$ by λ^{-2} and $d$vol by λ^d where d is the dimension of N. Thus if N is a surface, $E(f)$ depends only on the conformal structure on the domain.

Let $F : N \times (-\epsilon, \epsilon) \to M$ be a smooth 1-parameter family of maps with $f_t := F(\cdot, t)$ so that $f_0 = f$ and $f_t = f$ outside a compact set. Let $X = dF(\partial_t)$ and if e_i is an orthonormal basis of TN locally, let $Y_i(t) := df_t(e_i)$. Note that X and the Y_i are commuting sections of F^*TM and therefore $\nabla_X Y_i = \nabla_{Y_i} X$ for each i, where ∇ is pulled back from the Levi-Civita connection on TM.

By abuse of notation we use X and Y_i to refer to the restrictions of these vector fields in f^*TM over $N \times 0$.

We compute

$$\frac{d}{dt}\Big|_{t=0} E(f_t) = \int_N \sum_i \langle Y_i, \nabla_X Y_i\rangle d\text{vol} = \int_N \sum_i \langle Y_i, \nabla_{Y_i} X\rangle d\text{vol}$$

Area is invariant under tangential variations, but energy is not. We define a 1-form α on N which (up to the canonical identification of 1-forms and vector fields via the metric) is the pullback of the *tangential part* of X. Thus: let α be the 1-form on N defined pointwise

by $\alpha(V) := \langle X, df(V)\rangle$. Then $\alpha^\sharp$ is a vector field on N with divergence

$$\begin{aligned}\operatorname{div}\alpha^\sharp &= \sum_i e_i(\alpha(e_i)) - \alpha(\nabla_{e_i}e_i) = \sum_i e_i\langle X, Y_i\rangle - \langle X, df(\nabla_{e_i}e_i)\rangle \\ &= \sum_i \langle \nabla_{Y_i} X, Y_i\rangle + \langle X, \nabla_{Y_i}Y_i - df(\nabla_{e_i}e_i)\rangle\end{aligned}$$

Define the *tension* by the formula

$$\tau(f) := \sum_i \nabla_{df(e_i)} df(e_i) - df(\nabla_{e_i}e_i)$$

where the sum is taken over an orthonormal basis e_i pointwise (more intrinsically we could write $\tau(f) = \operatorname{tr}\nabla df$). Note that this is a vector field along $f(N)$. Since X is compactly supported, $\operatorname{div}\alpha^\sharp$ integrates to zero, and we obtain the first variation formula

$$\left.\frac{d}{dt}\right|_{t=0} E(f_t) = -\int_N \langle X, \tau(f)\rangle d\mathrm{vol}$$

for any smooth compactly supported variation tangent to X. In particular, f is a critical point for E if and only if $\tau = 0$. Such a map is called *harmonic*.

Example 4.8. If N and M are Kähler manifolds (for example, if they are nonsingular projective varieties with the induced metric), every holomorphic map $f : N \to M$ is harmonic. To see this, observe that at a point p in N, if we choose geodesic normal co-ordinates at p and $f(p)$ the tension operator reduces (in local co-ordinates) to the usual Laplacian. But for Kähler manifolds we can choose geodesic local co-ordinates which are the same time holomorphic co-ordinates, and then use the fact that for a Kähler manifold holomorphic functions are harmonic (in the usual sense).

For a conformal map $f : N \to M$, if e_i' is an orthonormal frame on M with $df(e_i) = \lambda e_i'$ then

$$\tau(f) = \sum_i \lambda^2 \nabla_{e_i'} e_i' + \lambda e_i'(\lambda) e_i' - df(\nabla_{e_i}e_i)$$

so the normal part to M of the tension field is pointwise proportional to the mean curvature. Comparing with the first variation formula for area (i.e. Proposition 2.1) we see that a harmonic map which is conformal is minimal.

Conversely, let us suppose that $N = \Sigma$ is 2-dimensional. Let $f : \Sigma \to M$ be harmonic but not conformal. There is another surface Σ' with a Riemannian metric and an area-preserving diffeomorphism $h : \Sigma' \to \Sigma$ so that $f \circ h : \Sigma' \to M$ is conformal. It is elementary that wherever f fails to be conformal, the energy density of $f \circ h$ is strictly smaller than that of f, and therefore $E(f \circ h) < E(f)$. In fact, one easily sees that for any $f : \Sigma \to M$ the area of f is less than or equal to the energy, with equality if and only if f is conformal.

We conclude that a map $f : \Sigma \to M$ that minimizes energy in its homotopy class over all choices of conformal structure on the domain is both harmonic and conformal, and is consequently a (locally least area) minimal surface. As a special case, any harmonic $f : S^2 \to M$ is conformal and thus minimal.

4.4.2. *Bubbling off.* For maps $f : N \to \mathbb{R}^n$ the energy is the sum of the L^2 norm of the coordinate functions, and a map is harmonic if each coordinate function is harmonic in the usual sense. If $f_i : N \to \mathbb{R}$ is a sequence of smooth functions with uniformly bounded energy, a limit (if it exists) might be no more regular that L^2. Rather than attempt to adapt the theory of Sobolev spaces to maps between Riemannian manifolds, it is convenient to appeal to a famous theorem of Nash that every Riemannian manifold M can be isometrically embedded in some high dimensional $\mathbb{R}^n$. Then for a smooth map $f : N \to M$ the energy of f is the same as the energy of the composition $N \to M \to \mathbb{R}^n$, and it makes sense to talk about L^2 limits of sequences of maps. Note that a map $f : N \to M \subset \mathbb{R}^n$ is harmonic (as a map to M) if the tension field (as a map to $\mathbb{R}^n$) is normal to M.

From now on we specialize to the case that the domain is Σ, a surface, mapping to a compact Riemannian manifold M.

Let's fix a homotopy class of map, and look for a sequence of maps $f_i : \Sigma \to M$ in this homotopy class whose energies converge to the infimum. Since $E(f_i)$ is bounded, there is a subsequence which converges weakly and pointwise almost everywhere to an L^2 map $f : \Sigma \to M$ for which $E(f) \leq \lim E(f_i)$. To say the subsequence converges weakly means that for every L^2 map g we have

$$\int_\Sigma \langle df_i, dg \rangle d\text{vol} \to \int_\Sigma \langle df, dg \rangle d\text{vol}$$

Note that at this point we do not know if f is continuous, or in case it is, that it is in the desired homotopy class.

Energy can go down in a limit if a definite amount of energy of f_i becomes concentrated in a sequence of smaller and smaller regions U_i that shrink to a collection of points. Conformally rescaling f_i near such a limit point gives a sequence of maps whose domains converge to $\mathbb{C}$, and after reparameterization by a diffeomorphism we may assume (reducing the energy in the process) these maps are themselves conformal. The energy density of the resulting maps might blow up and we can iterate this process, obtaining a 'tree' of finite energy harmonic maps from punctured spheres to M in the limit.

This situation is analyzed carefully by Sacks-Uhlenbeck. They show [16] Thm. 3.6 that a finite area harmonic map from a punctured disk to a Riemannian manifold may be extended smoothly to a harmonic map over the puncture, and thus the limit map fills in over the punctures of each sphere to give a harmonic map $S^2 \to M$. They are able to obtain therefore a limit $\overline{f}$ defined on a new domain $\overline{\Sigma}$, called a *bubble tree*. It is built from Σ by attaching 'depth 1' spheres at the points where the energy density blows up in Σ, then attaching 'depth 2' spheres where the energy density blows up in the the depth 1 spheres, and so on. The spheres in $\overline{\Sigma}$ are the bubbles, and they carry the energy that has 'bubbled off' from Σ.

The restriction of $\overline{f}$ to each sphere S^2 in $\overline{\Sigma}$ is a (nonconstant) harmonic map, and is therefore a minimal surface. If M is compact, there is a uniform upper bound on K_M and therefore by Gauss-Bonnet a positive *lower* bound on the area — hence the energy — of $\overline{f}$ restricted to each sphere in the tree. It follows that the number of bubbles in the tree is *finite*, and $\overline{\Sigma}$ admits the natural structure of a nodal Riemann surface.

Cutting off bubbles changes the map $\overline{f}$ by elements of π_2. Thus the restriction $\overline{f}$ to Σ is a harmonic map that might not be homotopic to the f_i, but induces the same map on π_1 (the details of the proof of this claim depend on a careful analysis of convergence to the bubble tree, ruling out the possibility that each bubble might be attached to the next by vanishingly thin 'necks' of positive diameter; see Parker [14]).

4.4.3. *Variation in Teichmüller space.* In the previous sections we have indicated the proof that for any continuous $f : \Sigma \to M$, and for each fixed marked conformal structure ϕ on Σ, there is a harmonic map $f_\phi : \Sigma \to M$ inducing the same map on π_1 as f. We may now ask how f_ϕ varies as a function of ϕ. If f_ϕ is not conformal, we may reduce its energy by a diffeomorphism of the domain whose composition with f_ϕ is conformal. Thus we will obtain a minimal surface if we can find a ϕ which realizes the infimum of $E(f_\phi)$.

Let ϕ_i be a sequence of marked conformal structures on Σ; i.e. homotopy classes of diffeomorphisms $\phi_i : \Sigma \to S_i$ where S_i is a Riemann surface, and let $f_i : S_i \to M$ be harmonic, so that $\phi_i \circ f_i$ induces the same map on π_1 as f. Suppose further that $E(f_i)$ converges to the infimum. We would like to argue that ϕ_i converges to some $\phi \in \mathcal{T}(\Sigma)$, the Teichmüller space of Σ.

There are two sources of non-compactness: that coming from the conformal structures on S_i, and that coming from the marking.

First we show that the S_i lie in a compact subset of moduli space. This means that there is a uniform upper bound on the moduli of embedded annuli in S_i. Let A be an annulus in some S_i conformally equivalent to a Euclidean cylinder $S^1 \times [0, R]$. We want a bound on R. Let's denote the restriction of f_i to A by $h : S^1 \times [0, R] \to M$. By hypothesis, the core circle of the annulus maps to a homotopically nontrivial loop in M; therefore there is a positive constant $\ell > 0$ so that the map $h_t : S^1 \to M$ defined by $h_t(\theta) = h(\theta, t)$ has length at least ℓ, and therefore $E(h_t) \geq \ell^2/4\pi$ by Cauchy-Schwarz. Hence

$$E(f_i) \geq E(h) \geq \int_0^R E(h_t)dt \geq R\ell^2/4\pi$$

so an upper bound for $E(f_i)$ gives an upper bound on R, and the claim is proved.

The noncompactness of Teichmüller space is more serious, since it can easily happen that energy is not proper on $\mathcal{T}$.

Example 4.9. Any $\alpha \in \pi_1(M)$ which normalizes the image of $\pi_1(\Sigma)$ will induce an outer automorphism of any marking of Σ. For example, if M is a surface bundle over a circle with fiber homeomorphic to Σ, dragging Σ around the circle direction induces the monodromy action on the marking.

It turns out that the conjugation action of the normalizer of $f_*\pi_1(\Sigma)$ in $\pi_1(M)$ accounts for all the non-properness of the energy functional E on $\mathcal{T}$.

Returning to our family of marked conformal structures $\phi_i : \Sigma \to S_i$ and harmonic maps $f_i : S_i \to M$ with $E(f_i)$ converging to the infimum, we have already argued that the S_i must lie in a compact subset of moduli space. Fix a subsequence with $S_i \to S$ in moduli space. For each i there is a family of simple based loops Γ_i in S_i that generate $\pi_1(S_i)$, and we can choose such loops so that $\Gamma_i \to \Gamma \subset S$; this induces a family of homotopy equivalences $\varphi_i : S_i \to S$; note that we do *not* assume these maps are consistent with the markings ϕ_i; i.e. $\varphi_i \circ \phi_i$ need not be homotopic to $\varphi_j \circ \phi_j$ for $i \neq j$.

As in § 4.4.2 some subsequence of the f_i converges away from finitely many points (where energy might bubble off in trees of spheres) to $\overline{f} : S \to M$. In particular, for large i, j the maps $f_i \circ \varphi_i^{-1}$ and $f_j \circ \varphi_j^{-1}$ induce the same maps on $\pi_1(S) \to \pi_1(M)$, and a free homotopy of the image of the basepoint under $f_i \circ \phi_i$ to the image under $f_j \circ \phi_j$ gives an element of $\pi_1(M)$ normalizing $f_*(\pi_1(\Sigma))$. It follows that we can find a new sequence of markings (ϕ_i', S_i) in a compact subset of Teichmüller space converging to (ϕ', S) so that $\overline{f} \circ \phi' : \Sigma \to M$ is a minimal surface satisfying the conclusion of Theorem 4.7.

5. Embedded minimal surfaces in 3-manifolds

Theorem 4.7 allows us to construct immersed minimal surfaces in 3-manifolds under very general hypotheses on the fundamental group. However, for applications it is sometimes important for surfaces to be *embedded*. In this section we describe results which guarantee the existence of embedded minimal surfaces certain isotopy classes.

5.1. Embedded minimal disks with prescribed boundary.

Let $\Gamma \subset M$ be an embedded Jordan curve in a closed 3-manifold M. If Γ is null-homotopic, Morrey's theorem shows that there is a least area $f : D \to M$ with $f : \partial D \to \Gamma$ a homeomorphism. Note by Gulliver-Osserman f has no interior branch points. It is natural to ask: under what conditions is f an embedding?

Even for $\Gamma \subset \mathbb{R}^3$ this is a subtle question. One necessary condition is that Γ should be unknotted — i.e. it should bound *some* embedded disk. However this condition is not sufficient.

Example 5.1 (Almgren-Thurston [1]). There is an unknotted $\Gamma \subset \mathbb{R}^3$ that does not bound an embedded disk within its convex hull.

Proposition 5.2 (Total curvature $\leq 4\pi$)**.** *Let $\Gamma \subset \mathbb{R}^3$ be a Jordan curve with total geodesic curvature at most 4π. Then Γ bounds an embedded minimal disk.*

5.2. Mean convex boundary.

Definition 5.3. Let M be a compact 3-manifold with smooth boundary ∂M. We say that M is *mean convex* if the mean curvature field H along ∂M does not point out of M, and M is *strictly mean convex* if H is nonzero everywhere on ∂M and points in to M.

For example, a convex subset of $\mathbb{R}^3$ is mean convex. Likewise, a small enough round ball in any Riemannian manifold is mean convex.

Mean convexity is the right boundary condition to put on a manifold in order to ensure the existence of minimal surfaces.

Theorem 5.4 (Meeks-Yau [10], Thm. 1)**.** *Let M be a mean convex 3-manifold, and let $\Gamma \subset \partial M$ be an embedded Jordan curve that is null-homotopic in M. Then Γ bounds a least area disk in M, and any such disk is properly embedded.*

Proof. The first step is to show that there is an immersed least area disk $f : B \to M$ with $f : \partial B \to \Gamma$. To do this we embed M as a codimension 0 submanifold of a Riemannian manifold N that is homogeneously regular in the sense of Morrey (see § 4.2.4) so that there is $f : B \to N$ with $f : \partial B \to \Gamma$, and then reason *a posteriori* that $f(B) \subset M$.

By perturbing the metric on M very slightly near the boundary, we may assume ∂M is strictly convex. We let $N := M \cup_{\partial M} \partial M \times [0, \infty)$ where we put the product metric on $\partial M \times [0, \infty)$. This is a C^0 Riemannian metric, and will not typically be smooth along $\partial M = \partial M \times 0$. We give a collar neighborhood of ∂M in M the structure of a product $\partial M \times (-\epsilon, 0]$ by exponentiating the unit normal field. Then the entire $\partial M \times (-\epsilon, \infty) \subset N$ has a metric of the form $g(t) \oplus dt^2$ where $g(t)$ is a C^0 family of metrics on ∂M, and $g(t) = g(0)$ for $t > 0$. We define a new C^∞ family of metrics $h(t)$ on ∂M by averaging g (pointwise as a function of t) with respect to a smoothly varying family of smooth probability measures ϕ_t centered at t and with support equal to $(-\epsilon, \infty)$. The mean curvature of $\partial M \times t$ at each point $(p, t) \in \partial M \times (-\epsilon, \infty)$ with respect to the $h(t)$ metric is (to first order) a weighted average of the mean curvatures in the $g(t)$ metric at nearby (p, s); in particular, the end of N is foliated by strictly mean convex surfaces. Then if $f : B \to N$ is minimal and not contained in M, it has an interior point of tangency with some strictly mean convex surface, violating minimality. We conclude $f : B \to M$. Since our metric on M is as close to the original metric as desired, we can obtain minimal $f : B \to M$ (in the original metric) as a suitable limit.

The second step is to show that any least area disk whose boundary is embedded is itself embedded. If the manifold and the map are simplicial with respect to some triangulation, one can build a tower as in Papakyriakopolos' proof of Dehn's Lemma, where the map of the disk lifts to each step of the tower, and where at the top step the map has only simple self-intersections; at this step one can perform cut-and-paste to build a new map with the same area but which is singular along the self-intersections — such a map cannot be minimal, thus *a posteriori* there could have been no self-intersections at the top of the tower, and therefore none at all: i.e. the original map was an embedding.

If M and Γ are real analytic, so is the minimal surface, and everything can be taken to be simplicial, and the argument goes through. Approximating the original metric and curve by real analytic ones we obtain the desired result. □

The first step of the argument has nothing to do with disks *per se*, and can be used to construct minimal surfaces in 3-manifolds with mean convex boundary.

A nice application is the so-called *bridge principle*:

Corollary 5.5 (Bridge Principle; Meeks-Yau [10]). *If X and Y are two strictly stable two-sided minimal surfaces in $\mathbb{R}^3$ and α is an arc joining their boundaries, there is a new stable minimal surface Z that is close to $X \cup Y$ attached by a 'bridge' along α.*

We indicate the idea of the proof. Suppose first that X and Y are embedded, and that α is disjoint from both except at the endpoints. We round α at the endpoints so it is normal to ∂X, ∂Y and then build a 3-manifold M which is the union of rounded thickened neighborhoods NX and NY of X and Y with a thickened neighborhood $N\alpha$ of α.

Since X is minimal it has mean curvature 0. Let f be an eigenfunction for the stability operator L of least eigenvalue, and recall that f is nowhere zero. We can push X in the normal direction by a distance $\pm tf$ for small t to produce $X(\pm t)$. Intuitively, since X is stable and f is an eigenfunction, the variation increases area locally to second order everywhere; hence $X(\pm\epsilon)$ for small enough ϵ has mean curvature pointing in to X everywhere, and together they bound NX which is mean convex (NY is constructed similarly).

A sufficiently thin tube $N\alpha$ around α has arbitrarily large principle curvatures along the meridians into the tube, and therefore the entire boundary is mean convex. Finally, the frontier where $N\alpha$ meets NX and NY can be rounded and modeled on catenoids; the result is the desired mean convex 3-manifold M.

We can build a Jordan curve Γ in ∂M by attaching a 1-handle along the tube from ∂X to ∂Y, and then span this by a surface Z in the desired homotopy class; we can assume Z is least area in M, and therefore Z is stable in $\mathbb{R}^3$.

If X and Y are not embedded we can nevertheless build M as above together with an immersion into $\mathbb{R}^3$.

Using the bridge principle, Hall [9] showed that one cannot replace 'least area' with 'stable' in Theorem 5.4.

Example 5.6 (Hall [9]). There is an embedded Jordan curve on S^2 that bounds an immersed stable minimal disk that is not embedded. To see this, take two copies of the equator and perturb them slightly. They bound a pair of (almost horizontal) disks; by the bridge principle there is a stable minimal disk D_1 obtained by attaching a small 1-handle to these two disks. This stable minimal disk is *not* least area; a least area disk D_2 is contained near a narrow strip around the equator. Now perturb the boundaries of D_1 and D_2 slightly so that D_1 intersects D_2 transversely, but their boundaries are disjoint in S^2. By the bridge principle we can tube these disks together to make D stable and not embedded, with embedded boundary. See Figure 8.

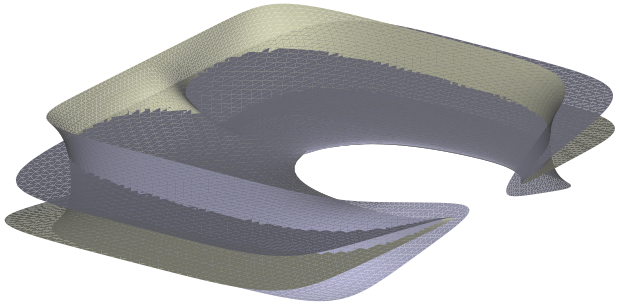

FIGURE 8. A self-intersecting stable disk with embedded boundary on the boundary of a convex set

5.3. **Embedded incompressible surfaces and the roundoff trick.**

Theorem 5.7 (Freedman-Hass-Scott [7] Thm. 5.1)**.** *Let M be a closed irreducible Riemannian 3-manifold, and let Σ be a closed surface of non-positive Euler characteristic. Let $f : \Sigma \to M$ be a least area immersion which is π_1-injective, and such that f is homotopic to a two-sided embedding g. Then either*

(1) f is an embedding; or
(2) f double covers a one-sided minimal surface K embedded in M, and $g(\Sigma)$ bounds a twisted I-bundle over a surface isotopic to K.

An irreducible 3-manifold has trivial π_2, by the sphere theorem. If $g : \Sigma \to M$ is a π_1-injective embedding, Theorem 4.7 guarantees the existence of a homotopic immersed minimal surface $f : \Sigma \to M$ in the same homotopy class. Conversely, a two-sided embedding $g : \Sigma \to M$ which is injective on essential simple curves is π_1-injective. Thus we can find f satisfying the hypothesis of Theorem 5.7 under many circumstances.

In the sequel we give a brief outline of the argument, following [7] which should be consulted for details.

5.3.1. *Cut and paste and the roundoff trick.* Let Σ be a 2-sided oriented immersed surface in a Riemannian 3-manifold M (we do not assume Σ is connected), and suppose Σ intersects itself transversely in double curves and isolated triple points.

If there are no triple points, the self-intersection locus Γ is a union of circles, and we can perform cut and paste along some circle γ. This means that we cut Σ along the preimage of γ, and reattach the local sheets by a permutation compatible with the orientations. The resulting surface is no longer immersed; it has a corner along the preimage of γ. This corner can be rounded, reducing the number of self-intersections and reducing area; this is called the *roundoff trick*. The resulting surface Σ' then satisfies $\text{area}(\Sigma') < \text{area}(\Sigma)$, has no triple points of self-intersection, and has fewer curves of self-intersection than Σ.

Under suitable topological conditions it might happen that Σ' as above is homotopic to Σ, so that we can deduce that any Σ as above is not least area in its homotopy class.

5.3.2. *Intersections near a tangency.* One of the subtleties of applying cut and paste and the roundoff trick is that (self)-intersections of minimal surfaces in 3-manifolds need not be in general position. Nevertheless one can give a normal form for the intersection near points of tangency, which we now describe.

Let Σ be a 2-sided oriented immersed minimal surface in a 3-manifold and let Σ_1, Σ_2 be two embedded sheets in Σ that intersect at a point p of common tangency.

It turns out there is a choice of local C^1 coordinates x, y, z with p at the origin, in which Σ_1 is the x–y plane, and Σ_2 looks like the graph of $\text{Re}\left((x+iy)^n\right)$ for some n.

For surfaces in $\mathbb{R}^3$ this can be deduced from the Weierstrass parameterization, and is analagous to (but simpler than) Osserman's model near an interior branch point; compare § 4.3. In an arbitrary Riemannian 3-manifold one must appeal to the general theory of quasilinear second order elliptic PDE; see [7] Lemma 1.4 and 1.5.

Using this local model, we can perform cut and paste and roundoff for Σ as follows. First, either Σ factors through a covering map, or the points of self-tangency are isolated (we assume we are in the second case). Let p be a point where Σ intersects itself transversely, and suppose there is a neighborhood $N(p)$ with the property that cut and paste and roundoff supported in $N(p)$ will reduce area by at least $\epsilon > 0$. We perturb Σ to Σ' by a perturbation supported outside $N(p)$ in such a way as to remove all the points of self-tangency; this can be done in such a way that $\text{area}(\Sigma') < \text{area}(\Sigma) + \epsilon$. Assuming there are no triple points, we can now perform cut and paste and roundoff on Σ', producing a new surface Σ'' with fewer double curves, where by hypothesis $\text{area}(\Sigma'') \leq \text{area}(\Sigma') - \epsilon < \text{area}(\Sigma)$.

5.3.3. *Embedded surfaces in homotopy equivalences.* If M is closed and irreducible and $f : \Sigma \to M$ is π_1-injective, we can lift f to $f_\Sigma : \Sigma \to M_\Sigma$, where M_Σ is the covering space of M with fundamental group equal to $f_*\pi_1(\Sigma)$. Then M_Σ is noncompact, but is complete

with respect to the metric pulled back from M, and f_Σ is a homotopy equivalence (in fact, M_Σ is necessarily homeomorphic to $\Sigma \times \mathbb{R}$, but that is not essential for what follows). In this context we have the following:

Proposition 5.8 ([7] Thm. 2.1)**.** *Let M be an oriented Riemannian 3-manifold without boundary, and let $f : \Sigma \to M$ be a homotopy equivalence, where Σ is a closed oriented surface other than S^2. If f is least area in its homotopy class, then it is an embedding.*

Proof. In the sequel we use homology with $\mathbb{Z}/2\mathbb{Z}$ coefficients.

If Σ is embedded, we are done, so we suppose to the contrary that Σ is not embedded.

Since f is a homotopy equivalence, it is an isomorphism on homology, and therefore if N is a regular neighborhood of $f(\Sigma)$, the map $\Sigma \to N$ is injective on homology. Since M is homotopic to Σ and without boundary, it has two ends, which are separated by the image of Σ. Thus ∂N contains an 'upper boundary' A which cobounds one of the ends, and $B := \partial N - A$ contains the 'lower boundary' which cobounds the other end. In particular, if $g : M \to \Sigma$ is a homotopy inverse to f, then $A \to M \to \Sigma$ and $B \to M \to \Sigma$ are both degree one, and therefore A and B both have rank H_1 at least as large as Σ.

We now argue separately depending on whether $H_1(\Sigma) \to H_1(N)$ is surjective or not.

Case 1: $H_1(\Sigma) \to H_1(N)$ is surjective.

For any compact 3-manifold, $H_1(\partial N) \to H_1(N)$ kills a half-dimensional subspace. We conclude that A and B have the same rank of H_1 as Σ, and since $A \to M \to \Sigma$ is degree one, A is homeomorphic to A' together with a collection of spheres, where $A' \to M \to \Sigma$ is a homotopy equivalence. Likewise there is $B' \subset B$ where $B' \to M \to \Sigma$ is a homotopy equivalence.

Since Σ is not embedded, by cut and paste and suitable roundoff we can find a neighborhood N with $\text{area}(\partial N) < 2\,\text{area}(\Sigma)$, so that either A' or B' violates the hypothesis that Σ is least area in its homotopy class.

Case 2: $H_1(\Sigma) \to H_1(N)$ is not surjective.

In this case we apply a tower argument. We relabel M, N and $f : \Sigma \to N$ as M_0, N_0 and $f_0 : \Sigma \to N_0$. Since $f_0 : H_1(\Sigma) \to H_1(N_0)$ is not surjective, there is a degree 2 cover $p_1 : M_1 \to N_0$ and a lift $f_1 : \Sigma \to M_1$ with regular neighborhood $N_1 \to M_1$. If $H_1(\Sigma) \to H_1(N_1)$ is not surjective, we pass to another cover, and repeat; each nontrivial lift produces f_j with fewer self-intersections than before, so that the tower terminates at $f_k : \Sigma \to M_k$ with neighborhood N_k and $H_1(\Sigma) \to H_1(N_k)$ an isomorphism.

Some homological algebra ([7] Cor. 2.3) shows that $\partial N_k = A_k \cup B_k$ each homologous to $f_k(\Sigma)$. Thus, as above, there are $A'_k \subset A_k$ and $B'_k \subset B_k$ each homeomorphic to Σ and mapping down to M in the homotopy class of f. At least one has area less than Σ unless f_k is an embedding. But if f_k is an embedding then since p_k has degree 2, the map $p_k \circ f_k : \Sigma \to M_{k-1}$ has no triple points, and after a small perturbation has only embedded double curves of self-intersections. Thus as in § 5.3.2 we can do cut and paste and roundoff to produce a surface in the same homotopy class with smaller area. This completes the argument. □

5.3.4. *Least area property under coverings.*

5.4. **Sweepouts and index one surfaces.**

5.5. **Combinatorial minimal surfaces.**

6. Acknowledgments

Danny Calegari was supported by NSF grant DMS 1005246.

References

[1] F. Almgren and W. Thurston, *Examples of unknotted curves which bound only surfaces of high genus within their convex hulls*, Ann. Math. **105** (1977), pp. 527–538

[2] M. do Carmo and C. Peng, *Stable complete minimal surfaces in $\mathbb{R}^3$ are planes*, Bull. AMS **1** (1979), no. 6, pp. 903–906

[3] T. Colding and W. Minicozzi, *A course in minimal surfaces*, AMS Graduate Texts in Mathematics **121**, American Mathematical Society, Providence, RI, 2011

[4] T. Colding and W. Minicozzi, *Estimates for parametric elliptic integrands*, IMRN **2002** (2002), pp. 291–297

[5] J. Douglas, *Solution of the problem of Plateau*, Trans. AMS **33** no. 1, pp. 263–321

[6] J. Eells and J. Sampson, *Harmonic mappings of Riemannian manifolds*, Amer. J. Math. **86** no. 1, (1964), pp. 109–160

[7] M. Freedman, J. Hass and P. Scott, *Lease area incompressible surfaces in 3-manifolds*, Invent. Math. **71** (1983), pp. 609–642

[8] R. Gulliver, *Regularity of minimizing surfaces of prescribed mean curvature*, Ann. Math. (2) **97** (1973), pp. 275–305

[9] P. Hall, *Two topological examples in minimal surface theory*, J. Diff. Geom. **19** (1984), no. 2, pp. 475–581

[10] W. Meeks and S.-T. Yau, *The Existence of embedded minimal surfaces and the problem of uniqueness*, Math. Z. **179** (1982), pp. 151–168

[11] C. Morrey, *The problem of Plateau on a Riemannian manifold*, Ann. Math. (2) **49**, no. 4 (1948), pp. 807–851

[12] R. Osserman, *A survey of minimal surfaces*, Second Edition. Dover Publications, Inc., NY, 1986

[13] R. Osserman, *A proof of the regularity everywhere of the classical solution to Plateau's problem*, Ann. Math. **91** no. 3, (1970), pp. 550–569

[14] T. Parker, *Bubble tree convergence for harmonic maps*, J. Diff. Geom. **44** (1996), pp. 595–633

[15] T. Radó, *On Plateau's problem*, Ann. Math. (2) **31** no. 3, (1930), pp. 457–469

[16] J. Sacks and K. Uhlenbeck, *The existence of minimal immersions of 2-spheres*, Ann. Math. **113** no. 1, (1981), pp. 1–24

[17] J. Sacks and K. Uhlenbeck, *Minimal immersions of closed Riemann surfaces*, Trans. AMS **271** (1982), no. 2, pp. 639–652

[18] A. Schoen, *Infinite periodic minimal surfaces without self-intersections*, NASA technical note TN D-5541, NASA, Washington D.C., May 1970

[19] R. Schoen, *Estimates for stable minimal surfaces in three-dimensional manifolds*, Seminar on minimal submanifolds, 111–126, Ann. of Math. Stud. **103**, Princeton Univ. Press, Princeton, NJ, 1983

[20] R. Schoen and S.-T. Yau, *Existence of incompressible minimal surfaces and the topology of three dimensional manifolds with non-negative scalar curvature*, Ann. Math. **110**, no. 1, (1979), pp. 127–142

[21] W. Thurston, *The Geometry and Topology of 3-Manifolds*, a.k.a "Thurston's Notes"; available from the MSRI at `http://www.msri.org/publications/books/gt3m`

[22] B. White, *Lectures on minimal surface theory*, preprint; arXiv:1308.3325

University of Chicago, Chicago, Ill 60637 USA
E-mail address: `dannyc@math.uchicago.edu`