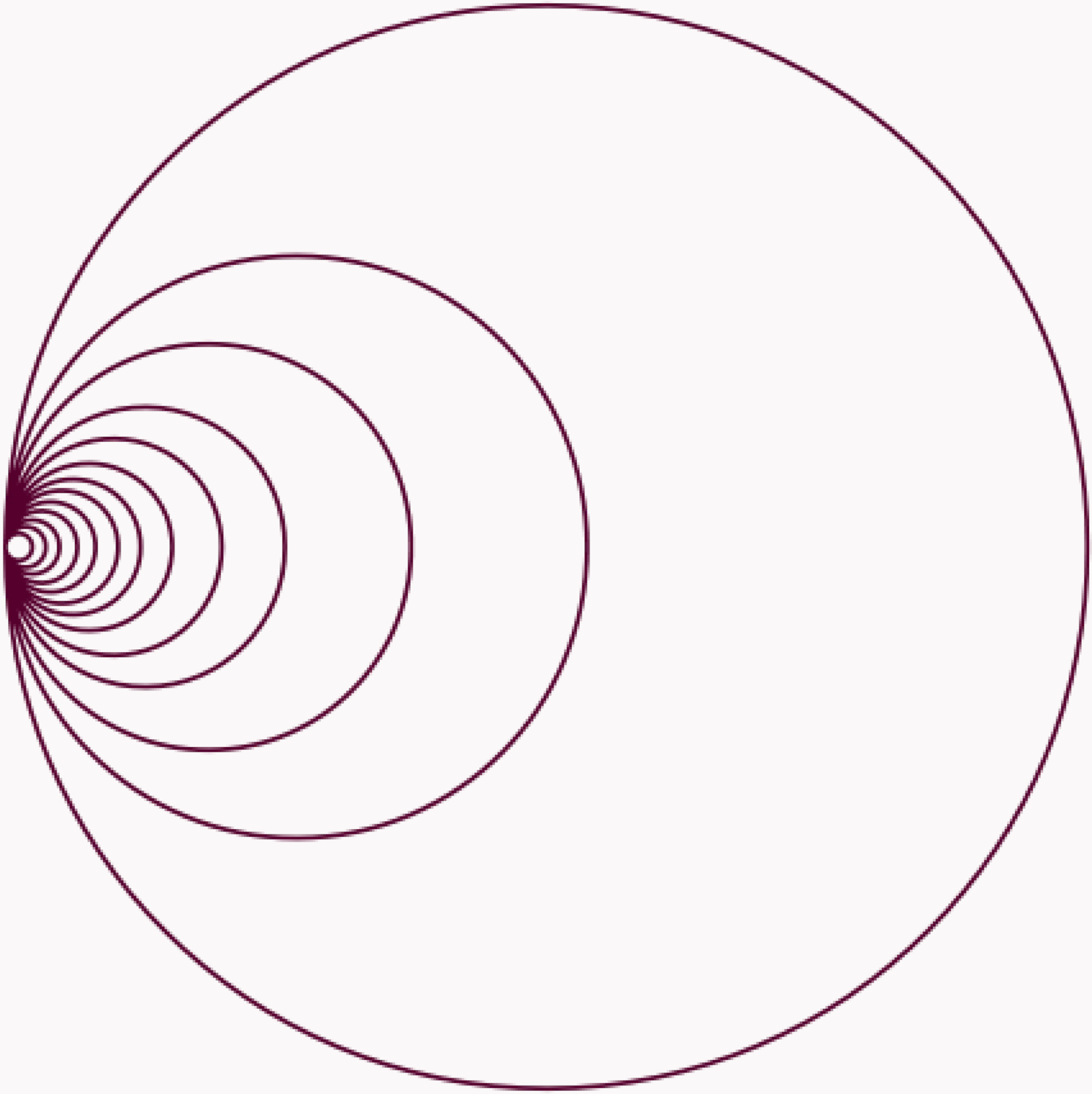


Columbia Journal *of* Undergraduate Mathematics

ISSUE 1 VOLUME 1 AUGUST 5, 2024

About the Journal

Aim and Scope. The primary goal of the *Columbia Journal of Undergraduate Mathematics* is to provide undergraduate readers with high-quality, accessible articles on challenging topics, or novel approaches to teaching more familiar concepts. Articles published are purely expository; we do not accept research papers. Most are under 20 pages in length, with the primary exceptions being senior theses written by students at Columbia and other universities alike. The journal also accepts and publishes mathematical artwork with clear pedagogical value.

Contact. Email columbiajournalofundergradmath@gmail.com with any questions, or visit https://journals.library.columbia.edu/index.php/cjum for more information.

Letter from the Editors

Undergraduate mathematics students are frequently tasked with writing expository pieces accessible to their peers. By the time they graduate, most math majors will have been assigned a paper on a topic not explored by a class they took, assembled a summary of a research niche they explored for an REU, or simply decided to write about a result or topic they felt they had a fresh way of talking about. The *Columbia Journal of Undergraduate Mathematics* was created when we asked ourselves a simple question about these projects: where do they go? The answer at the time seemed to be ‘nowhere’: since undergraduate expository work does not present original results or research, this writing is not eligible for most mathematics publications. We decided to change that. Our belief was that, since papers written *by* undergraduates themselves are uniquely understandable *for* undergraduates, such papers represent an untapped intellectual resource. We were confident that if we created a journal to showcase this work, we would uncover a treasure trove of creative and insightful mathematical exposition just waiting to come to light.

To our delight, our call for submissions was answered by a wide variety of talented undergraduate expositors originating from schools across the country (and one or two from abroad). Throughout the editing process, we were inspired by how these submissions offered novel ways of explaining known mathematical concepts with target audiences ranging from first-semester undergraduates to the most advanced graduating seniors. This issue offers the best of the best of those submissions.

Our opening article, “A Topological Proof of the Riemann-Hurwitz Formula,” approaches an algebro-geometric result through the lens of manifolds and algebraic topology, with exposition enhanced by diagrams drawn by the author herself. Without assuming much complex analysis, the paper introduces essential concepts in the study of Riemann surfaces and their branched covers, as well as the Euler characteristic, before proving the titular Riemann-Hurwitz formula. Our next piece, “Representations of Complex Tori and $\mathrm{GL}(2, \mathbb{C})$,” explores the representation theory of algebraic groups, culminating in a classification of the representations for complex tori and $\mathrm{GL}(2, \mathbb{C})$. Along the way, the paper introduces representation-theoretic tools such as Hopf algebras, weight space decompositions, and the theorems of the highest weight, and should be accessible to those with a minimal background in algebraic varieties and Lie groups. Returning to geometry, “The Gauss-Bonnet Theorem” explains a classical topology result while only assuming the reader has a knowledge of linear algebra and multivariable calculus. We recommend this article to introductory readers.

For more advanced readers, "The Peter-Weyl Theorem & Harmonic Analysis on S^n" assumes only group and integration theory to introduce the representation theory of topological groups and its relation to functional and harmonic analysis. The article culminates in a proof of the Peter-Weyl theorem, a characterization of all the representations of a compact group in terms of the square-integrable functions on it with applications to Fourier-type decompositions on spheres. Finally, "Elliptic bootstrapping and the non-linear Cauchy-Riemann equations" introduces the essential technique of elliptic bootstrapping in geometric analysis. Assuming knowledge of manifolds, L^p spaces, and some familiarity with partial differential equations and complex analysis, the paper discusses almost-complex and symplectic manifolds and introduces Sobolev spaces to prove a regularity theorem for J-holomorphic curves, with an explanation of their importance to moduli spaces in symplectic geometry.

Now as the editors-in-chief of this inaugural issue, we want to highlight our wonderful team of content editors, copy editors, graduate student editors, and our faculty advisors. Creating, editing, and reviewing a brand-new journal is no easy task, and we want to thank each of you for all the time and effort you have put into making this first issue. We are particularly thankful for the contributions of our Head Copy Editor Jazmyn Wang and Chief Confidentiality Officer J Xiang, without whom this issue would not exist. We would also like to acknowledge the support from the Columbia Libraries and Department of Mathematics, who were the technical and financial backbone to make this project come to life.

Finally, on behalf of the editorial team, we would like to thank every undergraduate student who submitted to our journal. We know firsthand the hard work and dedication required to create a novel piece of exposition, and we truly appreciate every submission we received, even if they did not make it into the final issue. In particular, we want to provide a special thanks to our published authors, who submitted inspiring work and worked tirelessly with our editorial and copy teams to turn already-amazing papers into the polished versions you see now. Our undergraduate editors grew tremendously as mathematical thinkers and writers from reading and editing these papers. We hope our readers will grow in the same way from reading this issue, discover a new favorite theorem or two, and perhaps be inspired to produce some undergraduate expository work of their own.

Sincerely,

Aiden Sagerman, Zachary Lihn, and Lisa Faulkner Valiente

Editors-in-Chief, *Columbia Journal of Undergraduate Mathematics*

Masthead

Contents

Page No.

Columbia Journal of Mathematics **1** (2024),
8–27

A topological proof of the Riemann–Hurwitz formula

By Mrinalini Sisodia Wadhwa

Abstract

The Riemann–Hurwitz formula is generally given as a result from algebraic geometry that provides a means of constraining branched covers of surfaces via their Euler characteristic. By restricting to the special case of compact Riemann surfaces, we develop an alternative proof of the formula that draws on topology and manifold theory as opposed to more advanced algebraic machinery. We first discuss the foundation in manifold theory, defining Riemann surfaces and providing an example of the complex projective line. We then discuss the local topological structure of holomorphic maps between Riemann surfaces, introducing the notion of a *branched cover* and of *branch points*. Next, we discuss triangulations of a topological space and use this to introduce the Euler characteristic of Riemann surfaces. Using these definitions, we explicate and prove the Riemann–Hurwitz formula on compact Riemann surfaces. To conclude, we discuss consequences of this formula for adjacent fields such as algebraic topology. We provide visual intuition and examples throughout, drawing primarily on Szameuly's *Galois Groups and Fundamental Groups* (2009), as well as Forster's *Lectures on Riemann Surfaces* (1981), Guillemin and Pollack's *Differential Topology* (1974), and a few other supplementary sources. The main prerequisite for this paper is a background in topology and covering spaces.

1. **Introduction**

We begin with a topological problem. Suppose we have two surfaces, each with certain properties—such as holes, punctures, boundaries, and so forth—that are invariant under homeomorphism. We call such properties *topological invariants*. Can we always obtain a surjective map between these two surfaces that preserves their local structure? In fact, we cannot: as we will see at the end of this paper, there is no such surjective map from the complex projective plane (a surface with no holes) to the torus (a surface with one hole).

This prompts a second question: do the topological invariants of these surfaces tell us something about whether or not we can obtain such a map between the two surfaces? There is, in fact, an intricate relationship between the surfaces' topological invariants and the existence of a surjective map between them that preserves their structure. This relationship is given by the Riemann–Hurwitz formula, first proposed by Bernard Riemann (1826–66) in his 1857 *Theorie der Abel'schen Functionen* [*Theory of Abelian Functions*] [Rie57, §7]. We give a preliminary statement of the formula below and will define the terms used in the formula statement carefully in subsequent sections of the paper.

THEOREM 1.1 (Riemann–Hurwitz Formula). *Let $\varphi\colon Y \to X$ be a holomorphic map of compact Riemann surfaces with degree d as a branched cover. The Euler characteristics $\chi(X)$ and $\chi(Y)$ of X and Y are related by the formula*

$$\chi(Y) = d \cdot \chi(X) - \sum_{y} (e_y - 1),$$

where the sum is over the branch points of φ and e_y is the ramification index corresponding to each branch point $y \in Y$.

A *branched cover* is a particularly well-structured surjective map between Y and X, and the ramification index corresponds to "sheets" of the cover intersecting with one another. As seen in the above statement, Riemann was looking at a specific class of surfaces and maps between them—namely, Riemann surfaces and holomorphic maps, concepts he had introduced in his 1851 doctoral dissertation that now serve as the foundation for the field of complex analysis. He appears to have died without offering a proof of this formula [Oor16, p.568–69]. The first proof was likely Adolf Hurwitz's (1858–1919) argument in his 1891 paper, *Über Riemann'sche Fläche mit gegebenen Verzweigungspunkten* [*On Riemann surfaces with Given Branch Points*] [Hur91, p.375–76].

This formula has subsequently been generalized to an algebraic-geometric version that takes X and Y to be smooth curves (rather than Riemann surfaces) and φ to be a morphism between them (rather than a holomorphic map) [Oor16, p.573–74]. It is in this abstract form—within the context of algebraic geometry—that most students now encounter the Riemann–Hurwitz formula. The usual proof of this version of the formula, given in [Sta18, Tag 0C1B], relies upon spectral sequences and other abstract-algebraic machinery. To avoid getting lost in the thickets of algebraic geometry, we will restrict to original case of Riemann surfaces and holomorphic maps between them. From this, we can develop a proof of the Riemann–Hurwitz formula that uses topology and manifold theory, providing the elusive topological and geometric intuition for the formula that draws us back to our motivating problem—how to understand the relationship between a surjective map between two surfaces and their topological invariants. The goal of this paper is to offer such a proof, following [Sza09, §3.6].

We proceed in four sections. Section 2 provides a foundation in manifold theory, defining Riemann surfaces and discussing the complex projective line as an example. Section 3 discusses holomorphic maps between Riemann surfaces and their local topological structure, introducing the notion of branch points and a branched cover. Section 4 discusses triangulation and the Euler characteristic of Riemann surfaces. Finally, Section 5 completes the proof of the Riemann–Hurwitz formula and discusses some interesting corollaries for algebraic topology.

This paper assumes a background in topology—specifically point-set topology and covering spaces. The reader does not need an extensive background in manifold theory or complex analysis. Rather, the relevant concepts from these fields—specifically holomorphisms and complex manifolds—are explained in the Section 2 with reference to the real case, identifying $\mathbb{R}^2$ with $\mathbb{C}$. It should also be noted that because complex analysis is not the main focus of this paper, we either assume or sketch complex-analytic results as needed to complete the major proofs in this paper, particularly in Section 4. Wherever possible, we provide pictures and visual intuition for definitions and proofs.

2. Riemann surfaces

This section grounds this paper in the relevant manifold theory, drawing on [For07, p.1–12] and [Sza09, §3.1–3.2]. We build up to a definition of Riemann surfaces and discuss some examples.

We begin by defining a holomorphic map between subsets of $\mathbb{C}^n$ and a complex atlas on a manifold.

Definition 2.1. Let $U \subset \mathbb{C}^n$ and $V \subset \mathbb{C}^m$. A map $f\colon U \to V$ is **holomorphic** if, for every $x \in U$, there exists a neighborhood $U_x \subset U$ of x such that f is complex-differentiable everywhere in U_x.

This is the complex analogue of a smooth map in real analysis, although being holomorphic is a vastly stronger condition: a complex-differentiable map is both infinitely differentiable and analytic (unlike in the real case, where C^1 maps are not necessarily C^∞, and where C^∞ maps are not necessarily analytic).

Definition 2.2. Let X be a topological 2-manifold. A **complex chart** on X is a pair $(U_i \subset X, f_i\colon U_i \to f_i(U_i) \subset \mathbb{C})$ such that U_i is an open subset of X and f_i is a homeomorphic mapping from U_i onto its image $f(U_i) \subset \mathbb{C}$.

We say a chart (U_i, f_i) is **centered** at $x \in U_i$ if $f_i(x) = 0$. Two charts (U, f), (V, g) are **holomorphically compatible** if their transition maps $f \circ g^{-1}$ and $g \circ f^{-1}$ are holomorphic where defined. This is illustrated in Figure 1.

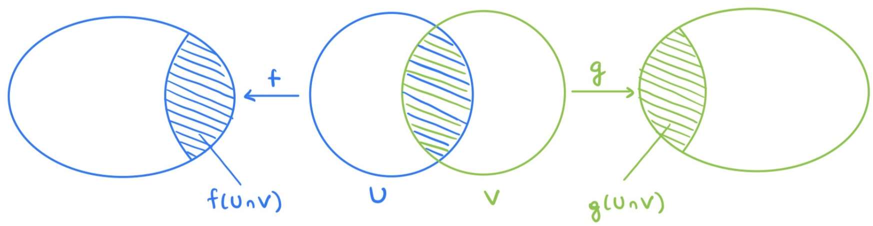


Figure 1. Two charts (U, f) and (V, g), with transition map $f \circ g^{-1}$ defined on $g(U \cap V)$ and $g \circ f^{-1}$ defined on $f(U \cap V)$, based on the illustration of the real case in [Tu07, §5.2, Fig 5.2].

Definition 2.3. A **complex atlas** $\mathcal{U}$ on X is a collection of holomorphically compatible charts (U_i, f_i) such that the $\{U_i\}$ form an open cover of X.

We say two atlases (U_i, f_i), (V_j, g_j) on X are **equivalent** if their union, defined by taking all U_i and V_j as a covering of X and all complex charts, is also a complex atlas on X. In particular, this implies that $f_i \circ g_j^{-1}$ and $g_j \circ f_i^{-1}$ are holomorphic on their respective domains for all i, j. We now proceed to define a Riemann surface by placing a complex structure on X, in a manner analogous to how [Tu07, §2.5] discusses placing a smooth structure in the real case.

Definition 2.4. A **Riemann surface** is a topological 2-manifold X with an equivalence class of complex atlases (which we call a **complex structure** on X).

As a trivial example, consider any open subset $U \subset \mathbb{C}$. Then U is a Riemann surface with the complex atlas $(U, i\colon U \hookrightarrow \mathbb{C})$, where i is the inclusion map. We consider one nontrivial example, the complex projective line, which we return to in subsequent sections of this paper.

Example 2.5 (Complex projective line $\mathbb{CP}^1$). Let $\mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$, where ∞ is an extra point not included in $\mathbb{C}$. We topologize $\mathbb{CP}^1$ as follows: the open sets are the usual open sets $U \subset \mathbb{C}$ and sets of the form $V \cup \{\infty\}$, where $V \subset \mathbb{C}$ is the complement of a compact set $K \subset \mathbb{C}$. We call $\mathbb{CP}^1$ with this topology the **complex projective line**, and we see that it is homeomorphic to the 2-sphere $S^2 \subset \mathbb{R}^3$ with antipodal points identified with 0 and ∞, as shown in Figure 2.

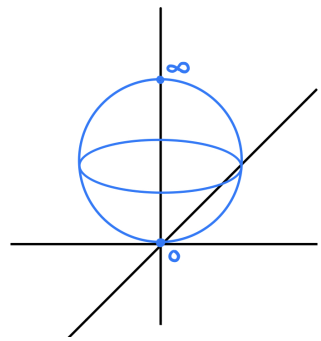


Figure 2. $\mathbb{CP}^1$ homeomorphic to S^2.

Now we define a complex atlas on $\mathbb{CP}^1$. Let $U_1 := \mathbb{CP}^1 \setminus \{\infty\} = \mathbb{C}$, and let f_1 be the identity map z on U_1. Then, let $U_2 := \mathbb{CP}^1 \setminus \{0\}$, and define the map f_2 as follows:

$$f_2(z) = \begin{cases} \frac{1}{z} & z \in U_2 \setminus \{\infty\} \\ 0 & z = \infty. \end{cases}$$

Then both f_1 and f_2 are well-defined homeomorphisms onto their images. The charts $(U_1, f_1), (U_2, f_2)$ cover $\mathbb{CP}^1$ and are holomorphically compatible, as their transition maps

$$f_1 \circ f_2^{-1} = f_2 \circ f_1^{-1} \colon U_1 \cap U_2 = \mathbb{C} \setminus \{0\} \to U_1 \cap U_2 = \mathbb{C} \setminus \{0\}$$

are given by $z \mapsto \frac{1}{z}$. Thus, they form a complex atlas on $\mathbb{CP}^1$.

If we consider $\mathbb{CP}^1$ under the homeomorphism to S^2, then the maps f_1 and f_2 correspond to the stereographic projection from $\mathbb{R}^2 \times \{0\}$ and $\mathbb{R}^2 \times \{1\}$, respectively, as shown in Figure 3.

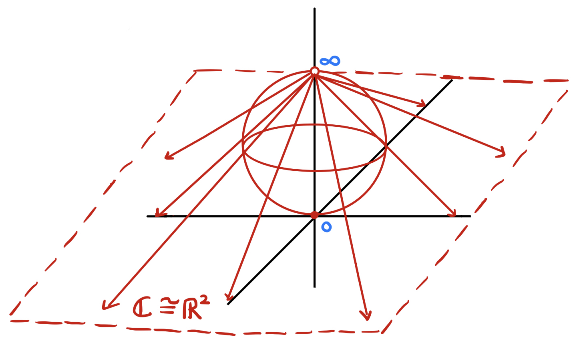
◊

3. Holomorphic maps and branched covers

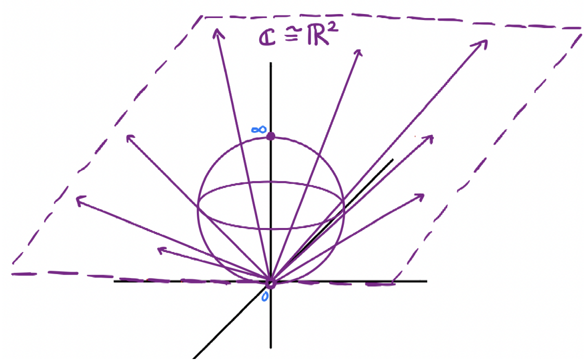
We now define holomorphic maps between Riemann surfaces, discuss their local topological structure, and introduce the notion of a branched cover.

Definition 3.1. Let X and Y be Riemann surfaces. A continuous map $\varphi \colon Y \to X$ is **holomorphic** if for every pair of charts (U, f) on Y and (V, g) on X such that $\varphi(U) \subset V$, the map $g \circ \varphi \circ f^{-1} \colon f(U) \subset \mathbb{C} \to g(V) \subset \mathbb{C}$ is holomorphic in the usual sense (i.e., the sense of Definition 2.1).

This definition is visualized in Figure 4. Note that this corresponds to our definition of a smooth map between smooth manifolds in the real case in [Tu07, §2.6].



(a) The chart (U_1, f_1) on $\mathbb{CP}^1$.



(b) The chart (U_2, f_2) on $\mathbb{CP}^1$.

Figure 3. A complex atlas on $\mathbb{CP}^1$, visualized under homeomorphism to S^2.

Henceforth, to avoid the trivial case, we assume that all holomorphic maps between Riemann surfaces in this paper are nonconstant on all connected components—i.e., that they do not map an entire connected component to a single point. We remarked after Definition 2.1 that being holomorphic is a stronger condition than being smooth, as holomorphic maps are both infinitely differentiable and analytic on subsets of $\mathbb{C}^n$. As we shall see, this means that we actually know a great deal more about the local structure of holomorphic maps than we do about smooth maps in the real case. This is summarized in the below proposition, which tells us that locally, every holomorphic map is just exponentiation.

Proposition 3.2. *Let $\varphi\colon Y \to X$ be a holomorphic map of Riemann surfaces and $y \in Y$ with image $\varphi(y) = x$ in X. Then there exist open neighborhoods $V_y \subset Y$ and $U_x \subset X$ of y and x respectively satisfying $\varphi(V_y) \subset U_x$,*

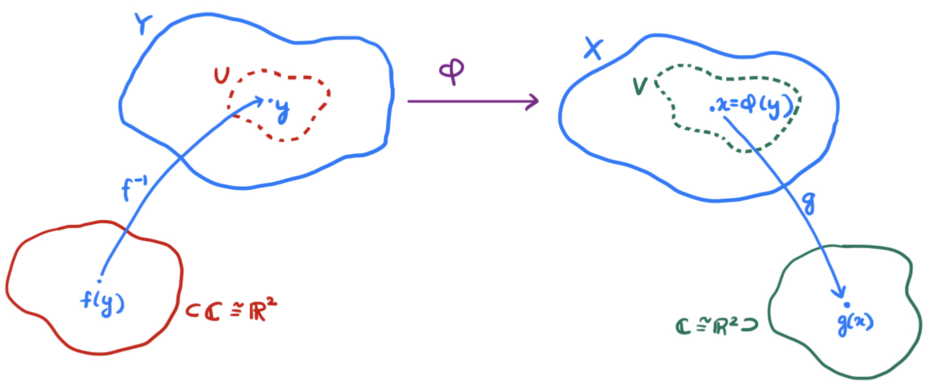


Figure 4. A holomorphic map $\varphi\colon Y \to X$ with charts (U, f) on Y and (V, g) on X, based on the illustration of the real case in [Tu07, §2.6, Figure 6.3].

as well as homeomorphisms $g_y\colon V_y \to g_y(V_y) \subset \mathbb{C}$ *and* $f_x\colon U_x \to f_x(U_x) \subset \mathbb{C}$ *satisfying* $f_x(x) = g_y(y) = 0$ *such that the diagram*

$$\begin{array}{ccc} V_y & \xrightarrow{\varphi} & U_x \\ {\scriptstyle g_y}\downarrow & & \downarrow{\scriptstyle f_x} \\ \mathbb{C} & \xrightarrow[z \mapsto z^{e_y}]{} & \mathbb{C} \end{array}$$

commutes for an appropriate positive integer e_y *chosen with respect to* y *that does not depend upon the choice of* g_y *or* f_x.

Figure 5 provides a geometric visualization of the commutative diagram. We proceed to sketch its proof, drawing on some results from complex analysis.

Proof sketch of Proposition 3.2. First, by selecting and shrinking neighborhoods U_x and V_y as necessary and performing linear transformations in $\mathbb{C}$, we can find charts (V_y, g'_y) and (U_x, f_x) centered at y and x, respectively. We will now modify these in order for the diagram to commute. As φ is a holomorphism from Y to X, we know by Definition 3.1 that $f_x \circ \varphi \circ g'^{-1}_y$ is holomorphic in a neighborhood of 0 and vanishes at 0. As holomorphic maps are necessarily analytic, complex analysis tells us that $f_x \circ \varphi \circ g'^{-1}_y$ must be of the form $z \mapsto z^{e_y} H(z)$, where H is a holomorphic function such that $H(0) \neq 0$.

We denote by log a fixed branch of the logarithm function in a neighborhood of $H(0)$. Now we apply complex analysis results to conclude: we shrink the neighborhood V_y as necessary so that $h := \exp((1/e_y)\log H)$ defines a holomorphic function h on $g'_y(V_y)$ such that $h^{e_y} = H$, and then we define g_y to be the composition of g'_y with the map $z \mapsto zh(z)$. This yields charts (V_y, g_y) and (U_x, f_x) centered at y and x respectively such that the diagram commutes.

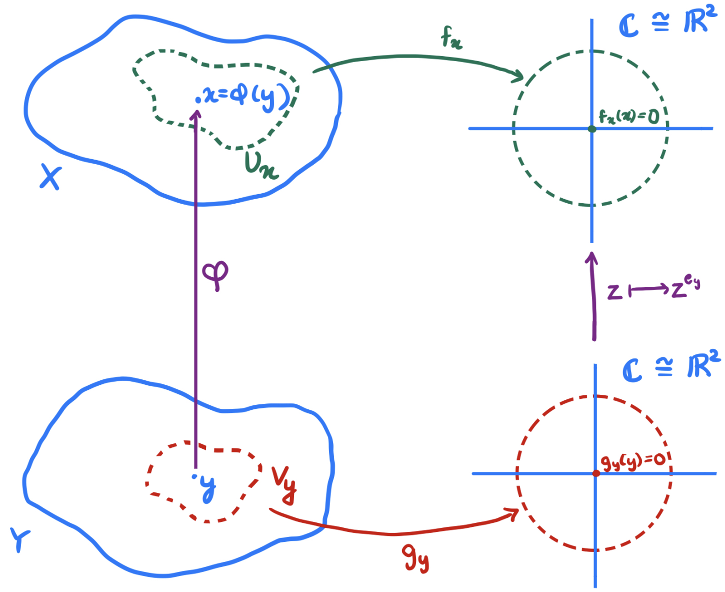


Figure 5. A geometric visualization of the local structure on holomorphic maps.

We observe moreover that e_y, defined in relation to an invertible holomorphic map, is necessarily a positive integer independent of the choice of g_y, f_x. □

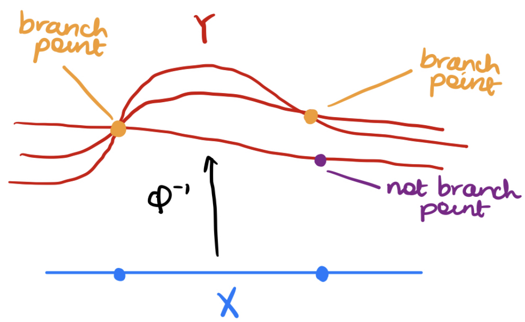
Having established this local structure, we introduce the notions of ramification index, branch points, and branched cover, following [Sza09, §3.2].

Definition 3.3. The positive integer e_y in Proposition 3.2 is called the **ramification index** of φ at y. The points $y \in Y$ such that $e_y > 1$ are called the **branch points** of φ. We denote the set of branch points of φ by S_φ.

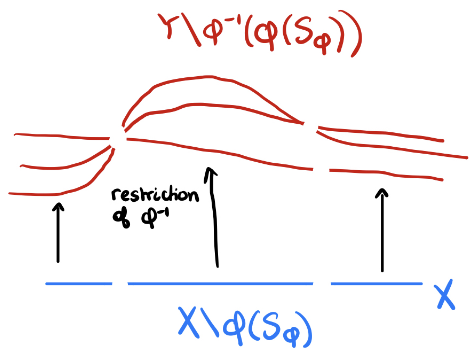
Remark 3.4. Note that S_φ is a discrete closed subset of Y. This follows from Proposition 3.2: given any $y \in Y$, there exists a punctured open neighborhood V_y of y that contains no branch points where φ has finitely many points in its preimage (due to the local structure of the map $z \mapsto z^{e_y}$).

From this observation, we proceed to introduce the notion of a branched cover and relate it to this local structure on holomorphic maps. First we must define a proper map.

Definition 3.5. A continuous map of locally compact topological spaces $\varphi\colon N \to M$ is **proper** if the preimage of each compact subset of M under φ is compact in N.



(a) A finite branched cover φ.



(b) Restricting φ to obtain a covering space.

Figure 6. Visualization of a branched cover.

Definition 3.6. Given locally compact Hausdorff spaces N and M, a proper surjective map $\varphi\colon N \to M$ is a **finite branched cover** if it restricts to a finite cover (of M) outside a discrete closed subset (of N).

Its **degree** is defined to be the degree of the finite cover obtained by its restriction.

We can think of a finite branched cover as essentially a covering space at all but a small number of points (namely, the branch points, which lie within the discrete closed subset). At the branch points, we can visualize the sheets of the cover merging together, so that the sheets of the cover "branch out" from them. Thus, when we remove these points, we obtain a covering space in its regular topological sense, as shown in Figure 6.

Finally, we relate this notion of a branched cover to holomorphic maps between Riemann surfaces with the following rather wonderful result.

THEOREM 3.7. *Let X be a connected Riemann surface, and let $\varphi\colon Y \to X$ be a proper holomorphic map. Then φ is a finite branched cover.*

Proof of Theorem 3.7. This result follows from Proposition 3.2 and Remark 3.4.

First, by definition, X and Y are locally compact Hausdorff spaces because they are Riemann surfaces.

Second, we claim φ is surjective because it is holomorphic and proper. Proposition 3.2 implies that as a holomorphic map between Riemann surfaces, φ is in fact an open map, since the map $z \mapsto z^{e_y}$ is open and f_x, g_y are homeomorphisms and therefore open maps. Thus $\varphi(Y)$ is open in X. Moreover, because φ is proper and X, Y are Hausdorff and locally compact, φ is a closed map, because in a locally compact Hausdorff space a subset is closed if and only if its intersection with every compact subset is closed. Thus $\varphi(Y)$ is closed in X. Then $\varphi(Y)$ is a nonempty clopen subset of X, so we must have $\varphi(Y) = X$ as X is connected, proving that φ is surjective.

Third, we have from Remark 3.4 that S_φ is a discrete closed subset of Y.

To conclude, we claim that the restriction of the map φ to $Y \setminus \varphi^{-1}(\varphi(S_\varphi))$ is a finite topological cover of $X \setminus \varphi(S_\varphi)$. This follows again from Proposition 3.2: given $x \in X \setminus \varphi(S_\varphi)$, each of the finitely many points in the preimage $\varphi^{-1}(x)$ has an open neighborhood that maps homeomorphically onto an open neighborhood of x. The intersection of these open neighborhoods is an open neighborhood of x that satisfies the definition of a finite topological cover (demonstrated in Figure 6). □

In light of this result, we will now take as a given that a holomorphic map φ as above yields a finite branched cover in subsequent sections of this paper.

4. Triangulation of Riemann surfaces

This section completes the setup for our topological proof of the Riemann–Hurwitz formula, carefully defining and providing geometric intuition for the various terms used in the formula statement. We define triangulation on a compact topological 2-manifold (and thus on any Riemann surface), prove that every compact Riemann surface has a triangulation, and introduce the concept of the Euler characteristic of a compact Riemann surface.

Intuitively, a triangulation divides up a space into smaller "triangles"—closed subsets of the space that map homeomorphically onto unit triangles in $\mathbb{R}^2$—that are glued together at edges or vertices. We formalize this notion below.

Definition 4.1. Let X be a compact topological 2-manifold. A **triangulation** of X consists of a finite system $\mathcal{T} = \{T_1, \ldots, T_n\}$ of closed subsets of

X whose union is the whole of X, and homeomorphisms $\varphi_i \colon \Delta \to T_i$, where Δ is the unit triangle in $\mathbb{R}^2$.

We say that the T_i are the **faces** of the triangulation, and that the images of the vertices (respectively edges) under φ_i of Δ are the **vertices** (respectively **edges**) of the triangulation. These must satisfy the following conditions:

(1) Each vertex (respectively edge) of $\mathcal{T}$ contained in a face T_i should be the image of a vertex (respectively edge) of Δ under φ_i;
(2) Any two different faces must either be disjoint, or intersect at a single vertex, or intersect at a single edge.

As an example, we consider a triangulation on the 2-sphere S^2, the underlying topological structure for the complex projective line $\mathbb{CP}^1$ discussed in Example 2.5.

Example 4.2 (Triangulation on the 2-sphere). By cutting S^2 along the equator and two meridians, we obtain a triangulation $\mathcal{T}$ with 6 vertices, 8 faces, and 12 edges. Figure 7 provides a visualization of $\mathcal{T}$ and the homeomorphic map from the unit triangle to one of its closed subsets T_1. Since S^2 is homeomorphic to $\mathbb{CP}^1$, this implies that there is a corresponding triangulation of $\mathbb{CP}^1$.

$\diamondsuit$

PROPOSITION 4.3 (Refinement of a triangulation). *Given a particular triangulation $\mathcal{T}$ of a compact topological space X and a point $x \in X$ that is not a vertex of $\mathcal{T}$, we can refine $\mathcal{T}$ to include x as a vertex.*

Proof of Proposition 4.3. There are two cases: either x lies in the interior of a face of $\mathcal{T}$ or it lies on an edge of $\mathcal{T}$.

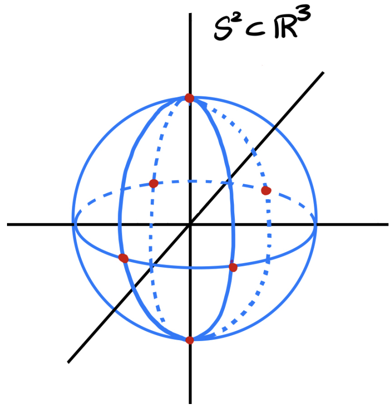
Case 1: Take the face $\varphi_i(\Delta)$ that contains x, and consider the natural subdivision of Δ that arises from joining $\varphi_i^{-1}(x)$ to the vertices and replace φ_i with its restrictions to the smaller triangles Δ_1, Δ_2 and Δ_3 that arise from the subdivision (where each Δ_i is homeomorphic to the unit triangle Δ in $\mathbb{R}^2$).

Case 2: Take the two faces $\varphi_i(\Delta)$ and $\varphi_j(\Delta)$ that meet at the edge on which x lies, and repeat the same process, considering the natural subdivision of Δ that arises from joining $\varphi_i^{-1}(x) = \varphi_j^{-1}(x)$ to the vertices and replace φ_i and φ_j with their restrictions to the smaller triangles $\Delta_1, \Delta_2, \Delta_3$ and Δ_4 (where, likewise, each Δ_i is homeomorphic to the unit triangle Δ in $\mathbb{R}^2$).

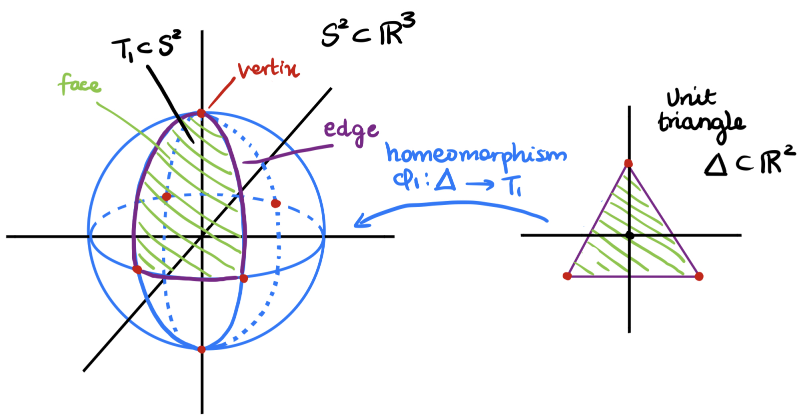
This process is illustrated in Figure 8.

$\square$

We will now prove an important result, following [Sza09, §3.6], which will set up our definition of the Euler characteristic.



(a) A triangulation $\mathcal{T}$of S^2.



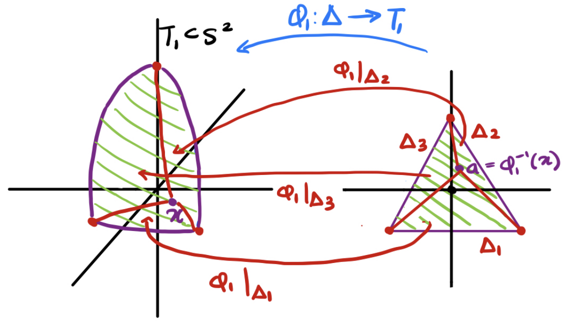
(b) Homeomorphic mapping of the unit triangle Δ to $T_1 \in \mathcal{T}$.

Figure 7. Illustration of Example 4.2.

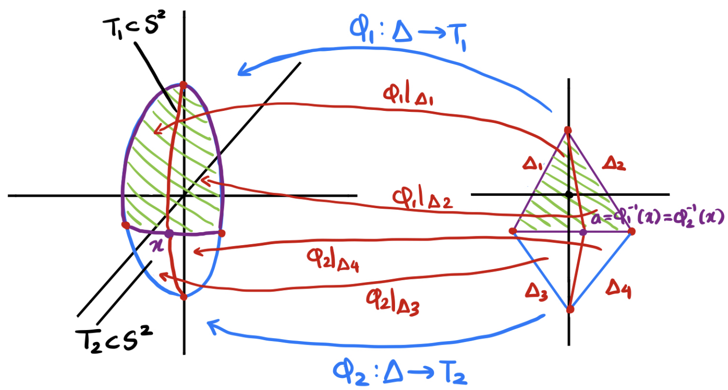
THEOREM 4.4. *Every compact Riemann surface has a triangulation.*

To prove this theorem, we begin with an arbitrary compact Riemann surface, and use results from complex analysis and topology to reduce this to the case of $\mathbb{CP}^1$, for which we know there is a triangulation by Example 4.2. We proceed in three steps. First, we show that a triangulation can be canonically lifted via a finite branched cover. Second, we sketch a proof using complex analysis that any compact Riemann surface yields a finite branched cover of $\mathbb{CP}^1$. Finally, we relate these findings and Example 4.2 to conclude.

LEMMA 4.5. *Let $\varphi\colon Y \to X$ be a finite branched cover of compact Riemann surfaces Y and X (in particular, following Theorem 3.7, consider the*



(a) Case 1, where x does not lie on an edge of $\mathcal{T}$.



(b) Case 2, where x lies on an edge of $\mathcal{T}$.

Figure 8. Two possible cases for refining a triangulation.

case where X is connected and φ is a proper surjective holomorphic map $Y \to X$). Then every triangulation of X can be lifted canonically to a triangulation of Y.

Proof of Lemma 4.5. Take some triangulation $\mathcal{T}$ on X with faces $\{T_i\}$ and homeomorphisms $\psi_i : \Delta \to T_i$, and let S_0 be the set of all vertices of $\mathcal{T}$. Following the process described in Proposition 4.3, we can refine $\mathcal{T}$ as necessary so that S_0 contains all images of branch points, i.e., for every $x \in X$ such that $x = \varphi(y)$ for some $y \in S_\varphi$, we have $x \in S_0$. Then the definition of a finite branched cover implies that the restriction of φ to $X \setminus \varphi^{-1}(S_0)$ is a cover.

Let Δ' be the subset of Δ obtained by omitting all vertices. We observe that Δ' is simply connected because it is contractible: as the triangle is filled-in and convex, we can take the straight-line homotopy to contract it to its center point. Then the fundamental group of Δ' is trivial and thus the restriction of the branched cover $\varphi\colon Y \to X$ above each of $\psi_i(\Delta')$ is trivial. This implies that we can canonically lift the restriction of each ψ_i to Δ' to each sheet of the cover $X \setminus \varphi^{-1}(S_0)$. We can also canonically lift all vertices of $\psi_i(\Delta)$ that are not the images of branch points. This demonstrates that the triangulation of X gives rise to a triangulation of Y away from the branch points.

It remains to show that we also have a triangulation of Y at the branch points. We revisit Proposition 3.2 regarding the local structure of holomorphic maps to consider the behaviour of φ near branch points. For any branch point $y \in Y$, Proposition 3.2 implies that there is a neighborhood of y on which φ locally looks like the continuous open map $z \mapsto z^{e_y}$. Then we can apply the process outlined in Proposition 4.3 to refine the triangulation again, adding each branch point y as a vertex. This yields the desired triangulation of Y.

□

This process of lifting a triangulation is illustrated in Figure 9, where we lift a triangle from X to Y via a branched cover of degree 3. We shall revisit this process and diagram in Section 5 while proving the Riemann–Hurwitz theorem.

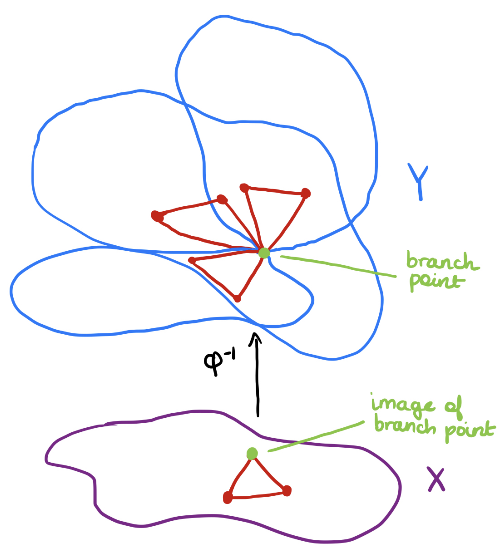


Figure 9. A triangulation of X lifts canonically to a triangulation of Y when $\varphi\colon Y \to X$ is a finite branched cover.

We move on to the second step of the proof of Theorem 4.4.

Lemma 4.6. *Given a connected compact Riemann surface Y, there exists a nonconstant holomorphic map $Y \to \mathbb{CP}^1$.*

The proof of this lemma utilizes two results from complex analysis that we will state below but not prove. Full proofs are given in [For07, Corollary 14.13, Theorem 1.8].

LEMMA 4.7 (Riemann's existence theorem). *Let X be a compact Riemann surface, $x_1, \dots, x_n \in X$ a finite set of points, and $a_1, \dots, a_n$ a sequence of complex numbers. Then there exists a function f on X that satisfies the following conditions*:

(1) *f is holomorphic everywhere on $X \setminus S$, where $S \subset X$ is a discrete closed subset, and for all complex charts $(U, \varphi\colon U \to \mathbb{C})$, the complex function $f \circ \varphi^{-1}$ is holomorphic everywhere except on a discrete closed subset of the domain*[1];
(2) *f is holomorphic at all the x_i, with $f(x_i) = a_i$ for all i from 1 to n.*

□

LEMMA 4.8 (Riemann's removable singularities theorem). *Let U be an open subset of a Riemann surface X, let $a \in U$, and let f be some function that is holomorphic on $U \setminus \{a\}$. Suppose f is bounded in some neighborhood of a. Then f can be extended uniquely to a function f' that is holomorphic on U.* □

Proof sketch of Lemma 4.6. Since Y is a compact Riemann surface, Lemma 4.7 gives a nonconstant function $f\colon Y \to \mathbb{C}$ that satisfies conditions (1) and (2) in its statement. We now define a map $\varphi_f\colon Y \to \mathbb{CP}^1 = \mathbb{C} \cup \{\infty\}$ as follows:

$$\varphi_f(y) = \begin{cases} f(y) & y \neq 0, \infty \\ \infty & y = 0, \infty. \end{cases}$$

We know f is holomorphic on all but a discrete set of points by condition (1), so given some $y \in Y$, we can choose a chart $(U, g\colon U \to \mathbb{C})$ centered around y such that f is holomorphic on $U \setminus \{y\}$ (shrinking U as necessary). Recall from Example 2.5 that the two standard complex charts on $\mathbb{CP}^1$ are given by z and $\frac{1}{z}$ on $\mathbb{CP}^1 \setminus \{\infty\} = \mathbb{C}$ and $\mathbb{CP}^1 \setminus \{0\}$, respectively. Now there are two cases. If f is holomorphic at y, then $z \circ \varphi_f \circ g^{-1}$ is holomorphic on $g(U)$. If f is not holomorphic at y, we apply Lemma 4.8: $(\frac{1}{z}) \circ \varphi_f \circ g^{-1}$ maps $g(U \setminus \{y\})$ to a bounded open subset of $\mathbb{C}$ and thus extends to a holomorphic function on $g(U)$. We conclude that φ_f is holomorphic. Moreover, since f is nonconstant, φ is also nonconstant by construction. □

[1] This is equivalent to f being a **meromorphic** function on X, which [Sza09, §3.3] discusses formally. We eschew further discussion of meromorphic functions here to avoid losing sight of the goal of this section: proving Theorem 4.4.

We finally piece together these results to complete the proof of Theorem 4.4.

Proof of Theorem 4.4. Take any compact Riemann surface Y, and consider its components (necessarily finitely many as Y is compact), which are connected compact Riemann surfaces $Y_1, \dots, Y_n$. By Lemma 4.6, there exist nonconstant holomorphic maps $\varphi_1, \dots, \varphi_n$ such that each φ_i maps Y_i into $\mathbb{CP}^1$. By Example 4.2, there is a triangulation on $\mathbb{CP}^1$. By Lemma 4.5, since each φ_i is a holomorphic map from a compact connected Riemann surface Y_i to $\mathbb{CP}^1$, another compact connected Riemann surface, we have that φ_i is a branched cover and that our triangulation of $\mathbb{CP}^1$ from Example 4.2 lifts to a triangulation of Y_i, say $\mathcal{T}_i$. We can then piece together these triangulations by taking their union to obtain a triangulation $\mathcal{T}$ on all of Y, completing the proof. □

Finally, we introduce the notion of an Euler characteristic, following the definition given in [Sza09, §3.6].

Definition 4.9. Given a triangulation $\mathcal{T}$ of a compact Riemann surface X, denote by S_0, S_1, and S_2 the set of vertices, edges, and faces of $\mathcal{T}$, respectively. Let s_0, s_1, and s_2 be their respective cardinalities. Then we define the **Euler characteristic** of X to be $\chi(X) := s_0 - s_1 + s_2$.

This is the classical definition of the Euler characteristic. It is in fact equivalent to the definition based on intersection theory in [GP78, p.116], though proving this equivalence casts beyond this paper's scope. Intuitively, this definition offers us a means of classifying compact Riemann surfaces based on their triangulations.

Note that we need Theorem 4.4 to ensure that the Euler characteristic is defined for all compact Riemann surfaces, as any compact Riemann surface must have a triangulation by Theorem 4.4 and therefore its Euler characteristic can be computed using the given formula. Moreover, the Euler characteristic is well-defined independent of the choice of triangulation on a given compact Riemann surface. To see this, notice that the Euler characteristic remains unchanged under the process of refining a triangulation described in Proposition 4.3 and illustrated in Figure 8. In both Case 1 (where we added x as a vertex when x was not on an edge) and Case 2 (where we added x as a vertex when x was on an edge), the Euler characteristic with the refined triangulation is $(s_0+1)-(s_1+3)+(s_2+2) = s_0 - s_1 + s_2$, the same as the original. Then, given any two triangulations of a compact Riemann surface, we can take their common refinement and thereby obtain the same value for its Euler characteristic throughout.

As an example, we return to the case of S^2, homeomorphic to $\mathbb{CP}^1$ as discussed in Examples 2.5 and 4.2.

Example 4.10. The triangulation $\mathcal{T}$ given in Example 4.2 and illustrated in Figure 7 has 8 faces, 6 vertices, and 12 edges, which implies that $\chi(S^2) = 6 - 12 + 8 = 2$. Indeed, any other triangulation of S^2 yields the same calculation for the Euler characteristic. For example, Figure 10 shows another triangulation $\mathcal{T}'$ of S^2 obtained by cutting along the equator and twice in the upper hemisphere. This triangulation has 4 faces, 4 vertices, and 6 edges, so again we compute $\chi(S^2) = 4 - 6 + 4 = 2$.

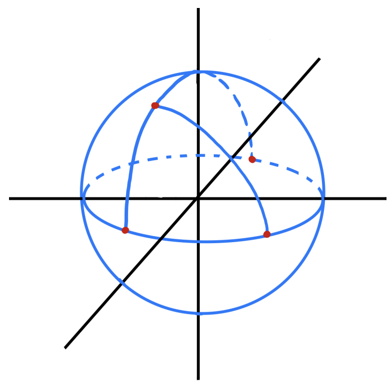


Figure 10. Another triangulation $\mathcal{T}'$ of S^2.

◊

5. **Proof of Riemann–Hurwitz formula**

Finally, we move to prove the Riemann–Hurwitz formula, given as Theorem 1.1 in the introduction, which we restate for convenience. Let $\varphi\colon Y \to X$ be a holomorphic map of compact Riemann surfaces with degree d as a branched cover. The Euler characteristics $\chi(X)$ and $\chi(Y)$ of X and Y are related by the formula

$$\chi(Y) = d \cdot \chi(X) - \sum_{y} (e_y - 1),$$

where the sum is over the branch points of φ and e_y is the ramification index corresponding to each branch point $y \in Y$.

The proof follows from closely revisiting the process of lifting a triangulation via a branched cover discussed in Lemma 4.5 and illustrated in Figure 9.

Proof of Theorem 1.1 (*The Riemann–Hurwitz formula*). Take any triangulation on X, and let s_0, s_1 and s_2 be the number of vertices, faces, and edges, respectively. Consider its canonical lifting to a triangulation of Y via the finite branched cover φ, given by the process outlined in the proof of Lemma 4.5. Notice that, by construction, all branch points $y \in Y$ of φ correspond to vertices of the lifted triangulation and therefore do not lie on edges or faces.

Thus, edges and faces are lifted canonically on the cover of degree d, so the number of edges and the number of faces of the lifted triangulation are equal to ds_1 and ds_2, respectively. For vertices on the lifted triangulation, there are two cases. Vertices that do not correspond to the images of branch points have d preimages as well, as the covering space is of degree d. However, at any branch point y, we have to account for the e_y sheets of the branched cover merging together, and thus the number of preimages is instead $d - (e_y - 1)$. Thus the number of vertices of the lifted triangulation can be written as $ds_0 - \sum_{y \in S_\varphi}(e_y - 1)$, and we can compute the Euler characteristic as follows:

$$\begin{aligned}\chi(Y) &= \left(ds_0 - \sum_{y \in S_\varphi}(e_y - 1)\right) - ds_1 + ds_2 \\ &= d(s_0 - s_1 + s_2) - \sum_{y \in S_\varphi}(e_y - 1) \\ &= d \cdot \chi(X) - \sum_{y \in S_\varphi}(e_y - 1).\end{aligned}$$

□

This is visually illustrated in the lifting of a triangle in Figure 9, where φ is a branched cover of degree 3. We notice that the 3 edges of the triangle in X are each lifted to 3 edges (for a total of 9 edges) in Y, and likewise that the 1 face of the triangle in X is lifted to 3 faces in Y. The 2 vertices in X that do not correspond to branch points are each lifted to 3 vertices in Y, but the vertex that corresponds to a branch point (with ramification index 3, as 3 sheets merge) is only lifted to $1 = 3 \cdot 1 - (3 - 1)$ vertex in Y. This gives a total of $7 = 3 \cdot 3 - (3 - 1)$ vertices in the lifted triangle, providing visual intuition for the proof.

This gives us the major result of this paper, and we conclude with a brief discussion of its implications. Because we are working with compact Riemann surfaces (instead of smooth curves in the algebraic geometry setting), we can apply some results from algebraic topology to restate the statement of Theorem 1.1 in more specific terms. In particular, any compact Riemann surface X is homeomorphic to a torus with g holes. The proof of this result, given in [Ful95, Theorem 17.4], utilizes the fact that compact Riemann surfaces are orientable topological 2-manifolds and a method of "cutting and pasting." We call g the **genus** of X, and can thus classify compact Riemann surfaces in terms of their genera, as depicted in Figure 11: compact Riemann surfaces of genus 0 are homeomorphic to S^2 and $\mathbb{CP}^1$, those of genus 1 are homeomorphic to the torus, those of genus 2 are homeomorphic to the 2-torus, and so forth.

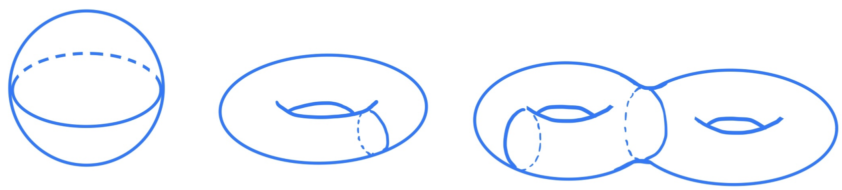


Figure 11. Tori with genera 0, 1, and 2, respectively.

Moreover, the genus of a compact Riemann surface gives us information about its Euler characteristic: a compact Riemann surface of genus g has Euler characteristic $2-2g$. This algebraic topology result, proven in [Ful95, p.244], follows by taking $g=0$ and $g=1$ as base cases and inducting on the genus. Note that we have already shown the $g=0$ case in Example 4.10, since we computed the Euler characteristic of $\mathbb{CP}^1$ to be $2=2-2\cdot 0$, where $g_{\mathbb{CP}^1}=0$. By restating Theorem 1.1 in these terms, we obtain the following corollary.

Corollary 5.1. *Let $\varphi\colon Y\to X$ be a holomorphic map of compact Riemann surfaces with degree d as a branched cover. Then*

$$2g_Y-2=d(2g_X-2)+\sum_y(e_y-1),$$

where the sum is over the branch points of φ, e_y is the ramification index corresponding to each branch point $y\in Y$, and g_X and g_Y are the genera of X and Y, respectively. □

This restatement of the Riemann–Hurwitz formula has a number of implications, one of which is discussed below, relating to our previous discussion of the case of $\mathbb{CP}^1$ in Examples 2.5 and 4.2.

Corollary 5.2. *If X is a compact Riemann surface of genus $g>0$, then there are no nonconstant holomorphic maps $\mathbb{CP}^1\to X$.*

Proof of Corollary 5.2. Suppose to the contrary that φ is a nonconstant holomorphic map $\mathbb{CP}^1\to X$. By Theorem 3.7, φ induces a branched cover, so by Corollary 5.1,

$$2g_{\mathbb{CP}^1}-2=d(2g-2)+\sum_y(e_y-1).$$

As $g_{\mathbb{CP}^1} = 0$, the left-hand side equals $2 \cdot 0 - 2 = -2$. But the right-hand side must be a positive value, as $g > 0$ by assumption, so

$$d(2g-2) + \sum_y (e_y - 1) > 0.$$

This gives us a contradiction, so no such φ can exist. □

This result is particularly interesting, as it reveals that the reverse of Lemma 4.6 does not hold: while, for any connected compact Riemann surface Y, we can have a nonconstant holomorphic map from Y into $\mathbb{CP}^1$, we cannot necessarily have a nonconstant holomorphic map out of $\mathbb{CP}^1$ into Y.

References

[For07] Otto Forster. *Lectures on Riemann Surfaces*. Springer, 2007.

[Ful95] William Fulton. *Algebraic Topology: A First Course*. Springer, 1995.

[GP78] Victor Guillemin and Alan Pollack. *Differential Topology*. AMS Chelsea Publishing, Providence, RI, 1978.

[Hur91] Adolf Hurwitz. Uber Riemann'sche Fläche mit gegebenen Verzweigungspunkten. *Mathematische Annalen*, 1891.

[Oor16] Frans Oort. The Riemann-Hurwitz Formula. *The Legacy of Bernhard Riemann After One Hundred and Fifty Years*, 2016.

[Rie57] Bernard Riemann. Theorie der Abel'schen Functionen. *Journal für die reine und angewandte Mathematik*, 1857.

[Sta18] The Stacks Project Authors. *Stacks Project*. https://stacks.math.columbia.edu, 2018.

[Sza09] Tamás Szamuely. *Galois Groups and Fundamental Groups*. Cambridge University Press, Cambridge, United Kingdom, 2009.

[Tu07] Loring W. Tu. *An Introduction to Manifolds*. Springer, 2007.

Department of Mathematics, Columbia University

E-mail: mw2706@columbia.edu

https://journals.library.columbia.edu/index.php/cjum

Columbia Journal of Mathematics **1** (2024),
28–42

Representations of complex tori and $\mathrm{GL}(2, \mathbb{C})$

By Songyu Ye

Abstract

Groups and their representations have been studied for a long time. One can extend the notion of a group by asking the group axioms to hold in other categories. A group in the category of smooth manifolds is a Lie group, and a group in the category of algebraic varieties is an algebraic group. In this paper, we discuss the representation theory of algebraic groups, in particular complex tori and GL(2,$\mathbb{C}$).

1. **Introduction**

The theme of this expository paper is to compare and contrast group objects in the settings of smooth manifolds and algebraic varieties. In particular, we begin by discussing the representation theory of tori in the smooth setting, and from our discussion it will become clear that some tools of Lie theory are not available to us in the algebraic setting. We remedy this by introducing different tools. One such tool we will introduce is the notion of a *Hopf algebra*, which axiomatizes the structure of the coordinate ring of an algebraic group. With a clear understanding of what is and is not available to us, we then discuss the representation theory of $(\mathbb{C}^*)^n$ and $\mathrm{GL}(2, \mathbb{C})$ in the algebraic setting.

2. **Representations of Tori**

2.1. *Real tori.* In this section, we study the representations of tori in the category of smooth manifolds. In particular, this means that the objects we are considering are smooth manifolds and the morphisms are smooth maps. A **real torus** T is a real Lie group which is isomorphic to the product of n circles. We say that T has **rank** n.

Let us first consider the case of $T = S^1$. We want to classify the finite-dimensional representations of S^1. It turns out that all finite-dimensional representations of S^1 are decomposable, i.e., can be written as a direct sum of irreducible representations.

Proposition 2.1. *Let K be a compact Lie group and let $\rho : K \to \mathrm{GL}(V)$ be a finite-dimensional complex representation. Then ρ is completely reducible.*

Proof sketch. The idea is to replicate the proof of Maschke's theorem for finite groups. Choose any inner product $\langle \cdot, \cdot \rangle$ on V and average over the group action to get a K-invariant inner product on V. In particular, put

$$\langle v, w \rangle_{\text{avg}} = \frac{1}{|K|} \int_K \langle \rho(k)v, \rho(k)w \rangle dk.$$

The existence of this inner product allows us to conclude that the orthogonal complement of a K-invariant subspace is also K-invariant. Inducting on the dimension of V allows us to completely decompose V into irreducible representations. ☐

We refer the reader to Chapter 9 of [FH91] for more detailed discussion.

Thus, it is enough to just consider the irreducible representations of S^1. By Schur's lemma (in particular, S^1 is abelian), they are all one-dimensional and therefore are indexed by characters $\chi : S^1 \to \mathbb{C}^*$. Since S^1 is compact, its image in $\mathbb{C}^*$ must also be compact; moreover, it is connected and contains the identity. Therefore, the image of χ must lie in S^1.

Proposition 2.2. *All characters of S^1 are isomorphic to $\chi_n : S^1 \to S^1$ given by $z \mapsto z^n$ for $n \in \mathbb{Z}$.*

Proof. Use the universal covering map $\exp : \mathbb{R} \to S^1$. Given a character $\chi : S^1 \to S^1$, we can lift it to a map $\tilde{\chi} : S^1 \to \mathbb{R}$. Since χ is a group homomorphism, it carries 1 to 1, and the fiber over 1 under exp is $\mathbb{Z}$. ☐

Since characters for $S^1 \times \cdots \times S^1$ are the same as products of characters for S^1, all characters of T are indexed by $\mathbb{Z}^n$. Explicitly, if T has rank n, then a character $\chi : T \to S^1$ is given by a tuple of integers $(n_1, \ldots, n_k)$, and the character is given by

$$(z_1, \ldots, z_k) \mapsto z_1^{n_1} \cdots z_k^{n_k}.$$

From our discussion above, we have the following classification statement for representations of real tori as Lie groups.

Theorem 2.3. *Let T be a real torus of rank n. Then every finite-dimensional representation V of T is isomorphic to a direct sum of one-dimensional irreducible representations with some multiplicities*

$$V \cong \bigoplus_{\chi \in \mathbb{Z}^n} W_\chi^{\oplus n_\chi},$$

where W_χ denotes the unique one-dimensional irreducible representation for which T acts by χ.

In particular, we can decompose V into eigenspaces for the action of T

$$V \cong \bigoplus_{\chi \in \mathbb{Z}^n} V_\chi,$$

where $V_\chi = \{v \in V \mid t \cdot v = \chi(t)v \text{ for all } v \in V \text{ and } t \in T\}$. This is referred to as the **weight space decomposition** of V, and we refer to the χ which appear in the decomposition as the **weights** of V. We say that $v \in V_\chi$ is a **weight vector** of weight χ.

2.2. *Complex tori.* We want an analagous story in algebraic geometry. To do so, we establish the following framework. Specifically, we are now dealing with the category of algebraic varieties over $\mathbb{C}$, where the objects are algebraic varieties and the morphisms are morphisms of algebraic varieties.

Definition 2.4. An **algebraic group** G over $\mathbb{C}$ is an algebraic variety over $\mathbb{C}$ with a group structure so that the multipliation map $G \times G \to G$ and the inversion map $G \to G$ are morphisms of algebraic varieties.

Definition 2.5. A **morphism of algebraic groups** $G \to H$ is a morphism of algebraic varieties that is also a group homomorphism.

Definition 2.6. Let G be an algebraic group. A **rational representation** of G is a morphism of algebraic groups $G \to \mathrm{GL}(V)$ for some vector space V. (For us, V will always be finite-dimensional over $\mathbb{C}$.)

We will consider complex algebraic tori $T = \mathbb{C}^* \times \cdots \times \mathbb{C}^*$. This is an algebraic group because T is the zero locus of the polynomial equations

$$\begin{aligned} T &= \operatorname{Spec} \mathbb{C}[x_1^{\pm 1}, \dots, x_n^{\pm 1}] \\ &:= \operatorname{Spec} \big(\mathbb{C}[x_1, \dots, x_n, y_1, \dots, y_n]/(x_1 y_1 - 1, \dots, x_n y_n - 1)\big). \end{aligned}$$

This is a group in the familiar way, and it is clear that the group law is indeed a morphism of algebraic varieties. The rest of this section will discuss the finite-dimensional rational representations of T as an algebraic group.

Remark 2.7. Why do we consider real tori as Lie groups and complex algebraic tori as algebraic groups? One good reason is that the real tori are not complex algebraic varieties. For example, there are no polynomials over $\mathbb{C}$ which define $S^1 \times S^1$ as a complex algebraic variety. Moreover, as Lie groups, the complex torus is the complexification of the real torus.

Example 2.8. $\mathrm{GL}(n, \mathbb{C})$ is a familiar group which can be endowed with the structure of an algebraic group. $\mathrm{GL}(n, \mathbb{C})$ is the zero locus of the polynomial equations

$$\mathrm{GL}(n, \mathbb{C}) = \mathrm{Spec}\left(\mathbb{C}[x_{ij}, \det^{-1}]\right) := \mathrm{Spec}\left(\mathbb{C}[x_{ij}, t]/(\det(x_{ij})t - 1)\right).$$

This variety becomes a group in the familiar way, and it is clear that the group law is indeed a morphism of algebraic varieties. ♢

The following theorem classifies the finite-dimensional rational representations of T as an algebraic group. The story is precisely that of the smooth manifold setting, but we introduce the language of Hopf algebras to demonstrate this.

THEOREM 2.9. *Let T be a complex torus of rank n. Then every finite-dimensional rational representation of T is isomorphic to a direct sum of one-dimensional irreducible representations with some multiplicities*

$$V \cong \bigoplus_{\chi \in \mathbb{Z}^n} W_\chi^{\oplus n_\chi},$$

where W_χ denotes the unique one-dimensional irreducible representation for which T acts by χ.

We will give an proof of this theorem after we introduce the language of Hopf algebras.

2.3. *Hopf algebras.* The notion of a Hopf algebra axiomatizes the structure of the ring of regular functions on an algebraic group. In particular, let G be an algebraic group and $\mathcal{O}(G)$ its ring of regular functions. Then the multiplication, inversion, and identity maps

$$\begin{aligned} \mu : G \times G &\to G \\ \iota : G &\to G \\ e : \mathrm{Spec}\, \mathbb{C} &\to G \end{aligned}$$

induce maps on the coordinate rings

$$\begin{aligned} \Delta : \mathcal{O}(G) &\to \mathcal{O}(G) \otimes \mathcal{O}(G) \\ \epsilon : \mathcal{O}(G) &\to \mathbb{C} \\ S : \mathcal{O}(G) &\to \mathcal{O}(G), \end{aligned}$$

where we made the identification $\mathcal{O}(G \times G) \cong \mathcal{O}(G) \otimes \mathcal{O}(G)$. Because the group axioms hold, these maps satisfy the following conditions and equip $\mathcal{O}(G)$ with the structure of a Hopf algebra.

Definition 2.10. Let A be a $\mathbb{C}$-algebra. Then we say A is a **Hopf algebra** if there are maps

$$\begin{array}{rl} \text{comultiplication} & \Delta : A \to A \otimes A \\ \text{counit (augmentation)} & \epsilon : A \to \mathbb{C} \\ \text{coinverse (antipode)} & S : A \to A \end{array}$$

so that the following diagrams commute:

$$\begin{array}{ccc} A & \xrightarrow{\Delta} & A \otimes A \\ \downarrow{\scriptstyle \Delta} & & \downarrow{\scriptstyle \Delta \otimes \mathrm{id}} \\ A \otimes A & \xrightarrow{\mathrm{id} \otimes \Delta} & A \otimes A \otimes A \end{array}$$

$$\begin{array}{ccc} A & \xrightarrow{\Delta} & A \otimes A \\ \downarrow{\scriptstyle \mathrm{id}} & & \downarrow{\scriptstyle \epsilon \otimes \mathrm{id}} \\ A & \xrightarrow{\cong} & \mathbb{C} \otimes A \end{array}$$

$$\begin{array}{ccc} A & \xrightarrow{\Delta} & A \otimes A \\ \downarrow{\scriptstyle \epsilon} & & \downarrow{\scriptstyle S \otimes \mathrm{id}} \\ \mathbb{C} & \longrightarrow & A \end{array} .$$

Remark 2.11. These maps can be worked out very explicitly. In particular, the points of G are in correspondence with the elements of $\mathrm{Hom}_{\mathrm{kAlg}}(\mathcal{O}(G), \mathbb{C})$. The correspondence can be written down explicitly as $g \mapsto \mathrm{ev}_g$, where $\mathrm{ev}_g : \mathcal{O}(G) \to \mathbb{C}$ is the evaluation map. The key idea is as follows. Let G be an arbitrary algebraic group G with x, y points of G, and write $f_x, f_y : \mathcal{O}(G) \to \mathbb{C}$ for the corresponding morphisms of $\mathbb{C}$-algebras. Then the composition $(f_x \otimes f_y) \circ \Delta$ is again a map $\mathcal{O}(G) \to \mathcal{O}(G) \otimes \mathcal{O}(G) \to \mathbb{C}$, and so we can ask if it is the map f_z corresponding to some $z \in G$. The condition that we require from comultiplication is precisely that the composition $(f_x \otimes f_y) \circ \Delta$ corresponds to the product $xy \in G$. In particular, the group law on G uniquely determines the comultiplication map on $\mathcal{O}(G)$.

Example 2.12. Recall that

$$\mathcal{O}(\mathfrak{G}_a(\mathbb{C})) = \mathbb{C}[x],$$

where $\mathfrak{G}_a(\mathbb{C})$ is the additive group of $\mathbb{C}$. Let $f, g \in \mathrm{Hom}_{\mathrm{kAlg}}(\mathcal{O}(G), \mathbb{C})$ with $f(x) = a$ and $g(x) = b$. We want to find a map

$$\Delta : \mathcal{O}(\mathfrak{G}_a(\mathbb{C})) \to \mathcal{O}(\mathfrak{G}_a(\mathbb{C})) \otimes \mathcal{O}(\mathfrak{G}_a(\mathbb{C}))$$

so that

$$((f \otimes g) \circ \Delta)(X) = (a + b).$$

We write down the map Δ explicitly as

$$\Delta(X) = X \otimes 1 + 1 \otimes X$$

and notice that it does the job. We see that Δ then must be the comultiplication map for $\mathfrak{G}_a(\mathbb{C})$ because such a map is unique (see the above remark), given the prescribed group law on $\mathfrak{G}_a(\mathbb{C})$. ◊

Example 2.13. By the same token, we can work out the Hopf algebra structure for $\mathbb{C}^*$ to be

$$\begin{aligned} \Delta(x) &= x \otimes x \\ \epsilon(x) &= 1 \\ S(x) &= x^{-1}. \end{aligned}$$

◊

Example 2.14. Consider the example of $\mathrm{GL}(2, \mathbb{C})$ as an algebraic group. The Hopf algebra structure is given by

$$\begin{aligned} \Delta(x_{ij}) &= \sum_{k=1}^{2} x_{ik} \otimes x_{kj} \\ \epsilon(x_{ij}) &= \delta_{ij} \\ S(x_{ij}) &= M_{ij}, \end{aligned}$$

where

$$M = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}^{-1} = \frac{1}{\det(M)} \begin{bmatrix} x_{22} & -x_{12} \\ -x_{21} & x_{11} \end{bmatrix}.$$

◊

Now we want to translate the representation theory of algebraic groups G into the language of comodules over Hopf algebras.

Theorem 2.15. *Let G be an algebraic group. Then rational representations V of G correspond to linear maps $\rho : V \to V \otimes \mathcal{O}(G)$ so that the following diagrams commute*:

$$\begin{array}{ccc} V & \xrightarrow{\rho} & V \otimes \mathcal{O}(G) \\ \downarrow{\scriptstyle \rho} & & \downarrow{\scriptstyle \rho \otimes \mathrm{id}} \\ V \otimes \mathcal{O}(G) & \xrightarrow{\mathrm{id} \otimes \Delta} & V \otimes \mathcal{O}(G) \otimes \mathcal{O}(G) \end{array}$$

$$\begin{array}{ccc} V & \xrightarrow{\rho} & V \otimes \mathcal{O}(G) \\ \big\downarrow \text{id} & & \big\downarrow \text{id} \otimes \epsilon \\ V & \xrightarrow{\cong} & V \otimes \mathbb{C} \end{array} .$$

Proof sketch. The first diagram says that the action of G on V is associative and the second diagram says that $e \in G$ acts by the identity transformation on V. These are precisely the conditions that say that V is a representation of G. □

We refer the reader to [Wat79] for a more detailed discussion of this theorem.

Definition 2.16. We call ρ a **comodule structure** on V.

Example 2.17. Consider the action of $\mathbb{C}^*$ on $\mathbb{C}^2$ given by

$$t \cdot (a, b) = (ta, t^{-1}b).$$

This is a rational representation of $\mathbb{C}^*$ which we can write as

$$\begin{aligned} \tau : \mathbb{C}^* &\to \mathrm{GL}(2, \mathbb{C}) \\ t &\mapsto \begin{bmatrix} t & 0 \\ 0 & t^{-1} \end{bmatrix}. \end{aligned}$$

This induces a comodule structure on $\mathbb{C}^2$ given by the map $\rho : \mathbb{C}^2 \to \mathbb{C}^2 \otimes \mathcal{O}(\mathbb{C}^*)$

$$\begin{aligned} \rho(a) &= a \otimes x \\ \rho(b) &= b \otimes x^{-1}, \end{aligned}$$

where x is the coordinate function on $\mathbb{C}^*$. ♢

2.4. *Weight space decomposition.* We are now ready to give a proof of Theorem 2.9 using the language of Hopf algebras.

Proof of 2.9. Let V be a finite-dimensional rational representation of T and let $\rho : V \to V \otimes \mathcal{O}(T)$ be the corresponding comodule structure. Recall that

$$\mathcal{O}(T) \cong \mathbb{C}[x_1, \ldots, x_n, {x_1}^{-1}, \ldots, {x_n}^{-1}].$$

We write as a vector space decomposition

$$V \otimes \mathcal{O}(T) \cong \bigoplus_{m \in \mathbb{Z}^n} V \otimes \mathbb{C} \cdot x^m.$$

Expanding $\rho(v)$ in terms of this basis, we find that

$$\begin{aligned}\rho(v) &= \sum_{m\in\mathbb{Z}^n} v_m \otimes x^m \quad \text{finitely many nonzero } v_m \\ \implies (\mathrm{id}\otimes\Delta)(\rho(v)) &= \sum_{m\in\mathbb{Z}^n} v_m \otimes x^m \otimes x^m \\ \implies (\rho\otimes\mathrm{id})(\rho(v)) &= \sum_{m\in\mathbb{Z}^n} \rho(v_m)\otimes x^m \\ \implies \rho(v_m) &= v_m \otimes x^m \quad \text{for those nonzero } v_m.\end{aligned}$$

The second step comes from our computation that $\Delta(x_i) = x_i \otimes x_i$ and the fact that Δ is a coalgebra homomorphism. The claim that Δ is a morphism of coalgebras is not immediate, but it ultimately reduces to the statement that if B is a k-algebra, then the multiplication map $B\otimes B \to B$ is a morphism of k-algebras if and only if B is commutative. We are working with (co)commutative (co)algebras, so this is not an issue. The fourth step comes from equating the second and third left-hand sides.

Finally, we apply the second diagram in 2.10 to get

$$(\mathrm{id}\otimes\epsilon)(\rho(v)) = v = \sum_{m\in\mathbb{Z}^n} v_m\epsilon(x^m) = \sum_{m\in\mathbb{Z}^n} v_m.$$

Thus we see that the comodule V decomposes as a direct sum of subcomodules

$$V = \bigoplus_{m\in\mathbb{Z}^n} V_m,$$

where $V_m := \{v \in V \mid \rho(v) = v\otimes x^m\}$. This is precisely saying that T acts on V_m by the character $\chi_m : T \to \mathbb{C}^*$ given by $t \mapsto t^m$. Moreover, picking a basis for each V_m gives a decomposition of V into a direct sum of one-dimensional irreducible representations

$$V \cong \bigoplus_{m\in\mathbb{Z}^n} W_m^{\oplus n_m},$$

where W_m is the unique one-dimensional irreducible representation for which T acts by χ_m. □

3. **Representations of** GL(2, ℂ)

3.1. *Reducibility.* We saw in Section 2 that every rational representation of T decomposes into a direct sum of irreducible representations and that the irreducible representations are indexed by $\mathbb{Z}^n$.

It turns out that rational representations of $\mathrm{GL}(2,\mathbb{C})$ also decompose into a direct sum of irreducible representations. This is because we can apply Weyl's unitary trick again. We consider $U(2) \subset \mathrm{GL}(2,\mathbb{C})$, the subgroup of unitary matrices. This is a compact subgroup that is also Zariski dense in $\mathrm{GL}(2,\mathbb{C})$. We can then apply the same averaging trick to obtain an inner product on V that will actually be $\mathrm{GL}(2,\mathbb{C})$ invariant, because $U(2)$ is Zariski dense in $\mathrm{GL}(2,\mathbb{C})$.

We refer the reader to Chapter 9 of [FH91] for more detailed discussion.

3.2. *Highest weight vectors.* To completely classify the rational representations of $\mathrm{GL}(2,\mathbb{C})$, we need to introduce highest weight vectors. Let $T \subset \mathrm{GL}(2,\mathbb{C})$ be the subgroup of diagonal matrices and $B \subset \mathrm{GL}(2,\mathbb{C})$ be the subgroup of upper triangular matrices. These ad hoc definitions will work for us, but in general T is a *maximal torus* and B is a *Borel subgroup* of $\mathrm{GL}(2,\mathbb{C})$.

Definition 3.1. Let V be a finite-dimensional rational representation of $\mathrm{GL}(2,\mathbb{C})$. A **highest weight vector** $v \in V$ is a weight vector so that $B \cdot v = \mathbb{C}^* \cdot v$. A **highest weight** is a weight which corresponds to a highest weight vector.

Example 3.2. The group $\mathrm{GL}(2,\mathbb{C})$ has a standard representation on $\mathbb{C}^2$ given by the matrix multiplication map. This action is transitive on the nonzero vectors, so $\mathbb{C}^2$ is irreducible. Considering the torus action $T \subset \mathrm{GL}(2,\mathbb{C})$, we see that $\mathbb{C}^2$ decomposes into a direct sum of weight spaces

$$\mathbb{C}^2 \cong \mathbb{C} \cdot e_1 \oplus \mathbb{C} \cdot e_2,$$

where e_1 and e_2 are the standard basis vectors with weights $(1,0)$ and $(0,1)$, respectively. Then $(1,0)$ is the unique highest weight, and the corresponding weight space is one-dimensional. The standard representation of $\mathrm{GL}(2,\mathbb{C})$ is irreducible. ◊

Example 3.3. Since $\mathrm{GL}(2,\mathbb{C})$ acts on $\mathbb{C}^2$, it also acts on $(\mathbb{C}^2)^{\otimes n}$ for $n \in \mathbb{Z}_{\geq 0}$ via

$$g \cdot (v_1 \otimes \cdots \otimes v_n) = (gv_1 \otimes \cdots \otimes gv_n).$$

This is known as the **tensor product representation** of $\mathrm{GL}(2,\mathbb{C})$. We can further quotient by the submodule generated by vectors of the form

$$v_1 \otimes \cdots \otimes v_i \otimes v_{i+1} \otimes \cdots \otimes v_n - v_1 \otimes \cdots \otimes v_{i+1} \otimes v_i \otimes \cdots \otimes v_n$$

for $1 \leq i \leq n-1$. This is known as the **symmetric power representation** of $\mathrm{GL}(2,\mathbb{C})$, denoted $\mathrm{Sym}^n \mathbb{C}^2$. Choosing a basis e_1, e_2 for $\mathbb{C}^2$ gives a basis for $\mathrm{Sym}^n \mathbb{C}^2$ given by

$$\{e_1^k e_2^{n-k} \mid 0 \leq k \leq n\},$$

and the action of GL(2, $\mathbb{C}$) on $\mathrm{Sym}^n\,\mathbb{C}^2$ is given by

$$g \cdot e_1^k e_2^{n-k} = (ge_1)^k (ge_2)^{n-k}.$$

Setting $g \in T$, we see that $e_1^k e_2^{n-k}$ is a weight vector with weight $(k, n-k)$. One can quickly check that $\mathrm{Sym}^n\,\mathbb{C}^2$ is irreducible for all $n \in \mathbb{Z}_{\geq 0}$. One can also check that the highest weight vector is e_1^n and that it has highest weight $(n, 0)$. ♢

Example 3.4. We have a familiar one-dimensional representation of GL(2, $\mathbb{C}$) given by the determinant map. The determinant of a diagonal matrix is the product of its diagonal entries, and so this representation has weight $(1, 1)$. We will denote the kth power of the determinant map by $\det^k$ for $k \in \mathbb{Z}$. This is a one-dimensional representation with weight (k, k). ♢

We are now ready to state the classification theorem for finite-dimensional rational irreducible representations of GL(2, $\mathbb{C}$).

Theorem 3.5. *Every finite-dimensional rational irreducible representation of* GL(2, $\mathbb{C}$) *is isomorphic to*

$$\mathrm{Sym}^n\,\mathbb{C}^2 \otimes \det{}^k$$

for some $n \in \mathbb{Z}_{\geq 0}$ and $k \in \mathbb{Z}$.

We will prove this theorem by considering the weights that appear in the weight space decomposition of $V|_T$, where $T \subset \mathrm{GL}(2, \mathbb{C})$ is the subgroup of diagonal matrices.

In particular, we appeal to the following facts from representation theory, collectively referred to as the *theorems of the highest weight.*

Theorem 3.6.

(1) *A finite-dimensional rational representation V of* GL(2, $\mathbb{C}$) *is irreducible if and only if it has a unique highest weight vector. In this case, it makes sense to talk about the highest weight of V, defined as the weight corresponding to the highest weight vector.*
(2) *Two finite-dimensional rational irreducible representations of* GL(2, $\mathbb{C}$) *are isomorphic if and only if they have the same highest weight.*
(3) *Let V be a finite-dimensional irreducible rational representation of* GL(2, $\mathbb{C}$) *with highest weight vector v. Then the highest weight of V is contained in the set*

$$\{(a, b) \in \mathbb{Z}^2 \mid a \geq b\}.$$

(4) *Every such weight above is a highest weight for some irreducible representation of* GL(2, $\mathbb{C}$).

We will give a short discussion of the proof of these theorems in the case of $\mathrm{GL}(n, \mathbb{C})$.

Remark 3.7. This theorem holds in great generality. Analagous statements are true for other algebraic groups such as $\mathrm{SL}(n, \mathbb{C})$ and $\mathrm{SO}(n)$, as well as representations of complex semisimple Lie algebras, but in order to make sense of such a theorem, one has to find the right notion of Borel subgroups and highest weight vector.

The proof in full generality is quite technical and we refer the reader to [Mil17] for a more detailed discussion.

The theorems of the highest weight immediately imply the classification theorem for finite-dimensional rational irreducible representations of $\mathrm{GL}(2, \mathbb{C})$. In particular, let V be a finite-dimensional rational irreducible representation of $\mathrm{GL}(2, \mathbb{C})$ with highest weight (a, b). Then by looking at the highest weights (observe that if v is a weight vector for V with weight μ and w is a weight vector for W with weight ν, then $v \otimes w$ is a weight vector for $V \otimes W$ with weight $\mu + \nu$), we see that $V \cong \mathrm{Sym}^{a-b} \mathbb{C}^2 \otimes \det^b$.

4. Theorems of the highest weight

In this section, we will discuss some aspects of Theorem 3.6 in the case of $\mathrm{GL}(2, \mathbb{C})$, in both the smooth setting and the algebraic setting.

4.1. *The smooth setting.* The exposition in this section follows Chapter 8 of [Ful97]. One of the main ingredients in the proof of 3.6 is considering the induced action of $\mathfrak{gl}(2, \mathbb{C})$ on V, where $\mathfrak{gl}(2, \mathbb{C})$ is the Lie algebra of $\mathrm{GL}(2, \mathbb{C})$. Recall that

$$\mathfrak{gl}(2, \mathbb{C}) = \mathrm{Mat}(2, \mathbb{C})$$

is a vector space equipped with a bracket operation given by the commutator. The Lie algebra $\mathfrak{gl}(2, \mathbb{C})$ can be identified with the tangent space of $\mathrm{GL}(2, \mathbb{C})$ at the identity matrix. We can then consider the action of $\mathfrak{gl}(2, \mathbb{C})$ on V given by

$$X \cdot v = \frac{d}{dt}\Big|_{t=0} \exp(tX) \cdot v,$$

where $\exp : \mathfrak{gl}(2, \mathbb{C}) \to \mathrm{GL}(2, \mathbb{C})$ is the exponential map. In particular, this map is the differential of the action of $\mathrm{GL}(2, \mathbb{C})$ on V.

Studying the action of $\mathfrak{gl}(2, \mathbb{C})$ on V is equivalent to studying the action of $\mathrm{GL}(2, \mathbb{C})$ on V because $\mathrm{GL}(2, \mathbb{C})$ is simply connected. This is a general principle which reflects the fact that any map of Lie groups $G \to H$ with G simply connected is determined by its differential at the identity. Then one can show the following lemma.

Lemma 4.1. *A subspace W of a representation of* $\mathrm{GL}(2,\mathbb{C})$ *is a subrepresentation if and only if W is stable under the action of* $\mathfrak{gl}(2,\mathbb{C})$.

We refer the reader to Chapter 3 of [Bou89] for a proof of this lemma. This discussion justifies our passing from the study of $\mathrm{GL}(2,\mathbb{C})$ to the study of $\mathfrak{gl}(2,\mathbb{C})$.

Just as we obtained a decomposition of V as a $\mathrm{GL}(2,\mathbb{C})$ into eigenspaces for the action of T, there is an analagous decomposition for the action of $\mathfrak{gl}(2,\mathbb{C})$. The object which replaces our maximal torus $T \subset \mathrm{GL}(2,\mathbb{C})$ is the **Cartan subalgebra** $\mathfrak{h} \subset \mathfrak{gl}(2,\mathbb{C})$. For us, $\mathfrak{h}$ will be the subspace of diagonal matrices in $\mathfrak{gl}(2,\mathbb{C})$. In general, $\mathfrak{h}$ is a maximal abelian subalgebra of $\mathfrak{gl}(2,\mathbb{C})$.

We can obtain a decomposition of V into eigenspaces for the action of $\mathfrak{h}$

$$V = \bigoplus_{\chi \in \mathfrak{h}^*} V_\chi,$$

where $V_\chi = \{v \in V \mid X \cdot v = \chi(X)v \text{ for all } X \in \mathfrak{h}\}$. Moreover $\mathfrak{gl}(2,\mathbb{C})$ acts on itself via the bracket (adjoint representation) and we can decompose this action as

$$\begin{aligned} \mathfrak{gl}(2,\mathbb{C}) &\cong \mathfrak{h} \oplus \mathfrak{g}_\alpha \oplus \mathfrak{g}_{-\alpha} \\ &\cong \mathfrak{h} \oplus \mathbb{C}e \oplus \mathbb{C}f, \end{aligned}$$

where $\alpha\left(\begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}\right) = d_1 - d_2$ and

$$e = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$
$$f = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

One can check that for all $h \in \mathfrak{h}$ we have

$$\begin{aligned} [h, e] &= \alpha(h)e \\ [h, f] &= -\alpha(h)f. \end{aligned}$$

The weights which appear in the adjoint representation of $\mathfrak{gl}(2,\mathbb{C})$ are called the **roots**. We will say α is a **positive root** and $-\alpha$ is a **negative root**, and call $\mathbb{C}e$ and $\mathbb{C}f$ the corresponding **positive** and **negative root spaces**. Then a weight vector $v \in V$ is a highest weight vector if and only if $e \cdot v = 0$.

We care about roots of the adjoint representation for the following reason. Let V be a finite-dimensional rational representation of $\mathfrak{gl}(2,\mathbb{C})$. Decompose V into eigenspaces for the action of $\mathfrak{h}$ as before:

$$V = \bigoplus_{\chi \in \mathfrak{h}^*} V_\chi.$$

Knowing how $\mathfrak{h}$ acts on V, we now need to investigate the actions of e and f. As it turns out, e and f are operators which translate between the weight spaces. Specifically, let v be an weight vector for the action of $\mathfrak{h}$ with weight χ. Then $e \cdot v$ is a weight vector with weight $\chi + \alpha$. Indeed—recalling that the action of the Lie algebra respects brackets—for $X \in \mathfrak{h}$ we have

$$\begin{aligned} X \cdot ev &= e \cdot Xv + [X, e] \cdot v \\ &= \chi(X)ev + \alpha(X)ev. \end{aligned}$$

A priori, we know nothing about the weights of V. Now we know that all of the weights of V are translates of each other by the roots of $\mathfrak{gl}(2,\mathbb{C})$. Now let μ be any weight which appears in the decomposition of V. Then we can consider the translates

$$\mu + \mathbb{Z}\alpha$$

and since V is finite-dimensional, only finitely many of the weight spaces of V corresponding to these weights are nonzero. Recall we picked a positive system, so now it makes to talk about the highest weight (it is the weight χ so that all of the weights $\chi + \mathbb{N}\alpha$ correspond to empty weight spaces).

If V is an irreducible representation then a highest weight vector must span its root space. This is because if v is a highest weight vector, then one can show that the subspace generated by v, $f \cdot v$, $f^2 \cdot v$, $\ldots$ is a subrepresentation. It follows that an irreducible representation can have only one highest weight vector (up to scale).

4.2. *The algebraic setting.* In order to justify the passage from $\mathrm{GL}(2,\mathbb{C})$ to $\mathfrak{gl}(2,\mathbb{C})$, we made use of the exponential map. This is not available in the category of varieties. However we can still make sense of the Lie algebra of an algebraic group and the induced action of the Lie algebra on a vector space. To do so, we need to pass to the *Zariski tangent space* of a variety.

Definition 4.2. Let A be a local ring and $\mathfrak{m}$ its maximal ideal. The residue field k of A is the field $A/\mathfrak{m}$ and the **Zariski cotangent space** of A is the k-vector space $\mathfrak{m}/\mathfrak{m}^2$. The **Zariski tangent space** of A is the dual vector space $\mathrm{Hom}_k(\mathfrak{m}/\mathfrak{m}^2, k)$. If X is a variety and $p \in X$, then we define the **Zariski tangent space** of X at p to be the Zariski tangent space of $\mathcal{O}_{X,p}$.

To make sense of this definition, we need to borrow a little motivation from the theory of differentiable manifolds. If M is a smooth manifold, tangent vectors at a point $p \in M$ are in one-to-one correspondence with derivations of the ring of germs of smooth functions at p, i.e., $\mathbb{R}$-linear maps $\mathcal{O}_{M,p} \to \mathbb{R}$ which satisfy the Leibniz rule

$$D(fg) = f(p)Dg + g(p)Df$$

for all $f, g \in \mathcal{O}_{M,p}$. We refer to Chapter 3 of [Lee03] for a more detailed discussion of this point of view.

Proposition 4.3. *Let X be a variety over a field k and let $p \in X$. Consider the local ring $\mathcal{O}_{X,x}$ and its maximal ideal $\mathfrak{m}$. Let $k(p)$ be the residue field of $\mathcal{O}_{X,p}$. It coincides with k. There is an isomorphism*

$$\operatorname{Hom}_k(\mathfrak{m}/\mathfrak{m}^2, k) \cong \operatorname{Der}_k(\mathcal{O}_{X,p}, k(x)).$$

Proof. A derivation is precisely the data of a k-linear map $\mathfrak{m} \to k$ which satisfies the Leibniz rule. This extends to a k-linear map $\mathcal{O}_{X,p} \to k$ by precomposing with $f \mapsto f - f(p)$. Moreover, $\mathfrak{m}^2$ maps to zero because if $f(p) = g(p) = 0$, then

$$D(fg) = f(p)Dg + g(p)Df = 0.$$

Therefore, a derivation induces an element of the tangent space of X at p.

Conversely, if we have a k-linear map $\mathfrak{m}/\mathfrak{m}^2 \to k$, precompose with the quotient map to get $D : \mathfrak{m} \to k$. Then we have to show that D satisfies the Leibniz rule. This is a straightforward computation. Let $f, g \in \mathcal{O}_{X,p}$. Then $(f - f(p))(g - g(p)) \in \mathfrak{m}^2$ and so

$$\begin{aligned} 0 = D((f - f(p))(g - g(p))) &= D(fg - f(p)g - fg(p) + f(p)g(p)) \\ \implies D(fg) &= f(p)Dg + g(p)Df, \end{aligned}$$

since constants derive to zero and so D is a derivation. It is clear that these two maps are inverses of each other. □

Now we can make sense of the Lie algebra of an algebraic group.

Definition 4.4. Let G be an algebraic group. The **Lie algebra** of G, denoted $\mathfrak{g}$, is the Zariski tangent space of G at the identity.

Note that if $\alpha : G \to W$ is a morphism of varieties, then there is an induced map $\mathcal{O}(W) \to \mathcal{O}(G)$ on coordinate rings, and this map is local in the sense that $\mathcal{O}_{W,\alpha(p)} \to \mathcal{O}_{G,p}$ is a local ring homomorphism for all $p \in G$. Geometrically, this is saying that if a regular function on W vanishes at a point $\alpha(p)$, then its pullback to G vanishes at p.

In particular, we see that a morphism of varieties α has a differential $d\alpha$ which takes a derivation $D : \mathcal{O}_{W,\alpha(p)} \to k$ to a derivation $d\alpha(D) : \mathcal{O}_{G,p} \to k$. Letting $W = \mathrm{GL}(V)$, we see that the differential of the action of G on V gives us a Lie algebra representation (in the sense that it respects the bracket) of $\mathfrak{g}$ on V.

Then again one proves that Lemma 4.1 holds in the algebraic setting and so we have reduced to the study of the action of $\mathfrak{g}$ on V. A reference for this proof can be found in Chapter 1 of [Bor91]. We can then proceed as in the smooth setting to prove Theorem 3.6.

References

[Bor91] Armand Borel. *Linear algebraic groups.* Springer-Verlag, New York, 1991.

[Bou89] Nicolas Bourbaki. *Lie groups and Lie algebras*: *Chapters* 1-3. Springer-Verlag, Berlin, 1989.

[FH91] William Fulton and Joe Harris. *Representation theory*: *a first course.* Springer-Verlag, New York, 1991.

[Ful97] William Fulton. *Young tableaux*: *with applications to representation theory and geometry.* Cambridge University Press, Cambridge, 1997.

[Lee03] John M. Lee. *Introduction to smooth manifolds.* Springer-Verlag, New York, 2003.

[Mil17] James S. Milne. *Algebraic groups*: *the theory of group schemes of finite type over a field.* Cambridge University Press, Cambridge, 2017.

[Wat79] William C. Waterhouse. *Introduction to affine group schemes.* Springer-Verlag, Berlin-New York, 1979.

Department of Mathematics, Cornell University
E-mail: sy459@cornell.edu
https://journals.library.columbia.edu/index.php/cjum

Columbia Journal of Mathematics **1** (2024),
43–69

The Gauss–Bonnet theorem

By Bonnie Yang

1. **Introduction**

The Gauss–Bonnet theorem is a crowning result of surface theory that gives a fundamental connection between geometry and topology. Roughly speaking, geometry refers to the "local" properties—lengths, angles, curvature—of some fixed object, while topology seeks to identify the "global" properties that are unchanged by a continuous deformation, such as stretching or twisting. The theorem formalizes an intuitive idea: continuous changes to curvature on one region of a surface will be balanced out elsewhere, so the *total* curvature of the surface stays the same.

Explicitly, the Gauss–Bonnet theorem says that a surface's total curvature, defined using its local Gaussian curvature, is directly proportional to the number of holes in the surface, which comes from an invariant quantity called its Euler characteristic. The Euler characteristic is a way of classifying which surfaces can be continuously deformed into one another; as an informal example, the classic joke that "a topologist is a person who cannot tell the difference between a coffee mug and a doughnut" comes from the fact that the objects each have one hole. Even though a coffee mug and a doughnut have visibly different geometric shapes, according to the Gauss–Bonnet theorem, both objects will have the same total curvature.

Our goal is to show

$$\int_{\mathcal{S}} K dA = 2\pi\chi(\mathcal{S}),$$

where $\mathcal{S}$ is a closed surface in $\mathbb{R}^3$, K is the Gaussian curvature, dA is the area element, and $\chi(\mathcal{S})$ is the Euler characteristic. The proof itself is delightfully systematic: we first find the total curvature of a curve on a plane, extend that result to curves on three-dimensional surfaces, extend *that* result to "polygons" on surfaces, and finally the entire surface.

In Section 2, we prove Hopf's Umlaufsatz for the total curvature of a simple closed curve in $\mathbb{R}^2$. Sections 3, 4, and 5 introduce concepts from differential geometry to define Gaussian curvature. In Section 6, we prove the local

Gauss–Bonnet theorem for the total curvature of a surface polygon. At last, in Section 7, we prove the global Gauss–Bonnet theorem for compact surfaces by covering the surface with polygons and applying the local Gauss–Bonnet theorem to each one.

Our discussion focuses on exposition, and references will be given in place of tedious computations when reasonable. This paper assumes a somewhat rigorous understanding of multivariable calculus and linear algebra, as well as some elementary group theory.

2. **Plane curves and Hopf's Umlaufsatz**

Hopf's Umlaufsatz[1] asserts that the total signed curvature of any simple closed curve in $\mathbb{R}^2$ is equal to $\pm 2\pi$, with sign depending on the curve's orientation. Although the theorem is about the curvature of a line and not a region with area, the Umlaufsatz does much of the heavy lifting for our later proof in $\mathbb{R}^3$. We begin with some preliminary theory of paths and curves.

Definition 2.1. A **(parametric) path** in $\mathbb{R}^n$ is a continuous function $\gamma : I \to \mathbb{R}^n$, where I is any interval of $\mathbb{R}$. The image of a path is called a **parametrized curve** in $\mathbb{R}^n$.

If γ is differentiable, the differential[2] $\dot{\gamma}(t)$ is called the **tangent vector** of γ at the point $\gamma(t)$. We say γ is **regular** if $\dot{\gamma}(t)$ is nonzero for all $t \in I$.

Remark 2.2. A particular curve can be the image of infinitely many paths. To see this, suppose γ_1 and γ_2 are two paths defined on the intervals I_1 and I_2, respectively. Since these are intervals of $\mathbb{R}$, we can define a bijection $\phi : I_1 \to I_2$ between their domains. Then if γ_1 and γ_2 are both injective with the same image curve, we can always *reparametrize* one path as the other by a composition $\gamma_2 = \gamma_1 \circ \phi$.

In practice, the terms *path* and *curve* are used interchangeably to mean either a continuous function $\gamma : [a, b] \to \mathbb{R}^n$ or its image. The correct interpretation should be clear from context.

Unless otherwise specified, all curves discussed in this paper are assumed to be regular and *smooth*, meaning there exist continuous partial derivatives of all orders.

---

[1] From German *umlauf* (rotation) and *satz* (theorem)—sometimes translated, unsurprisingly, to "rotation angle theorem."

[2] The "overdot" notation is conventially used for a derivative taken with respect to time (i.e., $\dot{\gamma} = d\gamma/dt$ and $\ddot{\gamma} = d^2\gamma/dt$).

Definition 2.3. If $\gamma : [a,b] \to \mathbb{R}^n$ is a parametrized curve, then for any $a \le t \le b$, the **arc length** of γ from a to t is given by the function

$$s(t) = \int_a^t \|\dot{\gamma}_t\| dt.$$

A regular curve γ is **unit-speed** if for all t, we have $\|\dot{\gamma}(t)\| = 1$. In this case, the arc length is $s(t) = t$, so γ is also said to be an **arc length parametrization**.

Remark 2.4. Every regular curve can be reparametrized to unit speed.

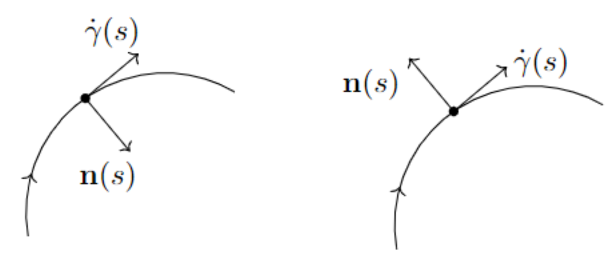
Hopf's Umlaufsatz involves an integral over the *curvature* of a plane curve, so we now focus our discussion on some geometric properties that are specific to curves in $\mathbb{R}^2$. For plane curves, which have two choices of unit normal vector for each tangent vector $\dot{\gamma}(s)$, we fix the **signed unit normal n** to be the vector obtained by rotating $\dot{\gamma}$ counterclockwise by $\pi/2$.

PROPOSITION 2.5. *Given a unit-speed plane curve γ, there exists a scalar κ called the **signed curvature** of γ such that*

$$\ddot{\gamma} = \kappa \mathbf{n},$$

where $\mathbf{n}$ is the signed unit normal of γ. Note that κ can be positive, negative, or zero for each point of the curve γ.

Proof. Recall that $\langle \dot{\gamma}, \dot{\gamma} \rangle = 1$, so we can differentiate to obtain $\langle \ddot{\gamma}, \dot{\gamma} \rangle + \langle \dot{\gamma}, \ddot{\gamma} \rangle = 0$. Thus, the vectors $\dot{\gamma}$ and $\ddot{\gamma}$ are perpendicular, so $\ddot{\gamma}$ must be a scalar multiple of $\mathbf{n}$. □



(a) $\kappa > 0$ (b) $\kappa < 0$

This formulation of curvature is strictly local, since it arises from the behavior of a curve at a specific point: if $\gamma(s)$ is a point on a unit-speed curve, then $\|\mathbf{n}(s)\| = 1$ and we have precisely $|\kappa(s)| = \|\ddot{\gamma}(s)\|$. To see how Hopf's Umlaufsatz relates local curvature to a curve's topology, we must next get a sense of what global properties a curve has.

We start with a geometric interpretation of the tangent vector for plane curves. When $\gamma : [a,b] \to \mathbb{R}^2$ is unit-speed, the direction of each vector $\dot{\gamma}(s)$

is determined by the angle $\theta(s)$ for which $\dot\gamma(s) = e^{i\theta(s)}$. It is straightforward to show that our choice of $\theta(s)$ is smooth: briefly, if $\dot\gamma$ is indeed defined on the complex unit circle, then the chain rule implies

$$\ddot\gamma(s) = i\dot\theta(s) \cdot e^{i\theta(s)} = \dot\theta(s);$$

one can recover $\dot\theta$ as the scalar in this expression, and then the continuous map θ by taking an antiderivative.

Definition 2.6. Let $f : [a, b] \to S^1$ be any path in the unit circle, and let $p : \mathbb{R} \to S^1$ be defined by $p(t) = e^{it}$. An **angle function** for f is a smooth map $\theta : [a, b] \to \mathbb{R}$ which satisfies

$$f(s) = p \circ \theta = e^{i\theta(s)}.$$

If $f = \dot\gamma$ for some unit-speed plane curve γ, then θ is called a **tangent angle function** for γ.

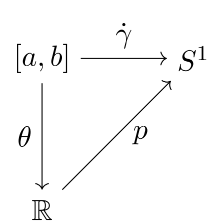
PROPOSITION 2.7. *Given a unit-speed curve $\gamma : [a, b] \to \mathbb{R}^2$ with a tangent angle function θ, the* ***signed curvature*** *of γ is defined by*

$$\kappa = \dot\theta,$$

the rate at which the tangent vector $\dot\gamma$ rotates. (*See* [Pre10, Proposition 2.2.1] *for a proof.*)

The upshot of this discussion is that we can express the tangent $\dot\gamma$ of any plane curve γ as a path in the unit circle! This is useful because every path in S^1 has a fixed *degree*, which counts how many times the curve "goes around" the circle counterclockwise. Defining a path $\dot\gamma : [a, b] \to S^1$ this way allows us to treat the degree of the tangent as a topological property of γ itself. Later, we will see that the proof of the Umlaufsatz is essentially an argument about the degree of $\dot\gamma$ in a specific case: when γ is a simple closed curve.

A tangent angle function θ takes each point $\dot\gamma(s)$ on the circle to a number on the "unfolded" real line, which we call a *lift* of $\dot\gamma$ to $\mathbb{R}$. Notice that a tangent curve $\dot\gamma$ which winds around the circle n times will have its tangent angle function increase by n. Using the fact that each corresponding angle $\theta(s)$ is unique up to an integer multiple of 2π, we can recover the degree of $\dot\gamma$ from this unfolding process.



Relationship between domains for the tangent curve $\dot{\gamma}$, tangent angle function θ, and the unit circle S^1.

PROPOSITION 2.8. *Let* $f : [a, b] \to S^1$ *be a path in the circle and* $\theta, \phi : [a, b] \to \mathbb{R}$ *be any two angle functions for* f*. Then we have*

$$\theta(b) - \theta(a) = \phi(b) - \phi(a).$$

Equivalently, for a chosen tangent angle $\theta(s_0)$ *with* $s_0 \in [a, b]$, *there exists a* unique *angle function* θ_0 *such that* $f(s_0) = e^{i\theta_0(s_0)}$.

Proof. We will show that for $e^{i\theta(s)}$ and $e^{i\phi(s)}$ to agree, the values $\theta(s)$ and $\phi(s)$ must differ by an integer multiple of 2π, and by continuity, the integer must be the same for all s.

First, since both expressions for $\dot{\gamma}(s)$ are points in S^1, the angles $\theta(s)$ and $\gamma(s)$ clearly differ by full rotations about unit circle. Formally, this means there exists some integer $n(s)$ such that for all $s \in [a, b]$, we have

$$\phi(s) - \theta(s) = 2\pi n(s).$$

Because θ and ϕ are continuous functions, n is continuous on the domain $[a, b]$ as well, and we apply the intermediate value theorem to conclude that n is a constant that does not depend on s. Thus, the integer term cancels, and we see

$$\phi(b) - \phi(a) = \theta(b) + 2\pi n(s) - \theta(a) - 2\pi n(s) = \theta(b) - \theta(a)$$

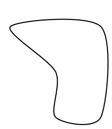
as desired. □

Definition 2.9. Let $f : [a, b] \to S^1$ be a path in the circle and let $\theta : [a, b] \to \mathbb{R}$ be a tangent angle function of γ. The **degree** of f is defined as

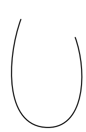
$$\frac{\theta(b) - \theta(a)}{2\pi}.$$

If $\gamma : [a, b] \to \mathbb{R}^2$ is a unit-speed plane curve, then the degree of its tangent $\dot{\gamma}$ is called the **rotation index** of γ and denoted $\mathbf{ind}(\gamma)$.

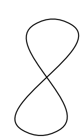
Definition 2.10. Given a compact interval $[a, b] \subset \mathbb{R}$, we say $\gamma : [a, b] \to \mathbb{R}^n$ is a **closed curve** of period $b - a$ if $\gamma(a) = \gamma(b)$. If γ is injective on the open interval (a, b), then γ is called **simple**.



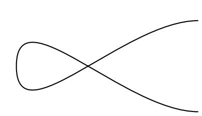
Simple and closed



Simple, not closed



Not simple, closed



Not simple and not closed

The Jordan curve theorem from topology tells us that any simple closed curve on a plane has an "interior" and an "exterior." Precisely, if γ is a simple closed curve in $\mathbb{R}^2$, then the *complement* of its image is the union of two subsets of $\mathbb{R}^2$, denoted $\text{int}(\gamma)$ and $\text{ext}(\gamma)$, which satisfy the following:

- $\text{int}(\gamma)$ and $\text{ext}(\gamma)$ are disjoint, so $\text{int}(\gamma) \cap \text{ext}(\gamma) = \emptyset$;
- $\text{int}(\gamma)$ is bounded and $\text{ext}(\gamma)$ is unbounded;
- Both $\text{int}(\gamma)$ and $\text{ext}(\gamma)$ are *connected*, so any two points in the same subset can be joined by a curve contained entirely in that subset.

This gives us a way to distinguish between two possible orientations of γ using geometry: we say γ is **positively-oriented** if the signed unit normal $\mathbf{n}$ points into $\text{int}(\gamma)$ at every point in the curve.

Now, when we claim that a property like the rotation index is global, we mean that it is invariant under a "continuous deformation." The following definition formalizes this notion for closed curves in $\mathbb{R}^2$.

Definition 2.11. An **isotopy** of simple closed plane curves of period ℓ is a family of curves $\gamma_t : \mathbb{R} \to \mathbb{R}^2$ such that

(i) Each curve γ_t is period ℓ;
(ii) For all $0 \leq t \leq 1$, the map $h : \mathbb{R} \times [0,1] \to \mathbb{R}^2$ defined by $h(s,t) = \gamma_t(s)$ is also a regular, smooth, and closed plane curve of period ℓ;
(iii) We have $h(s,0) = \gamma_0(s)$ and $h(s,1) = \gamma_1(s)$.

If such a family exists, we say that γ_0 is **isotopic** to γ_1.

Example 2.12. We have already seen an example of such a family: the reparametrizations discussed at the beginning of this section are given by isotopies of the form $h(s,t) = \gamma(s + s_0 t)$, where $s_0 \in \mathbb{R}$ is a constant. ♢

Example 2.13. A **translation** of a plane curve is an isotopy of the form

$$h(s,t) = \gamma(s) + t\vec{x}$$

for some point $\vec{x} \in \mathbb{R}^2$. ♢

Lemma 2.14. *If γ_0 and γ_1 are closed plane curves connected by an isotopy, then $I(\gamma_0) = I(\gamma_1)$.*

Proof. Similar to the proof of Proposition 2.8, we show that the rotation index is an integer constant by continuity. First, notice that the rotation index for a closed curve is indeed an integer. Now let h be an isotopy from γ_0 to γ_1, and fix $\gamma_t(s) = h(s,t)$. Then the map from s to $I(\gamma_s)$ given by the equation in Definition 2.9 is a continuous function $[0,1] \to \mathbb{Z}$, so we apply the intermediate value theorem to conclude that $I(\gamma_s)$ is constant. □

THEOREM 2.15 (Hopf's Umlaufsatz). *Let $\gamma : [a,b] \to \mathbb{R}^2$ be a unit-speed, simple closed curve on a plane. Then the total signed curvature is given by*

$$\int_\gamma \kappa ds = \pm 2\pi.$$

As promised, this reduces to a claim about the rotation index! Since $\kappa = \dot{\theta}$ for any curve by Proposition 2.7, the total signed curvature can be computed as

$$\int_\gamma \kappa ds = \int_a^b \dot{\theta}(s) ds = \theta(b) - \theta(a) = 2\pi \cdot \mathbf{ind}(\gamma).$$

Thus, the point of the Umlaufsatz is that for *simple* closed curves, we have $\mathbf{ind}(\gamma) = \pm 1$.

Proof of Theorem 2.15. Our strategy is to replace $\dot{\gamma} : [a,b] \to S^1$ with another map to the circle, the *secant line* between two points on a curve. Crucially, the degree of the secant line is straightforward to compute, so we will use it to obtain $\mathbf{ind}(\gamma)$ indirectly.

Both the secant line and its angle function take two parameter inputs. When the two parameters are equal, the secant is precisely the tangent line, and the secant angle function is continuously extended to the tangent angle function of γ at a single point. The domain of this secant map can be interpreted geometrically as a triangle formed by the points (a,a), (a,b), and (b,b), and the restriction of the secant map to the diagonal is exactly the tangent map $\dot{\gamma}$. A continuous deformation of the diagonal to the other two sides of the triangle preserves the endpoints (a,a) and (b,b), so the total change of the secant angle function is the same along this deformed path. Then to find $I(\gamma)$, it suffices to compute the degree of the secant map coming from the non-diagonal sides.

We begin by assuming, without loss of generality, that $\gamma(a)$ is the lowest point on the curve and is located at the origin $(0,0)$. Since γ is assumed to be continuous, the projection of γ to its y-coordinate is continuous on $[a,b]$ as well, so we know there exists a $t_0 \in [a,b]$ such that the y-coordinate of $\gamma(t_0)$ is minimal. The remaining assumptions follow because the rotation index is invariant under isotopy, including the reparametrizations and translations given as examples of Definition 2.11. Finally, because γ is unit-speed, we also have $\dot{\gamma}(a) = \pm e_1$, the first standard basis vector of $\mathbb{R}^2$.

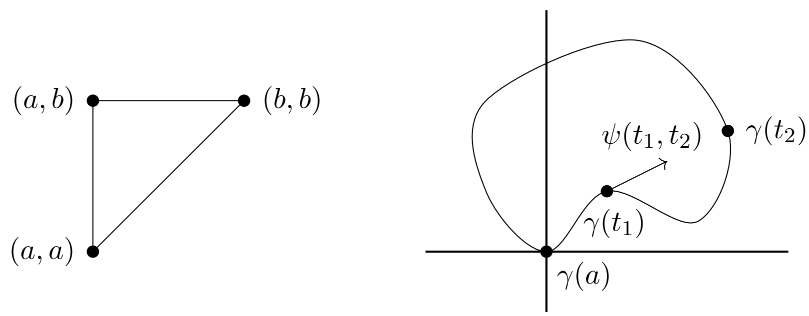
Now we are ready to define the secant map. Let

$$\triangle = \{(t_1, t_2) \mid a \leq t_1 \leq t_2 \leq b\},$$

and define the continuous function $\psi : \triangle \to S^1$ by

$$\psi(t_1, t_2) = \begin{cases} \dot{\gamma}(t_1) & t_1 = t_2 \\ -\dot{\gamma}(a) & (t_1, t_2) = (a, b) \\ \dfrac{\gamma(t_2) - \gamma(t_1)}{\|\gamma(t_2) - \gamma(t_1)\|} & \text{otherwise.} \end{cases}$$

This is a smooth function (see [Ben17, pages 22-24]), and the first two cases are straightforward to visualize. For parameters (t_1, t_2) which satisfy the third case, the vector $\psi(t_1, t_2)$ is precisely the unit vector with origin $\gamma(t_1)$ and pointing towards $\gamma(t_2)$. In particular, if (t_1, t_2) lies on a non-diagonal side of the triangle $\triangle$, then $\gamma(t_1)$ is fixed as $\gamma(t_2)$ travels along the curve (see [Kni06] for nice animations).



By applying Proposition 2.8 in each coordinate, we see that there exists a smooth function $\tilde{\theta} : \mathbb{R}^2 \to S^1$ which gives the angle $\tilde{\theta}(t_1, t_2)$ between $\psi(t_1, t_2)$ and the horizontal. Because we defined $\psi = \dot{\gamma}$ along the diagonal, by Proposition 2.8, we know that

$$2\pi \cdot \mathbf{ind}(\gamma) = \theta(b) - \theta(a) = \tilde{\theta}(b, b) - \tilde{\theta}(a, a),$$

so $\mathbf{ind}(\gamma)$ is equal to the degree of ψ! Further, it is visually clear that we can compute the total change of $\tilde{\theta}$ the diagonal by computing the change from (a, a) to (a, b) and (a, b) to (b, b) separately, then taking a sum. Thus, we have

$$2\pi \cdot \mathbf{ind}(\gamma) = \tilde{\theta}(b, b) - \tilde{\theta}(a, a) = \left(\tilde{\theta}(a, b) - \tilde{\theta}(a, a)\right) + \left(\tilde{\theta}(b, b) - \tilde{\theta}(a, b)\right).$$

The last step is to compute the degree of ψ over the two non-diagonal segments. We will suppose γ is positively-oriented, so $\dot{\gamma}(a) = e_1$ and the secant angle is $\tilde{\theta}(a, a) = 0$ (an analogous argument holds for the opposite orientation, where $\tilde{\theta}(a, a) = \pi$). For the segment from (a, a) to (a, b), we know that the corresponding line $\psi(a, t)$ lies in the upper half-plane for all $t \in [a, b]$, so we must have $0 \leq \tilde{\theta}(a, t) \leq \pi$. Thus, we find $\tilde{\theta}(a, b) = \pi$. Meanwhile, on the segment from (a, b) to (b, b), we have the corresponding line $\psi(t, b) = -\psi(a, t)$, which implies $\tilde{\theta}(b,b) - \tilde{\theta}(a,b) = \pi$ as well. The degree of ψ is therefore

$$\frac{(\tilde{\theta}(a,b) - \tilde{\theta}(a,a)) + (\tilde{\theta}(b,b) - \tilde{\theta}(a,b))}{2\pi} = \frac{\pi + \pi}{2\pi} = 1$$

and -1 if the orientation of ψ is reversed. This shows $\mathbf{ind}(\gamma) = \pm 1$ as desired.

Altogether, we conclude that if θ is any tangent angle function for γ, then

$$\int_a^b \kappa ds = \theta(b) - \theta(a) = 2\pi \cdot \mathbf{ind}(\gamma) = \pm 2\pi,$$

which completes the proof. □

3. Regular surfaces and tangent planes

In the previous section, we showed that the two-dimensional circle can be locally unfolded to the one-dimensional real line using the function e^{it}, which gives a continuous deformation on sufficiently small intervals. Similarly, we interpret surfaces as three-dimensional objects which can be "flattened" to $\mathbb{R}^2$.

Definition 3.1. Given any subsets $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$, a invertible map $f : X \to Y$ is called a **homeomorphism** if both f and its inverse $f^{-1} : Y \to X$ are continuous. If such a map exists, we say X and Y are **homeomorphic**.

Remark 3.2. The paths defined in Section 2 are homeomorphisms from an interval of $\mathbb{R}$ to a curve in $\mathbb{R}^n$. In general, isotopies, which we only defined for simple closed plane curves, are continuous families of homeomorphisms.

Definition 3.3. A **regular surface** is a subset $\mathcal{S} \subset \mathbb{R}^3$ where for each point $p \in \mathcal{S}$, there exists an open neighborhood $V \subset \mathbb{R}^3$ containing p, an open subset $U \subset \mathbb{R}^2$, and a map $\sigma : U \to V \cap \mathcal{S}$ with the following properties:

(i) σ is a smooth function on U;
(ii) σ is a homeomorphism;
(iii) For all $q \in U$, the differential $d\sigma_q$ is injective.

In this case, the map σ is called a **surface patch** or **local parametrization** of the coordinate neighborhood $V \cap \mathcal{S}$. We will also only consider **connected** surfaces, meaning any two points in $\mathcal{S}$ can be joined by a curve lying entirely in $\mathcal{S}$.

Unless otherwise specified, all surfaces discussed in this paper are assumed to be regular.

Remark 3.4. Like paths, multiple surface patches may have the same image. Suppose the surface patches σ_1 and σ_2 are defined on the open subsets $U_1, U_2 \subset \mathbb{R}^2$ respectively. We say that two surface patches σ_1, σ_2 are **reparametrizations** of one another if there exists a homeomorphism $\Phi : U_1 \to U_2$ such that $\sigma_2 = \sigma_1 \circ \Phi$. In this case, the bijection Φ is called a

reparametrization map. The upshot is that we can define any geometric property of a smooth surface by defining it up to reparametrization!

Condition (i) is basic for doing calculus on surfaces, like understanding what it means for a function on a surface to be differentiable. Condition (ii) ensures that the inverse $\sigma^{-1} : V \cap \sigma(U) \to U$ is continuous, so the surface has no self-intersections and the tangent to each point is unique. Condition (iii), sometimes called the *regularity condition*, allows us to apply the immersion theorem to conclude that σ is indeed "locally invertible" when the codomain is restricted to $V \cap \sigma(U)$.

Example 3.5. A surface is often the image of multiple surface patches. Given the unit sphere S^2, which has radius 1, we can define the smooth maps $\sigma_1, \sigma_2 : U \to S^2$ by

$$\sigma_1 \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos(u)\cos(v) \\ \cos(u)\sin(v) \\ \sin(u) \end{pmatrix} \qquad \sigma_2 \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -\cos(u)\cos(v) \\ \sin(u) \\ -\cos(u)\sin(u) \end{pmatrix},$$

where u and v are angles corresponding to something like latitude and longitude, respectively. That is, if p is a point on the sphere, then we can draw a line through p which is parallel to the z-axis and intersects the xy-plane at a point q. Then u is the angle between p and q, while v is the angle between q and the positive x-axis.

To ensure σ_1 and σ_2 are homeomorphisms, we take the domain to be the open set $U = (-\pi/2, \pi/2) \times (0, 2\pi) \subset \mathbb{R}^2$. Notice that neither σ_1 nor σ_2 cover all of S^2 when restricting the domain to U: the image of σ_1 misses points of the form $(x, 0, z)$ with $x \geq 0$, while the image of σ_2 misses points of the form $(x, y, 0)$ with $x \leq 0$. However, we have $S^2 = \sigma_1(U) \cup \sigma_2(U)$, so S^2 satisfies the definition of a surface.

Thus, the construction of a surface can be somewhat ad hoc. Our strategy also happens to be unnecessarily complicated for the sphere, which has a neat geometric origin we will introduce in the next example. $\diamondsuit$

Example 3.6. A *surface of revolution* is obtained by rotating a simple plane curve, called the *profile curve*, around a straight line in the plane. Typically, the axis of revolution is the z-axis, and we define a path $\gamma : I \to \mathbb{R}^3$ on the xz-plane by $\gamma(u) = (f(u), 0, g(u))$. The surface obtained by rotating γ about the z-axis is parametrized with $\sigma : I \times [0, 2\pi) \to \mathbb{R}^3$ given by

$$\sigma(u, v) = (f(u)\cos v,\ f(u)\sin v,\ g(u)),$$

where v is the angle of rotation. To check for Definition 3.3 (iii), notice

$$\sigma_u \times \sigma_v = f(u)(-\dot{g}(u)\cos v,\ -\dot{g}(u)\sin v,\ \dot{f}(u)),$$

so $\sigma_u \times \sigma_v$ is nonzero if and only if $f(u) \neq 0$ and $\dot{f}, \dot{g}$ are not both zero; the nonzero vector product implies that σ_u and σ_v are linearly independent, which we will show is crucial for doing calculus on surfaces in the following discussion of tangent planes. Thus, the surface of revolution is indeed a surface when γ does not intersect the z-axis and is indeed regular. In practice, we assume $f(u) > 0$ so that $f(u)$ is the distance between $\sigma(u, v)$ and the axis of rotation. $\Diamond$

Example 3.7. The unit sphere S^2 in latitude-longitude coordinates, as in the first example, is a surface of revolution with profile curve functions $f(u) = \cos(u)$ and $g(u) = \sin(u)$. $\Diamond$

Example 3.8. A *torus* is formed by rotating a circle in the xz-plane with center $(R, 0, 0)$ and radius r about the z-axis, with $R > r > 0$. This is a surface of revolution with profile curve

$$\gamma(\theta) = (R + r\cos\theta,\ 0,\ r\sin\theta),$$

and the parametrization is $\sigma : [0, 2\pi) \times [0, 2\pi) \to \mathbb{R}^3$ defined by

$$\sigma\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} (R + r\cos(u))\cos(v) \\ (R + r\cos(u))\sin(v) \\ r\sin(v) \end{pmatrix},$$

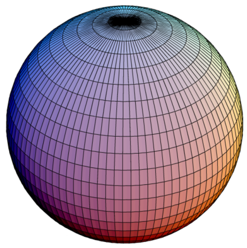
where u is the angle in γ and v is the angle about the z-axis. $\Diamond$

Example 3.9 (Non-example). Consider a line passing through the origin that forms an angle α with the xy-plane, such that the length of the line above the plane is the same as the length below. Rotating this line about the z-axis generates a *circular cone* with vertex at the origin. For example, if $\alpha = \pi/4$, the cone is parametrized by

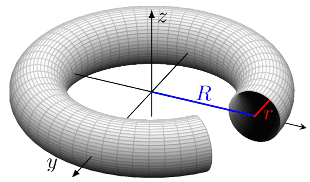
$$\mathcal{S} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = z^2\}.$$

We give an abridged argument for why this is not a regular surface (for full explanation and diagrams, see [Pre10, Example 4.1.5]). Let $U \subset \mathbb{R}^2$ be an open ball and $\sigma : U \to V \cap \mathcal{S}$ be a surface patch that contains the vertex $(0, 0, 0)$. Further, let $\vec{a} \in U$ be the point at the center of U such that $\sigma(a) = (0, 0, 0)$. The open set $V \cap \mathcal{S}$ must contain a point $\vec{p}$ in the upper half of the cone where $z > 0$, as well as a point $\vec{q}$ in the lower half where $z < 0$; let $\vec{a}, \vec{b} \in U$ be the points with $\sigma(\vec{a}) = \vec{p}$ and $\sigma(\vec{b}) = \vec{q}$. We can find a curve $\beta : I \to U$ that passes through $\vec{b}$ and $\vec{c}$, but not $\vec{a}$; this implies the existence of a continuous curve $\gamma = \sigma \circ \beta$ that passes through $\vec{p}$ and $\vec{q}$ but not $(0, 0, 0)$, which contradicts the definition of a surface patch σ. $\Diamond$

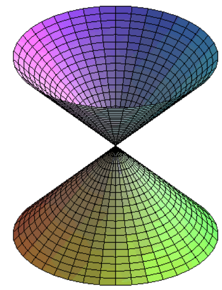
Now, condition (iii) of Definition 3.3 is also precisely what allows us to find the tangent *plane* to a point. It implies that the partials σ_u and σ_v are



Sphere with latitude-longitude grid [Com07].



Torus generated by rotating a circle in the xz-plane about the z-axis.



A circular cone is not a regular surface [Com06].

linearly independent, so their span must be a two-dimensional linear subspace. We begin defining the tangent by considering smooth curves on the surface.

Definition 3.10. Let p be any point on a surface $\mathcal{S} \subset \mathbb{R}^3$. If $\gamma : (-\epsilon, \epsilon) \to \mathcal{S}$ is a path with $\gamma(0) = p$, then **tangent vector** to $\mathcal{S}$ at p is precisely $\dot{\gamma}(0)$, the tangent vector to γ at p. The **tangent space** of $\mathcal{S}$ at p, denoted $T_p\mathcal{S}$, is the set of all vectors tangent to $\mathcal{S}$ at p.

PROPOSITION 3.11. *Let p be a point on a surface $\mathcal{S} \subset \mathbb{R}^3$, and suppose $\sigma : U \to \mathbb{R}^3$ is a surface patch whose image contains p, say $p = \sigma(u_0, v_0)$. Then the tangent space of $\mathcal{S}$ at p is the vector subspace*

$$T_p\mathcal{S} = \operatorname{span}(\sigma_u, \sigma_v),$$

where σ_u, σ_v are the partial derivatives evaluated at p.

Proof. We will prove these two spaces are equal using double containment. First, if γ is a path in the image of a surface patch σ, then we have

$$\gamma(t) = \sigma(u(t), v(t))$$

for some smooth functions $u(t)$ and $v(t)$. The existence of such smooth functions follows from properties (i)–(iii) of a surface, which imply σ^{-1} is smooth. Differentiating with the chain rule, we have

$$\dot{\gamma} = \sigma_u du + \sigma_v dv,$$

so every tangent vector of $\mathcal{S}$ can be written as a linear combination of the partials σ_u and σ_v. Thus, we have $T_p\mathcal{S} \subset \operatorname{span}(\sigma_u, \sigma_v)$.

On the other hand, we can write every vector $\vec{v} \in \operatorname{span}(\sigma_u, \sigma_v)$ as a linear combination $\vec{v} = a_1\sigma_u + a_2\sigma_v$ for some coefficients $a_1, a_2 \in \mathbb{R}$. Then we can define a curve

$$\gamma(t) = \sigma(u_0 + a_1 t, v_0 + a_2 t).$$

At the point $p = \gamma(0) \in \mathcal{S}$, we have

$$\dot{\gamma}(0) = a_1\sigma_u + a_2\sigma_v = \vec{v},$$

so every vector in the span is the tangent vector of $\mathcal{S}$ at some point p. This shows $\text{span}(\sigma_u, \sigma_v) \subset T_p\mathcal{S}$, so we must have exactly $\text{span}(\sigma_u, \sigma_v) = T_p\mathcal{S}$. □

4. **The first fundamental form and surface area**

To describe the local geometry of a surface, we need a way to make local measurements like lengths, angles, and areas. The first fundamental form allows us to compute the length of a curve on a surface using tangent vectors.

Definition 4.1. Let $p \in \mathcal{S}$ be any point of a surface. The **first fundamental form** of $\mathcal{S}$ at p is given by

$$\mathbf{I}_p(\vec{v}, \vec{w}) = \langle \vec{v}, \vec{w} \rangle,$$

where $\vec{v}, \vec{w} \in T_p\mathcal{S}$ are tangent vectors. That is, the first fundamental form I_p is the standard inner product on $\mathbb{R}^3$ restricted to the tangent space $T_p\mathcal{S}$.

In practice, this form is expressed in terms of surface patches. Suppose $p = \sigma(u_0, v_0)$ for some surface patch σ so that partial derivatives $\{\sigma_u, \sigma_v\}$ evaluated at p form a basis for the tangent plane $T_p\mathcal{S}$. Then any tangent vector $\vec{v} \in T_p\mathcal{S}$ is tangent to a curve γ in the image of σ given by $\gamma(t) = \sigma(u(t), v(t))$. As shown in the proof of Proposition 3.11, we can express the tangent vector as a linear combination $\vec{v} = \dot{\gamma}(0) = \sigma_u du + \sigma_v dv$.

We use the fact that the inner product is symmetric bilinear to expand $\mathbf{I}_p$ as the quadratic form

$$\begin{aligned}\mathbf{I}_p(\vec{v}, \vec{v}) &= \langle \sigma_u du + \sigma_v dv, \sigma_u du + \sigma_v dv \rangle \\ &= \langle \sigma_u, \sigma_u \rangle (du)^2 + 2\langle \sigma_u, \sigma_v \rangle du dv + \langle \sigma_v, \sigma_v \rangle (dv)^2.\end{aligned}$$

Traditionally, the inner product components of this form are denoted

$$E = \langle \sigma_u, \sigma_u \rangle \qquad F = \langle \sigma_u, \sigma_v \rangle \qquad G = \langle \sigma_v, \sigma_v \rangle,$$

and the expression $Edu^2 + 2Fdudv + Gdv^2$ is called the first fundamental form of the surface patch $\sigma(u, v)$. Note that the linear maps du, dv and metric coefficients E, F, G depend on choice of parametrization σ, but the form itself only depends on $\mathcal{S}$ and point p.

Finally, when γ is a curve in the image of a patch σ, we can substitute the first fundamental form of σ in the arc length formula to compute

$$\int \|\dot{\gamma}(t)\| dt = \int \sqrt{\langle \dot{\gamma}, \dot{\gamma} \rangle} dt = \int \sqrt{Edu^2 + 2Fdudv + Gdv^2} dt.$$

Example 4.2. For a surface of revolution with unit-speed profile curve $u \mapsto (f(u), 0, g(u))$, we have

$$\sigma_u = (\dot{f}\cos v,\ \dot{f}\sin v,\ \dot{g}) \qquad \sigma_v = (-f\sin v,\ f\cos v,\ 0).$$

Using the fact that $\dot{f}^2 + \dot{g}^2 = 1$ for the unit-speed curve, we compute the coefficients $E = 1$, $F = 0$, and $G = f^2$. Thus, the first fundamental form is $du^2 + f(u^2)dv^2$. ♢

Example 4.3. For the parametrization of S^2 as a surface of revolution, we have $f(u) = \cos(u)$ and $g(u) = \sin(u)$. The corresponding first fundamental form is $du^2 + \cos^2(u)dv^2$. ♢

Since the Gauss–Bonnet theorem involves integrating over a surface, we will briefly discuss areas of surface regions.

Definition 4.4. Given a surface patch $\sigma : U \to \mathbb{R}^3$ and a subset $R \subseteq U$, the **area** $A_\sigma(R)$ of the surface region $\sigma(R)$ is

$$A_\sigma(R) = \int_R \|\sigma_u \times \sigma_v\| dudv.$$

Using the first fundamental form to compute $\|\sigma_u \times \sigma_v\| = \sqrt{EG - F^2}$, we can further write

$$dA = \sqrt{EG - F^2} dudv.$$

Importantly, since the value $EG - F^2 = \det(\mathbf{I}_p)$ does not depend on choice of basis, the area of a surface region does not depend on choice of patch σ. This agrees with the remark about reparametrizations and geometric properties at the beginning of Section 3.

Example 4.5. Recall that a general surface of revolution has parametrization $\sigma : I \times [0, 2\pi)$ for some interval $I \subset \mathbb{R}$, so the surface area is computed by

$$A(\mathcal{S}) = \int_{\mathcal{S}} 1 dA = \int_{I \times [0,2\pi)} \sqrt{G - 0} dudv = \int_0^{2\pi} \int_I f(u) dudv.$$

♢

Example 4.6. The surface area of the unit sphere is

$$A(S^2) = \int_{S^2} 1 dA = \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \cos(u) dudv = 4\pi.$$

♢

5. The second fundamental form and surface curvature

In the same way that a plane curve's signed curvature $\kappa = d\theta/ds$ is a ratio defined by associating an infinitesimal change $\dot{\gamma}$ with an infinitesimal angle $\dot{\theta}$

on the unit circle, the curvature of a surface in $\mathbb{R}^3$ is defined by associating an infinitesimal area element $dA = dudv$ with another infinitesimal area element $d\sigma$ on the unit sphere. The **Gaussian curvature** is precisely the ratio $K = dA/d\sigma$.

In practice, we can measure curvature by considering how the the unit normal $\mathbf{N}$ varies as we move around the surface. For the tangent plane $T_p\mathcal{S}$, Proposition 3.11 makes a choice of normal vector straightforward: if $\sigma : U \to \mathbb{R}^3$ is a surface patch which contains p, then the unit vector

$$\mathbf{N}_\sigma = \frac{\sigma_u \times \sigma_v}{\|\sigma_u \times \sigma_v\|}$$

is perpendicular to every linear combination of σ_u and σ_v. We call $\mathbf{N}_\sigma$ the **standard unit normal** of the patch σ at point p.

While $\pm\mathbf{N}$ does not depend on choice of surface patch σ, the parametrization determines the sign. In order for the integration of functions to be well-defined, we will only consider surfaces which are *orientable*, meaning we have a smooth choice of normal $\mathbf{N}$. Informally, an orientable surface has two sides; the typical example of a non-orientable surface is the Mobius strip (see [Pre10, Example 4.5.3]). Importantly, working with orientable surfaces means **we assume that all surface patches discussed in the paper will have a standard unit normal that is the same as the chosen normal N.**

The values of $\mathbf{N}$ are given by the **Gauss map G** $: \mathcal{S} \to S^2$, which sends each point $p \in \mathcal{S}$ to its standard unit normal $\mathbf{N}_p$ in the unit sphere. Since we are interested in the rate of change of $\mathbf{N}$, we need to define the derivative $d\mathbf{G}_p$ at each point. In general, given a map f between two surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$, the derivative of f is the linear map $df_p : T_p\mathcal{S}_1 \to T_{f(p)}\mathcal{S}_2$ which "pushes forward" the tangent vector to the curve $p = \gamma(0)$ in $\mathcal{S}_1$ to the tangent at $(f \circ \gamma)(0)$ in $\mathcal{S}_2$. Thus, the derivative of the Gauss map is a function

$$d\mathbf{G}_p : T_p\mathcal{S} \to T_{\mathbf{G}(p)}S^2.$$

Now by definition, $T_{\mathbf{N}_p}S^2$ is the plane through the origin perpendicular to the point $\mathbf{G}(p) = \mathbf{N}_p$, which is precisely $T_p\mathcal{S}$, so the derivative $d\mathbf{G}_p$ is actually a map from $T_p\mathcal{S}$ to itself.

Definition 5.1. Let $\mathcal{S}$ be an orientable surface with Gauss map $\mathbf{G}$. For each $p \in \mathcal{S}$, the **Weingarten map** of $\mathcal{S}$ at p is the linear map $\mathbf{W} : T_p\mathcal{S} \to T_p\mathcal{S}$ is given by

$$\mathbf{W}_p = -d\mathbf{G}_p.$$

Definition 5.2. If $\mathbf{W}_p$ is the Weingarten map at a point $p \in \mathcal{S}$, the **Gaussian curvature** K of $\mathcal{S}$ at p is given by

$$K = \det(\mathbf{W}_p).$$

Remark 5.3. The Gaussian curvature does not depend on orientation of the tangent plane, as the determinant of the 2×2 matrix $\mathbf{W}_p$ is the same when every entry changes sign.

Example 5.4. The Gaussian curvature of S^2 is 1 everywhere, because the Gauss map at every point in S^2 is the precisely the identity map. Thus, the Weingarten map at every point is also the identity, and so $K = \det(I) = 1$. $\Diamond$

Unfortunately, most Weingarten maps are not so obvious. To get an explicit formula for K, we need to define a metric for curvature on a surface patch σ.

Definition 5.5. The **second fundamental form** of $\mathcal{S}$ at p is the bilinear map $\mathbf{II}_p : T_p\mathcal{S} \to \mathbb{R}$ defined by

$$\mathbf{II}_p = \langle \mathbf{W}_p(\vec{v}), \vec{w} \rangle$$

for some tangent vectors $\vec{v}, \vec{w} \in T_p\mathcal{S}$.

Unlike with the form $\mathbf{I}_p$, it is not immediately clear that $\mathbf{II}_p$ has a corresponding quadratic function.

Proposition 5.6. *The second fundamental form is* symmetric bilinear. *That is, for all tangent vectors $\vec{v}, \vec{w} \in T_p\mathcal{S}$, we have $\mathbf{II}_p(\vec{v}, \vec{w}) = \mathbf{II}_p(\vec{w}, \vec{v})$.*

Proof. First, let $p \in \mathcal{S}$ be a point in the image of a surface patch σ. Suppose $\gamma(t) = \sigma(u(t), v(t))$ is a curve in the patch with $\gamma(0) = p$, so

$$\dot{\gamma}(0) = \sigma_u du(0) + \sigma_v dv(0)$$

is tangent to $\mathcal{S}$ at p. Then

$$\begin{aligned} \mathbf{W}_p(\dot{\gamma}(0)) &= -d\mathbf{G}_p(\sigma_u du(0) + \sigma_v dv(0)) \\ &= -\frac{d}{dt}\mathbf{G}(u(t), v(t))\Big|_{t=0} \\ &= -\left(\mathbf{G}_u du(0) + \mathbf{G}_v dv(0)\right). \end{aligned}$$

In particular, since

$$du(\sigma_u) = dv(\sigma_v) = 1 \qquad du(\sigma_v) = dv(\sigma_u) = 0,$$

we have $\mathbf{W}_p(\sigma_u) = -\mathbf{G}_u$ and $\mathbf{W}_p(\sigma_v) = -\mathbf{G}_v$.

Since $\{\sigma_u, \sigma_v\}$ is a basis for $T_p\mathcal{S}$, we can write our tangent vectors as linear combinations $\vec{v} = a_1\sigma_u + a_2\sigma_v$ and $\vec{w} = b_1\sigma_u + b_2\sigma_v$ for some $a_1, a_2, b_1, b_2 \in \mathbb{R}$.

Using the fact that a bilinear form on $\mathbb{R}^n$ is linear in both inputs, we compute

$$\begin{aligned}\mathbf{II}_p(\vec{v},\vec{w}) &= \langle \mathbf{W}_p(\vec{v}),\vec{w}\rangle = \langle -a_1\mathbf{G}_u - a_2\mathbf{G}_v, b_1\sigma_u + b_2\sigma_v\rangle \\ &= -a_1b_1\langle \mathbf{G}_u,\sigma_u\rangle - a_1b_2\langle \mathbf{G}_u,\sigma_v\rangle - a_2b_1\langle \mathbf{G}_v,\sigma_u\rangle - a_2b_2\langle \mathbf{G}_v,\sigma_v\rangle \\ &= \langle -b_1\mathbf{G}_u - b_2\mathbf{G}_v, a_1\sigma_u + a_2\sigma_v\rangle \\ &= \langle \mathbf{W}_p(\vec{w}),\vec{v}\rangle = \mathbf{II}_p(\vec{w},\vec{v}),\end{aligned}$$

which shows the desired equality. □

We now obtain a quadratic form: given a tangent vector $\vec{v} = \sigma_u du + \sigma_v dv$, we have

$$\mathbf{II}_p(\vec{v},\vec{v}) = -\langle \mathbf{G}_u,\sigma_u\rangle(du)^2 - 2\langle \mathbf{G}_u,\sigma_v\rangle du dv - \langle \mathbf{G}_v,\sigma_v\rangle(dv)^2,$$

where the middle term uses the fact that $\langle \mathbf{G}_u,\sigma_v\rangle = \langle \mathbf{G}_v,\sigma_u\rangle$.

The metric coefficients are traditionally denoted

$$L = -\langle \mathbf{G}_u,\sigma_u\rangle \qquad M = -\langle \mathbf{G}_u,\sigma_v\rangle \qquad N = -\langle \mathbf{G}_v,\sigma_v\rangle,$$

and we say $Ldu^2 + 2Mdudv + Ndv^2$ is the second fundamental form of the surface patch $\sigma(u,v)$.

Together with the first fundamental form, the second fundamental form gives us a very useful formula for Gaussian curvature. If we write $-\mathbf{G}_u$ and $-\mathbf{G}_v$ in terms of the basis $\{\sigma_u,\sigma_v\}$, then the explicit matrix for the Weingarten map with respect to this basis is

$$\begin{pmatrix} E & F \\ F & G\end{pmatrix}^{-1}\begin{pmatrix} L & M \\ M & N\end{pmatrix}$$

(for the full derivation, see [Pre10, Proposition 8.1.2]). Thus, we have

$$K = \frac{LM - M^2}{EG - F^2}.$$

Example 5.7. A sphere of radius c has Gaussian curvature $1/c^2$ everywhere. This is because when a surface is scaled by some constant c, the coefficients E, F, G are multiplied by a factor of c^2 and the coefficients L, M, N are multiplied by a factor of c, so K changes by a factor of $1/c^2$.

Further, since the surface area changes by a factor of c^2, we find the *total curvature* of any sphere $\mathcal{S}$ is

$$\int_{\mathcal{S}} K dA = \frac{1}{a^2}\cdot 4\pi a^2 = 4\pi.$$

◊

Example 5.8. Using the parametrization of the torus from Section 3, we compute the partials

$$\sigma_u = \begin{pmatrix} -r\sin(u)\cos(v) \\ -r\sin(u)\sin(v) \\ r\cos(v) \end{pmatrix} \qquad \sigma_v = \begin{pmatrix} -(R + r\cos(u)\sin(v) \\ (R + r\cos(u))\cos(v) \\ 0 \end{pmatrix}.$$

The coefficients for the first fundamental form are $E = r^2$, $F = 0$, and $G = (R + r\cos(u))^2$, and the coefficients for the second are $L = r$, $M = 0$, and $N = (R + r\cos\theta)\cos\theta$. The Gaussian curvature is then

$$K = \frac{\cos(u)}{r(R + r\cos(u))}.$$

Interestingly, the torus has both positive and negative curvature: we have $K \geq 0$ when $\pi/2 \leq u \leq \pi/2$, and $K \leq 0$ when $\pi/2 \leq v \leq 3\pi/2$. $\diamondsuit$

6. The local Gauss–Bonnet theorem

The most basic version of the Gauss–Bonnet theorem applies to simple closed curves on a surface. In Section 2, we considered the particular case where the surface is a plane, where the Gaussian curvature is 0. Our next step is to extend the Umlaufsatz to curved surfaces.

Definition 6.1. Given an open subset $U \subset \mathbb{R}^2$ and a local parametrization $\sigma : U \to \mathcal{S}$, we say $\gamma : [a, b] \to \mathbb{R}^3$ is a **simple closed curve** in the patch $\sigma(U)$ if there exists a simple closed plane curve $\beta(t) = (u(t), v(t))$ such that $\gamma = \sigma \circ \beta$.

In this case, γ is **positively-oriented** if the signed unit normal $\mathbf{n}$ of β points into $\text{int}(\beta) \subset \mathbb{R}^2$ at every point of β. Finally, $\text{int}(\gamma) \subset \mathbb{R}^3$ is defined as the image of $\text{int}(\beta)$ under the map σ.

Lemma 6.2. *In the situation above, we have*

$$\int_\gamma \dot{\theta}(s)ds = \pm 2\pi.$$

Proof. Briefly, we can find an isotopy between γ and any another simple closed curve $\tilde{\gamma}$ that is completely contained in $\text{int}(\gamma)$. We choose $\tilde{\gamma}$ to be the image under surface patch σ of a *very* small circle in $\text{int}(\beta)$, so the interior of $\tilde{\gamma}$ is essentially a subset of the plane in $\mathbb{R}^2$. Then using Lemma 2.14, we can replace γ with $\tilde{\gamma}$ in the above integral, and the equality follows from Hopf's Umlaufsatz.

For the first isotopy, let $p = \sigma(u_0, v_0)$ be a point in $\text{int}(\gamma) = \sigma(\text{int}(\beta))$. By property (iii) of regular surfaces, we can scale the axes of $\mathbb{R}^3$ to obtain a patch $\tilde{\sigma}(V) \subset \sigma(U)$ containing p with

$$\tilde{\sigma}(x, y) = (x, y, f(x, y))$$

for some smooth map f. By the same property, we may translate the surface so that $p = \tilde{\sigma}(u_0, v_0)$. Then $\sigma^{-1}(\tilde{\sigma}(V))$ is an open subset of $U \subset \mathbb{R}^2$, so there exists an $\epsilon > 0$ such that $\sigma(B_\epsilon(p)) \subset \tilde{\sigma}(V)$.

Now, consider the isotopy of curves given by

$$h_1(s,t) = \sigma(t \cdot u(s), t \cdot v(s)).$$

By choosing sufficiently small t, such as $t = \epsilon/2$, we can find an isotopy between our original curve $\gamma = h_1(s, 1)$ and a curve in $\tilde{\sigma}(V)$. Note that such a curve has the form $\gamma_{\epsilon/2}(s) = (x(s), y(s), f\,(x(s), y(s)))$ for some smooth functions $x(s)$ and $y(s)$.

Using this, we define a second isotopy of curves in $\tilde{\sigma}(V)$ by

$$h_2(s,t) = (x(s), y(s), t \cdot f(x(s), y(s))).$$

This gives an isotopy between $\gamma_{\epsilon/2} = h_1(s, \epsilon/2) = h_2(s, 1)$ and the simple plane curve $\tilde{\gamma} = h_2(s, 0)$. Then by Lemma 2.14, we have

$$\int_\gamma \dot{\theta} ds = \int_{\gamma_{\epsilon/2}} \dot{\theta} ds = \int_{\tilde{\gamma}} \dot{\theta} ds,$$

and the final integral is equal to $\pm 2\pi$ by Theorem 2.15. □

Remark 6.3. A more sophisticated version of this proof will define the *relative index* of a curve with respect to an orthonormal basis, then use the Gram–Schmidt process to produce a smooth family of bases for curves in $\tilde{\sigma}(V)$. After obtaining the plane curve $\tilde{\gamma}$, the final step is to show that the relative index of $\tilde{\gamma}$ coincides with the formula for $I(\tilde{\gamma})$ (see [Swa, Theorem 6.6]).

Our definition of a curve's curvature also requires adjustment. Notice that given any curve γ on a surface $\mathcal{S}$, the set $\{\dot{\gamma}, \mathbf{N}, \mathbf{N} \times \dot{\gamma}\}$ is an orthonormal basis for $\mathbb{R}^3$. Recall that when γ is unit-speed, its absolute curvature is given by $\kappa = \|\ddot{\gamma}\|$. There is a particular term for the projection of κ on the the tangent plane of $\mathcal{S}$.

Definition 6.4. If γ is a unit-speed curve on an surface $\mathcal{S}$, then the **geodesic curvature** of γ is defined by

$$\kappa_g = \ddot{\gamma} \cdot (\mathbf{N} \times \dot{\gamma}).$$

Remark 6.5. Informally, κ_g measures how far the curve is from being the shortest path between two points on a surface. When the surface is a plane, the shortest path is a straight line, so a plane curve in $\mathbb{R}^3$ has $\kappa_g = \kappa$ up to a sign. In general, the sign of the geodesic curvature κ_g of a curve depends on the orientation of both the surface and the curve itself.

We are now ready to prove the Gauss–Bonnet theorem for simple closed curves.

THEOREM 6.6. *Let γ be a unit-speed simple closed curve on a surface patch σ, and suppose γ is positively-oriented. Then*

$$\int_{\gamma} \kappa_g ds = 2\pi - \int_{int(\gamma)} K dA,$$

*where κ_g is the geodesic curvature of γ, K is the Gaussian curvature of σ, and dA is the area element of σ. The integral over the area element is called the **total curvature** of the region $int(\gamma)$.*

Proof. The argument is entirely computational. First, we will use a basis of the tangent plane to find an orthonormal basis for $\mathbb{R}^3$, then expand $\dot{\gamma}$ and $\ddot{\gamma}$ in terms of this basis. We then use this to compute κ_g, which allows us to rewrite the integral of κ_g over the curve γ as the difference of two integrals. Finally, we evaluate the integrals separately to obtain the expression on the right; the 2π term will come from a direct application of Hopf's Umlaufsatz for surface curves, while the area integral uses both fundamental forms of the surface patch σ.

Let $\{\mathbf{e}_1, \mathbf{e}_2\}$ be a smooth[3] orthonormal basis for the tangent plane at each point in the image of σ; one such choice is $\mathbf{e}_1 = \sigma_u / \|\sigma_u\|$ and $\mathbf{e}_2 = \mathbf{N} \times \mathbf{e}_1$. Then $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{N}\}$ is an orthonormal basis for $\mathbb{R}^3$. Note that since we can always swap values of $\mathbf{e}_1$ and $\mathbf{e}_2$ if necessary, we assume $\mathbf{N} = \mathbf{e}_1 \times \mathbf{e}_2$ without loss of generality.

Now, let $\theta(s)$ be the *oriented angle* between the tangent vector $\dot{\gamma}(s)$ and the basis vector $\mathbf{e}_1$. This is the angle by which $\mathbf{e}_1$ must be rotated to be parallel to $\dot{\gamma}$, when viewing the side of the surface which $\mathbf{N}$ points away from. That is, from this side, $\theta(s)$ is precisely the tangent angle from Definition 2.6 taken with respect to $\mathbf{e}_1$ instead of the standard basis. Thus, we have

$$\dot{\gamma} = \cos\theta \mathbf{e}_1 + \sin\theta \mathbf{e}_2$$

$$\ddot{\gamma} = \cos\theta \dot{\mathbf{e}}_1 + \sin\theta \dot{\mathbf{e}}_2 + \dot{\theta}(-\sin\theta \mathbf{e}_1 + \cos\theta \mathbf{e}_2),$$

where the expression for $\ddot{\gamma}$ uses the chain rule. Substituting these expressions and $\mathbf{N} = \mathbf{e}_1 \times \mathbf{e}_2$ into the formula for geodesic curvature, we find that

$$\kappa_g = \dot{\theta} - \mathbf{e}_1 \cdot \dot{\mathbf{e}}_2$$

(for full computations, see [Pre10, Theorem 13.1.2]). We can therefore compute the left side of the claimed equality as

$$\int_{\gamma} \kappa_g ds = \int_{\gamma} \dot{\theta} ds - \int_{\gamma} \mathbf{e}_1 \cdot \dot{\mathbf{e}}_2 ds.$$

[3] Here, "smooth" means that $\mathbf{e}_1, \mathbf{e}_2$ are smooth functions of the surface parameters (u, v).

First, we know know from Lemma 6.2 that the integral of $\dot{\theta}$ around γ is equal to $\pm 2\pi$; since γ is positively-oriented, this is exactly 2π. It remains to show that

$$\int_\gamma \mathbf{e}_1 \cdot \dot{\mathbf{e}}_2 ds = \int_{\text{int}(\gamma)} K dA.$$

Differentiating $\mathbf{e}_2$, we have

$$\int_\gamma \mathbf{e}_1 \cdot \dot{\mathbf{e}}_2 \ ds = \int_\gamma \mathbf{e}_1 \cdot ((\mathbf{e}_2)_u \dot{u} + (\mathbf{e}_2)_v \dot{v}) \ ds = \int_\beta (\mathbf{e}_1 \cdot (\mathbf{e}_2)_u) du + (\mathbf{e}_1 \cdot (\mathbf{e}_2)_v) dv$$
$$= \int_{\text{int}(\beta)} [(\mathbf{e}_1 \cdot (\mathbf{e}_2)_v)_u - (\mathbf{e}_1 \cdot (\mathbf{e}_2)_u)_v] \ dudv,$$

where the last equality uses Green's theorem (see [Shi, Appendix 2, Theorem 2.6]). Now given the first and second fundamental forms of σ,

$$Edu^2 + 2Fdudv + Gdv^2 \qquad Ldu^2 + 2Mdudv + Ndv^2,$$

we can write the partial derivatives of $\mathbf{e}_1$ and $\mathbf{e}_2$ in terms of the basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{N}\}$ to see that

$$(\mathbf{e}_1)_u \cdot (\mathbf{e}_2)_v - (\mathbf{e}_1)_v \cdot (\mathbf{e}_2)_u = \frac{LN - M^2}{(EG - F^2)^{1/2}}$$

(for full computations with coefficients, see [Pre10, Lemma 13.1.3]). Then applying the formulas for dA and K, this integral becomes

$$\int_\gamma \mathbf{e}_1 \cdot \dot{\mathbf{e}}_2 ds = \int_{\text{int}(\beta)} \frac{LN - M^2}{(EG - F^2)^{1/2}} dudv = \int_{\text{int}(\gamma)} \frac{LN - M^2}{EG - F^2} dA = \int_{\text{int}(\gamma)} K dA,$$

where β is the simple closed plane curve specified in Definition 6.1. This completes the proof. □

For the remainder of this paper, our discussion will be in terms of regions on surfaces rather than curves. By **region**, we mean a compact, simply connected subset $\triangle$ of a surface $\mathcal{S}$. We will only consider regions with *piecewise smooth* boundaries, which means the boundary $\partial\triangle$ looks like a polygon with curved sides, or possibly a simple closed curve with no vertices.

Definition 6.7. The boundary $\partial\triangle$ is **positively-oriented** if, for all t such that $\gamma_i(t)$ is not a vertex, the signed unit normal $\mathbf{n}$ obtained by rotating $\dot{\gamma}_i$ counterclockwise by $\pi/2$ points into $\triangle$.

The next version of the Gauss–Bonnet theorem accounts for boundary vertices, where a single oriented angle is undefined, by using exterior angles. Given a vertex v of the polygon, we have one curved edge γ_i traveling towards v and another edge γ_j traveling away. As in the beginning of the proof of Theorem 6.6, take $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{N}\}$ to be a smooth orthonormal basis of $\mathbb{R}^3$, and let θ_i and θ_j be the oriented angles of $\dot{\gamma}_i$ and $\dot{\gamma}_j$ at v, respectively. The **exterior**

angle at υ is given by $\delta = \theta_j - \theta_i$. Since this is only well-defined up to multiples of 2π, we assume $-\pi < \delta < \pi$.

THEOREM 6.8 (Local Gauss–Bonnet). *Let R be a simply connected region with piecewise smooth boundary in a surface path σ. If the boundary $\partial\triangle$ is positively-oriented, then we have*

$$\int_{\partial\triangle} \kappa_g ds = 2\pi - \sum_{i=1}^{n} \delta_i - \int_{\triangle} K dA,$$

where δ_i is the exterior angle for some vertex $i = 1, \dots, n$.

Proof. This is essentially a generalization of Theorem 6.6 to curves with "corners." Applying the same argument as before, we find

$$\int_{\partial\triangle} \kappa_g ds = \int_{\partial\triangle} \dot{\theta} ds - \int_{\triangle} K dA.$$

It remains to show that

$$\int_{\partial\triangle} \dot{\theta} ds = 2\pi - \sum_{i=1}^{n} \delta_i.$$

The strategy is to approximate $\partial\triangle$ with a smooth curve γ which rounds off the corners. We know by Lemma 6.2 that the total turning angle going once around γ is exactly 2π. Now notice that since $\partial\triangle$ is piecewise smooth, the integral on the left-hand side of the equality is really the sum of n integrals along the edges of the polygon, with the turning angle at each vertex excluded from the total. We therefore take γ to be a close-enough approximation such that the difference between 2π and $\int_{\partial\triangle} \dot{\theta}$ is only due to these vertex angles, and the equality follows (for a more rigorous argument, see [Pre10, Theorem 13.2.2]). ☐

Example 6.9. Consider an n-gon on the plane with straight edges. In this case, we have $K = 0$ and $\kappa_g = 0$ for each side of the polygon. An **internal angle** of the polygon is given by $\alpha_i = \pi - \delta_i$ for $i = 1, \dots, n$ and $0 < \alpha_i < 2\pi$. Then Theorem 6.8 implies

$$\sum_{i=1}^{n} \alpha_i = (n-2)\pi,$$

a well-known formula from elementary geometry. ◊

7. The global Gauss–Bonnet theorem

The most general version of the Gauss–Bonnet theorem applies to compact, oriented surfaces with piecewise smooth boundary. We will take **compact** to mean closed and bounded, although compactness is technically a generalization of these properties to higher-dimensional Euclidian subsets. Roughly speaking, any such surface may be covered with a specific arrangement of finitely many "polygons," and we can find the entire surface's curvature by applying the local Gauss–Bonnet theorem to each polygon and taking the sum.

Definition 7.1. A surface $\mathcal{S} \subset \mathbb{R}^3$ can be **triangulated** if it is possible to write $\mathcal{S} = \bigcup_{\lambda=1}^{F} \triangle_\lambda$, where

(i) Each $\triangle_\lambda$ is the image of a triangle under a local parametrization σ;
(ii) For all $\lambda \neq \mu$, the intersection $\triangle_\lambda \cap \triangle_\mu$ is either empty, a single vertex, or a single edge;
(iii) When $\triangle_\lambda \cap \triangle_\mu$ is a single edge, the orientations of the edge are opposite in $\triangle_\lambda$ and $\triangle_\mu$;
(iv) For all λ, at most one edge $\triangle_\lambda$ is contained in $\partial\mathcal{S}$.

In this case, each region $\triangle_\lambda$ is called a **face**, and a collection of such faces is called a **triangulation** of $\mathcal{S}$.

Remark 7.2. The choice of compatible orientation in (iii) gives us an orientation on the boundary of $\mathcal{S}$, which comes from the normal $\mathbf{N}$ and orientation of $\mathcal{S}$ itself. However, we do not need to worry about boundary orientation when integrating κ_g in the theorem. If we have instead $-\mathbf{N}$, then the orientation on $\partial\mathcal{S}$ swaps while $\mathbf{N} \times \dot{\gamma}$ is unchanged, so the sign of κ_g on $\partial\mathcal{S}$ does not depend on choice of orientation on $\mathcal{S}$.

THEOREM 7.3. *Every compact surface has a triangulation with finitely many faces.*

The proof of this theorem, which comes from algebraic topology, has a relatively simple idea. For every point $p \in S$, we can find a small disc containing p, and we know $\mathcal{S}$ can be covered by a finite collection of these discs because the surface is compact. We can triangulate the interior of each disc, then paste them together to make a surface homeomorphic to $\mathcal{S}$. The challenge with a formal proof is adjusting for how the discs may overlap (see [DM68]).

We now define the topological invariant of interest in the final Gauss–Bonnet theorem.

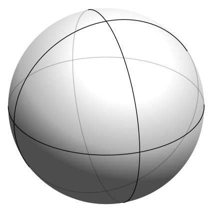
Definition 7.4. For any triangulation of a surface $\mathcal{S}$, the **Euler characteristic** of the triangulation is given by

$$\chi(\mathcal{S}) = V - E + F,$$

where V, E, and F denote the total number of vertices, edges, and faces, respectively.

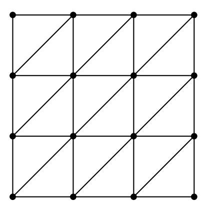
THEOREM 7.5. *Let $\mathcal{S}$ be a surface equipped with a triangulation. If $\mathcal{S}$ is homeomorphic to another surface $\mathcal{S}'$, then $\chi(\mathcal{S}) = \chi(\mathcal{S}')$.*

Example 7.6. One triangulation of S^2 is found by intersecting the sphere with three coordinate planes.



This triangulation has eight faces, and its Euler characteristic is $6-12+8=2$. $\diamondsuit$

Example 7.7. To triangulate the torus, we use the fact that the torus is homeomorphic to a square: roll the square into a tube, then stretch the tube so that the two ends meet as a donut. A triangulation of the square is shown below.



Taking into account that opposite sides of the squares will meet once rolled into the torus, we find the Euler characteristic of this triangulation to be $9 - 27 + 18 = 0$. $\diamondsuit$

While different triangulations of a surface $\mathcal{S}$ may have different numbers of vertices, edges, and faces, the Euler characteristic $\chi(\mathcal{S})$ only depends on the surface itself. This important property is a consequence of the final Gauss–Bonnet theorem.

THEOREM 7.8 (Global Gauss–Bonnet). *Let $\mathcal{S} \subset \mathbb{R}^3$ be a compact, oriented surface with piecewise smooth boundary. Then*

$$\int_{\partial \mathcal{S}} \kappa_g ds + \int_{\mathcal{S}} K dA + \sum_{i=1}^{n} \delta_i = 2\pi\chi(\mathcal{S}),$$

where δ_i with $i = 1, \ldots, n$ is an exterior angle of $\partial\mathcal{S}$.

Notice that because the left-hand side of the equality has nothing to do with a chosen triangulation, our proof will hold for any choice of triangulation for $\mathcal{S}$. Theorem 7.8 therefore implies the following corollary.

COROLLARY 7.9. *The Euler characteristic $\chi(\mathcal{S})$ of a compact surface $\mathcal{S}$ is independent of the choice of triangulation.*

Proof of Theorem 7.8. As mentioned, the main idea is to apply the local Gauss–Bonnet theorem to each $\triangle_\lambda$ of the triangulation, then use the total to compute each term on the left-hand side. The integral values are easy to find, but we need some additional geometric reasoning to find the difference between the total exterior angle of $\partial\mathcal{S}$ and the sum of the total exterior angles for all polygons of the triangulation.

We begin by expressing the integrals over $\mathcal{S}$ in terms of the triangulation. For the integral of κ_g on the boundary, we know from Definition 7.1 (iii) that any edge of the triangulation which is not in $\partial\mathcal{S}$ will be paired with an edge of the opposite orientation. Because κ_g changes sign when the orientation of the curve is reversed, the integral of κ_g on non-boundary edges cancels out in pairs. As for the area integral, the area of $\mathcal{S}$ is the sum of each $\triangle_\lambda$ by definition. Thus, we have

$$\int_{\partial\mathcal{S}} \kappa_g ds = \sum_{\lambda=1}^{F} \int_{\partial\triangle_\lambda} \kappa_g ds \qquad \int_{\mathcal{S}} K dA = \sum_{\lambda=1}^{F} \int_{\partial\triangle_\lambda} K dA.$$

Now, we compute the total curvature of each region $\triangle_\lambda$. Let δ_{λ_j} for $j = 1, 2, 3$ denote an exterior angle of $\triangle_\lambda$. Applying Theorem 6.8, we have

$$\int_{\partial\triangle_\lambda} \kappa_g ds + \int_{\triangle_\lambda} K dA + \sum_{j=1}^{3} \delta_{\lambda_j} = 2\pi,$$

and the sum over all of the $\triangle_\lambda$ is

$$\int_{\partial\mathcal{S}} \kappa_g ds + \int_{\mathcal{S}} K dA + \sum_{\lambda=1}^{F} \sum_{j=1}^{3} \delta_{\lambda_j} = 2\pi F.$$

To complete the proof, we just need to show that the difference between the sum in the previous expression and the total exterior angle of $\partial\mathcal{S}$ is exactly

$$\sum_{\lambda,j} \delta_{\lambda_j} - \sum_{i=1}^{n} \delta_i = 2\pi(E - V).$$

This is merely a matter of counting. We first make a distinction between vertices of triangulation on the boundary and in the interior of $\mathcal{S}$, denoting the respective totals by V_B and V_I. We do the same for edges of the triangulation that are on the boundary, edges in the interior, and edges that join a boundary vertex to an interior vertex, denoting these totals by E_B, E_I, and E_{IB}.

Letting α_{λ_j} denote the interior angles of the region $\triangle_\lambda$, we have

$$\sum_{\substack{\text{interior}\\\text{vertices}}} \delta_{\lambda_j} = \sum_{\substack{\text{interior}\\\text{vertices}}} (\pi - \alpha_{\lambda_j}) = \pi(2E_I + E_{IB}) - 2\pi V_I.$$

This is because each interior edge contributes two interior vertices to the total, while each interior/boundary edge contributes one. Further, the interior angles at each interior vertex sum to 2π.

On the other hand, given a boundary vertex υ, we will denote the associated angle or number with a superscript (υ). Every boundary vertex υ is contained in $E_{IB}^{(\upsilon)}+1$ faces. Moreover, the total interior angle at any boundary vertex is π if the vertex is on a smooth curve, and $\pi - \delta_i$ if the vertex is a "corner" of $\partial\mathcal{S}$ with exterior angle δ_i. Thus,

$$\begin{aligned}\sum_{\substack{\text{boundary}\\\text{vertices } \upsilon}} \delta_{\lambda_j} = \sum_{\substack{\text{boundary}\\\text{vertices } \upsilon}} (\pi - \alpha_{\lambda_j}) &= \sum_{\substack{\text{boundary}\\\text{vertices } \upsilon}} \pi(E_{IB}^{(\upsilon)} + 1) - \left(\sum_{\text{smooth } \upsilon} \alpha_{\lambda_j} + \sum_{\text{corner } \upsilon} \alpha_{\lambda_j} \right)\\ &= \pi E_{IB} + \sum_{i=1}^{n} \delta_i.\end{aligned}$$

Using the fact that $V_B = E_B$ for the closed polygon $\partial\mathcal{S}$, we find

$$\begin{aligned}\sum_{\lambda,j} \delta_{\lambda_j} &= \sum_{\substack{\text{interior}\\\text{vertices}}} \delta_{\lambda_j} + \sum_{\substack{\text{boundary}\\\text{vertices}}} \delta_{\upsilon_j} = 2\pi(E_I + E_{IB} - V_I) + \sum_{i=1}^{n} \delta_i\\ &= 2\pi(E_I + E_{IB} + E_B - V_I - V_B) + \sum_{i=1}^{n} \delta_i = 2\pi(E - V) + \sum_{i=1}^{n} \delta_i,\end{aligned}$$

as desired. At last, we conclude

$$\int_{\partial\mathcal{S}} \kappa_g ds + \int_{\mathcal{S}} K dA + \sum_{i=1}^{n} \delta_i = 2\pi F - 2\pi(E - V) = 2\pi\chi(\mathcal{S}).$$

□

For surfaces without boundary, sometimes called **closed** surfaces, we have the following remarkable result.

COROLLARY 7.10. *When $\mathcal{S} \subset \mathbb{R}^3$ is a compact, oriented surface without boundary, the total curvature of $\mathcal{S}$ is*

$$\int_{\mathcal{S}} K dA = 2\pi\chi(\mathcal{S}).$$

Example 7.11. If $\mathcal{S}$ is any sphere, we know $\chi(\mathcal{S}) = 2$, so the Gauss–Bonnet theorem says

$$\int_{\mathcal{S}} K dA = 4\pi.$$

This agrees with our computation at the end of Section 5. ♢

Example 7.12. If $\mathcal{S}$ is a torus, then $\chi(\mathcal{S}) = 0$ and the Gauss–Bonnet theorem says

$$\int_{\mathcal{S}} K dA = 0.$$

Earlier, we saw that the torus has both positively and negatively curved regions; we now know the positive and negative contributions cancel each other out. ♢

In this paper, we showed that the total curvature of a surface does not change with a deformation of the surface. Beyond our discussion, it is a theorem of topology that every compact, oriented surface without boundary is homeomorphic to a g-torus for some $g \geq 0$, where g is, roughly, the number of holes in the surface. Thus, the integral $\int_{\mathcal{S}} K dA$ is precisely what determines the topological classification of a surface.

References

[Ben17] Rachad Bentbib. *The Gauss-Bonnet Theorem*. PhD thesis, Universite Cadi Ayyad, June 2017.

[Com06] Wikimedia Commons. Cône surface réglée. https://commons.wikimedia.org/wiki/File:Cne_surface_rgle.png, 2006.

[Com07] Wikimedia Commons. Sphere (shaded). https://commons.wikimedia.org/wiki/File:Sphere_(Shaded).png, 2007.

[DM68] P. H. Doyle and D. A. Moran. A short proof that compact 2-manifolds can be triangulated. *Inventiones Mathematicae*, 5(2):160–162, June 1968.

[Kni06] Oliver Knill. Proof of the hopf umlaufsatz by deformation, Jun 2006.

[Pre10] Andrew Pressley. *Elementary Differential Geometry*. Springer Undergraduate Mathematics Series. Springer, London, New York, 2nd edition, 2010.

[Shi] Theodore Shifrin. Differential Geometry: A First Course in Curves and Surfaces.

[Swa] Andrew Swann. Supplementary Notes for MM08 Geometry I - Hopf's Umlaufsatz.

Department of Mathematics, Barnard College, Columbia University
E-mail: by2328@barnard.edu
https://journals.library.columbia.edu/index.php/cjum

Columbia Journal of Mathematics **1** (2024),
70–98

The Peter–Weyl theorem & harmonic analysis on S^n

By Luca Nashabeh

Abstract

For finite groups, the Artin–Wedderburn theorem gives a precise decomposition of the algebra of all $\mathbb{C}$-valued functions into matrix algebras. Specialized to the case of cyclic groups, this produces the classical discrete Fourier transform. In this paper, we endeavor to extend these techniques to compact topological groups, proving similar harmonic decompositions on S^1, S^2, and S^3.

1. **Introduction**

The representation theory of finite groups provides us with many powerful tools that not only allow us to directly study the properties and structures of groups, but also give insight into algebras defined on those groups. One of the most powerful results is the following theorem, which gives a relationship between the algebra of $\mathbb{C}$-valued functions on a finite group G and its irreducible representations.

Theorem 1.1 (Artin–Wedderburn theorem). *Let G be a finite group and $\mathbb{C}[G]$ its group algebra with the convolution product*

$$(f_1 * f_2)(g) = \sum_{h \in G} f_1(h) f_2(h^{-1}g).$$

Furthermore, let $\rho_i : G \to \mathrm{GL}(V_i)$ for $1 \leq i \leq k$ be the irreducible representations of G, and $\tilde{\rho}_i : \mathbb{C}[G] \to \mathrm{End}(V_i)$ the linear extensions to the group algebra. Then, the map

$$\tilde{\rho} = \bigoplus_{i=1}^{k} \rho_i, \quad \tilde{\rho} : \mathbb{C}[G] \to \bigoplus_{i=1}^{k} \mathrm{End}(V_i)$$

is an isomorphism.

A simple example of the power of the Artin–Wedderburn theorem comes from specializing to the cyclic case $G = \mathbb{Z}/n\mathbb{Z}$. Here, we can describe ρ_m and $\tilde{\rho}_m$ explicitly as

$$\rho_m(x) = \zeta_n^{mx} = \exp\left(\frac{2\pi i}{n} mx\right) \quad \text{and} \quad \tilde{\rho}_m(f) = \sum_{x\in G} f(x) \exp\left(\frac{2\pi i}{n} mx\right).$$

The Artin–Wedderburn theorem then gives us the following classical result.

COROLLARY 1.2 (Discrete Fourier transform). *Let $f \in \mathbb{C}[G] \cong \mathbb{C}^n$. Then f can be uniquely decomposed into pure frequencies with amplitudes*

$$F_m = \sum_{x=1}^{n} f(x) \exp\left(\frac{2\pi i}{n} mx\right).$$

More generally, the Artin–Wedderburn theorem allows us to do a Fourier decomposition on any finite group, including non-abelian ones. However, while the Artin–Wedderburn theorem is certainly a powerful result, the requirement of finiteness prevents us from getting a Fourier decomposition for many interesting continuous groups.

The Peter–Weyl theorem is one path to generalizing the Artin–Wedderburn theorem, proving a very similar result not just for finite groups, but indeed for all *compact* groups. In doing so, we obtain not only the classical Fourier series, which is simply a decomposition on the compact circle group, but also analogous decompositions on all n-spheres. However, before we move to proving these exciting results, we will begin with a necessary discussion of the representation theory of compact groups.

2. Preliminaries on compact groups

To begin, we should answer the question of what a compact group actually is. As one might guess, in order to make sense of compactness on a group, we need to introduce a topology on the group. Moreover, for this topology to be at all useful, it would be smart to have the topology interact well with the group structure. These ideas motivate the following definition.

Definition 2.1. A **topological group** G is a group equipped with a topology τ such that

(1) The group product is continuous as a function $G \times G \to G$, with the product topology on $G \times G$;
(2) The inverse function $^{-1} : G \to G$ is continuous as a function on G.

If, in addition, G is compact and Hausdorff, then it is a **compact group**.

Remark 2.2. The Hausdorff condition is not universal, but we will include it here for simplicity.

Example 2.3. Any finite group equipped with the discrete topology is a compact group. ♢

Example 2.4. More interestingly, consider the group

$$\mathrm{U}(1) = \{e^{i\theta} \mid \theta \in [0, 2\pi)\} \subseteq \mathbb{C},$$

with the usual topology inherited from $\mathbb{C}$. Since complex multiplication and conjugation are continuous, this is a Hausdorff topological group. Furthermore, since the unit circle is a compact subset of $\mathbb{C}$, this is a compact group. ♢

Example 2.5. Consider the group

$$\mathrm{SU}(2) = \left\{ \begin{bmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{bmatrix} \;\middle|\; |\alpha|^2 + |\beta|^2 = 1 \right\} \subseteq \mathbb{C}^4.$$

Again, since matrix multiplication and inversion are rational functions on $\mathbb{C}^n$, this is a Hausdorff topological group. Moreover, writing $\alpha = x + iy$ and $\beta = z + iw$, we see that the restriction is

$$x^2 + y^2 + z^2 + w^2 = 1.$$

In particular, as a topological space, this group is homeomorphic to the 3-sphere S^3, which is certainly compact. ♢

The most important result about compact topological groups, for our purposes, is the existence of a so-called *Haar measure* μ. We give a brief statement of the result.

Theorem 2.6 (Haar measure on compact groups). *Let G be a compact group. Then, there exists a measure μ on (Borel) subsets $S \subseteq G$ such that*

(1) *μ is left translation invariant, i.e., for any $g \in G$, $\mu(gS) = \mu(S)$;*
(2) *μ is right translation invariant, i.e., for any $g \in G$, $\mu(Sg) = \mu(S)$;*
(3) *$\mu(G) = 1$.*

Remark 2.7. The original theorem actually applies to locally compact groups and gives some additional regularity properties of this measure.

We will not prove this theorem, as it is not really an exercise in representation theory. However, the interested reader can consult [vdB93, Sec. 1].

As with any measure, the Haar measure allows us to perform integration on a compact group. Moreover, this integration is compatible with the group structure, in the sense that

$$\int_S f(x)\,\mathrm{d}\mu(x) = \int_{g^{-1}S} f(gx)\,\mathrm{d}\mu(x) = \int_{Sg^{-1}} f(xg)\,\mathrm{d}\mu(x)\,.$$

As such, choosing $S = G$, we can use the Haar integral to perform an averaging trick similar to the one used with sums in the case of finite groups.

3. **Representation theory of compact groups**

3.1. *Continuous representations*. Having set up the preliminary background on compact groups, we can now move to the actual subject of their representations. As with finite groups, it is most convenient to work with complex vector spaces, so, unless otherwise mentioned, we will take any vector space to be over $\mathbb{C}$. Unlike in the finite case, however, we do impose a slight extra condition of continuity on representations of infinite groups.

Definition 3.1. Let G be a topological group. A **continuous representation** of G is a homomorphism $\rho : G \to \mathrm{GL}(V)$ for some topological Hausdorff vector space V, such that the map $(g, v) \mapsto \rho(g)v$ is continuous as a map $G \times V \to V$. If V is also finite, then we have a **finite continuous representation**.

A **subrepresentation** of V is a subspace W fixed by the action of G, so that $\rho|_W$ is also a representation. An **irreducible representation** V is a representation with no nontrivial subrepresentations (i.e., no subrepresentations except 0 and V itself).

Remark 3.2. The reason to consider ρ as a map $G \times V \to V$ instead when discussing continuity is so that we do not need to define a topology on $\mathrm{GL}(V)$.

Furthermore, though we will not prove this, the continuity of ρ in this sense is equivalent to the a priori weaker condition that, for any fixed $v \in V$, the map $g \mapsto \rho(g)v$ is continuous as a function $G \to V$ (see [Mor19, Sec. V.2]).

For the rest of this paper, we will only consider finite continuous representations over $\mathbb{C}$ unless otherwise mentioned. The advantage of doing so is that much of the theory in the finite case carries over completely analogously. For example, we have the following lemma.

LEMMA 3.3 (Schur's lemma, part 1). *Let G be a compact group, and V_1, V_2 two complex irreducible representations. Then the space of all homomorphisms from V_1 to V_2 commuting with the actions of G is*

$$\mathrm{Hom}_G(V_1, V_2) = \begin{cases} 0 & V_1 \ncong V_2 \\ \mathbb{C} & V_1 \cong V_2. \end{cases}$$

Proof. Let $\rho : V_1 \to V_2$ be a homomorphism commuting with G. Then, as can easily be checked, $\ker \rho$ and $\rho(V_1)$ are subrepresentations of V_1 and V_2, respectively. Since V_1 and V_2 are irreducible, either $\rho = 0$ or $V_1 \cong V_2$. In the latter case, we can then consider an eigenvalue λ of ρ; since $\rho - \lambda$ has nontrivial kernel, it must be the 0 map, showing that $\rho = \lambda$. □

COROLLARY 3.4. *The irreducible representations of compact abelian groups are all one-dimensional.*

Proof. Let V be an irreducible representation of an abelian group, and let $\rho(g)$ be the representation of some element g. Since the group is abelian, $\rho(g)$ commutes with the action of any other group element so—taking $V_1 = V_2 = V$ in Schur's lemma—we conclude that $\rho(g)$ is a scalar. Since this is true for any g, for V to be irreducible, it must be one-dimensional. □

Example 3.5. We can already determine all the irreducible representations of U(1). By Schur's lemma, we know these are all one-dimensional. Parametrizing U(1) as $\exp(i\theta)$, any irreducible representation must therefore be a continuous function satisfying

$$\rho(x+y) = \rho(x)\rho(y) \quad \text{and} \quad \rho(0) = \rho(2\pi) = 1.$$

Since $\rho(g) \neq 0$, setting $f(x) = \log(\rho(x))$ gives

$$f(x+y) = f(x) + f(y) \quad \text{and} \quad f(0) = f(2\pi) + 2\pi i n,\ n \in \mathbb{Z}.$$

Choosing $f(0) = 0$ for convenience, we see that the only continuous functions satisfying these conditions are

$$f_n(\theta) = in\theta.$$

Thus, all irreducible representations of U(1) have the form

$$\rho_n(x) = e^{in\theta}.$$

◊

3.2. *Unitary representations.* In the case of finite groups, defining a G-invariant inner product on our representations was ultimately a rather useful tool. Motivated by the technique of averaging there, we can do something similar for compact groups.

Proposition 3.6. *Let G be a compact group and (ρ, V) a finite representation. Then there exists an inner product on V such that $\rho(g)$ is unitary for all $g \in G$ (i.e., the inner product is G-invariant).*

Proof. Using the Haar measure, we can imitate the proof from the case of finite groups. Specifically, let $\langle \cdot, \cdot \rangle$ be any inner product on V, and define the new inner product $\langle \cdot, \cdot \rangle_G$ by

$$\langle v, w \rangle_G = \int_G \langle gv, gw \rangle \, \mathrm{d}g \,.$$

Note that this is indeed an inner product, as $\langle v, v \rangle_G$ is the integral of a continuous, nonnegative quantity which is only identically zero if $v = 0$. Furthermore, this inner product is G-invariant, as

$$\langle hv, hw \rangle_G = \int_G \langle ghv, ghw \rangle \, \mathrm{d}g = \int_{Gh} \langle gv, gw \rangle \, \mathrm{d}g = \langle v, w \rangle_G.$$

□

Corollary 3.7. *Let G be a compact group and V a finite representation. Then V is semisimple (i.e., decomposes as a sum of irreducibles).*

Proof. If V is irreducible, we are done. Thus, let $W \subseteq V$ be an irreducible subspace fixed by G, and consider $W^\perp$ as given by the invariant inner product. We wish to show that $W^\perp$ is fixed by G. But, for any $v \in W^\perp$ and $w \in W$, we know that $\langle gv, w\rangle = \langle v, g^{-1}w\rangle = 0$, since W is fixed by G. Thus, gv is orthogonal to everything in W, so it is in $W^\perp$, showing that $W^\perp$ is also fixed by G. We can thus write $V = W \oplus W^\perp$ and induct on $W^\perp$ to get a decomposition into irreducible subspaces. □

Corollary 3.8 (Schur's lemma, part 2). *Let G be a compact group and V a finite representation such that $\mathrm{End}_G(V) = \mathbb{C}$. Then V is irreducible.*

Proof. Suppose that V were reducible, so that $V = V_1 \oplus V_2$ with V_1 and V_2 nontrivial. Let $P : V \to V_2$ be the orthogonal projection map onto V_2 given the unitary structure of the proposition. Then $P \in \mathrm{End}_G(V)$, so either $P = 0$ or $V \cong V_2$. But $V_2 \neq 0$, so P cannot be 0 and $V \cong V_2$, a contradiction. □

From now on, we will also assume any representation is unitary and denote its inner product as simply $\langle \cdot, \cdot \rangle$.

3.3. *Matrix coefficients and Schur orthogonality.* In our ultimate discussion of the Peter–Weyl theorem, it will be useful to have a more concrete understanding of the endomorphisms of the representations of G. To that end, it would be useful to consider matrix representations of these endomorphisms. However, rather than having to choose a basis for our representations, it is convenient to use the slightly more abstract notion of matrix coefficients.

Definition 3.9. Let G be a compact group and (ρ, V) a finite representation. A **matrix coefficient** is any function $m^\rho_{v,w} : G \to \mathbb{C}$ of the form

$$m^\rho_{v,w}(g) = m_{v,w}(g) = \langle \rho(g)v, w\rangle \quad \text{with} \quad v, w \in V.$$

The span of all matrix coefficients will be denoted $C(G)_\rho$. If a specific basis v_i is implied, these may also just be written as m_{ij}.

Note that this naming makes the most sense if we choose v, w to be from an orthonormal basis, in which case the individual matrix coefficients are just those of the matrix representation of g. However, more generally, the matrix coefficients so defined will always be the elements of the matrix representation of g with respect to some basis. The converse, namely that the elements of any matrix representation are actually matrix coefficients, also holds by linearity, so this definition really is not introducing anything new; it is perhaps just a bit easier to work with.

For finite groups, we had a very strong result on the orthogonality of these matrix coefficients. As one might expect by now, this result raises practically unchanged to the compact case.

THEOREM 3.10 (Orthogonality of matrix coefficients). *Let G be a compact group and (ρ_1, V) and (ρ_2, W) two irreducible finite representations. Let $v_1, v_2 \in V$ and $w_1, w_2 \in W$. Then we have*

$$\int_G m_{v_1,v_2}(g)\overline{m_{w_1,w_2}(g)}\,\mathrm{d}g = \begin{cases} \frac{1}{\dim V}\langle v_1, w_1\rangle\overline{\langle v_2, w_2\rangle} & \rho_1 \cong \rho_2 \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. For any $v \in V$, $w \in W$, consider the operators

$$L_{v,w}(x) = \langle x, v\rangle w \quad \text{and} \quad T_{v,w} = \int_G gL_{v,w}g^{-1}\,\mathrm{d}g\,.$$

Both of these are elements of $\mathrm{Hom}(V, W)$. Furthermore, note that $T_{v,w}$ commutes with the action of G, as

$$T_{v,w}g = \int_G hL_{v,w}(h^{-1}g)\,\mathrm{d}h = \int_G (gh)L_{v,w}h^{-1}\,\mathrm{d}h = gT_{v,w}.$$

As such, Schur's lemma tells us $T_{v,w}$ is a scalar if and only if $\rho_1 \cong \rho_2$ and is 0 otherwise. To determine this scalar, we can take the trace:

$$\mathrm{Tr}\,T_{v,w}(g) = \int_G \mathrm{Tr}\,hL_{v,w}h^{-1}\,\mathrm{d}h = \int_G \mathrm{Tr}\,L_{v,w}\,\mathrm{d}h = \mathrm{Tr}\,L_{v,w}.$$

The trace of $L_{v,w}$ is most easily evaluated by using an orthonormal basis e_i of V, yielding

$$\mathrm{Tr}\,L_{v,w} = \sum_{i=1}^{\dim V}\langle L_{v,w}(e_i), e_i\rangle = \sum_{i=1}^{\dim V}\langle e_i, v\rangle\langle w, e_i\rangle = \langle w, v\rangle.$$

Thus, we have $T_{v,w}(g) = \frac{1}{\dim V}\langle w, v\rangle$. Finally, we can answer our original question by noting that

$$\begin{aligned}
\int_G m_{v_1,v_2}(g)\overline{m_{w_1,w_2}(g)}\,\mathrm{d}g &= \int_G \langle gv_1, v_2\rangle\overline{\langle gw_1, w_2\rangle}\,\mathrm{d}g \\
&= \int_G \langle gv_1, v_2\rangle\langle g^{-1}w_2, w_1\rangle\,\mathrm{d}g \\
&= \int_G \langle g\,\langle g^{-1}w_2, w_1\rangle\,v_1, v_2\rangle\,\mathrm{d}g \\
&= \left\langle \int_G g\langle g^{-1}w_2, w_1\rangle v_1\,\mathrm{d}g\,, v_2\right\rangle \\
&= \langle L_{w_1,v_1}w_2, v_2\rangle.
\end{aligned}$$

Using our classification for L_{w_1,v_1}, we can ultimately conclude

$$\int_G m_{v_1,v_2}(g)\overline{m_{w_1,w_2}(g)}\,\mathrm{d}g = \begin{cases} \frac{1}{\dim V}\langle v_1, w_1\rangle\overline{\langle v_2, w_2\rangle} & \rho_1 \cong \rho_2 \\ 0 & \text{otherwise.} \end{cases}$$

□

3.4. *Characters*. Having just proved Schur orthogonality, it is worth taking a brief digression to discuss characters.

Definition 3.11. Let G be a compact group and (ρ, V) a representation. The **character** χ of ρ is defined by

$$\chi(g) = \operatorname{Tr}\rho(g).$$

If ρ is an irreducible representation, χ is called an **irreducible character**.

Characters function largely the same as for finite groups. Indeed, the character of the sum of two representations is simply the sum of characters, and therefore any character breaks down into a sum of irreducible characters. We reproduce the following two familiar results.

Corollary 3.12 (Character orthogonality). *Let G be a compact group and let V, W be two irreducible representations with characters χ_V, χ_W. Then*

$$\int_G \chi_V(g)\overline{\chi_W(g)}\,\mathrm{d}g = \begin{cases} 1 & V \cong W \\ 0 & V \ncong W. \end{cases}$$

Proof. Choose orthonormal bases v_i and w_j of V and W. Then, we have

$$\chi_V(g) = \sum_{i=1}^{\dim V}\langle gv_i, v_i\rangle = \sum_{i=1}^{\dim V} m_{v_i,v_i}(g)$$

and similarly for χ_W. Thus,

$$\int_G \chi_V(g)\overline{\chi_W(g)}\,\mathrm{d}g = \sum_{i=1}^{\dim V}\sum_{j=1}^{\dim W}\int_G m_{v_i,v_i}(g)\overline{m_{w_j,w_j}(g)}\,\mathrm{d}g\,.$$

If $V \ncong W$, we already know this is 0 by Theorem 3.10. On the other hand, if $V \cong W$, we can take $v_i = w_i$, yielding

$$\sum_{i=1}^{\dim V}\sum_{j=1}^{\dim V}\int_G m_{v_i,v_i}(g)\overline{m_{v_j,v_j}(g)}\,\mathrm{d}g = \frac{1}{\dim V}\sum_{i=1}^{\dim V}\sum_{j=1}^{\dim V}|\langle v_i, v_j\rangle|^2 = 1.$$

□

Corollary 3.13. *Let G be a compact group and χ a character of a finite representation. Write χ as a sum of irreducible characters $\chi = \sum_{i=1}^k n_i\chi_i$.*

Then

$$\int_G |\chi(g)|^2 \, \mathrm{d}g = \sum_{i=1}^{k} n_i^2.$$

Namely, χ is irreducible if and only if the integral is 1.

Example 3.14. If we parametrize the circle group in terms of an angle $\theta \in [0, 2\pi)$, one can check that the Haar measure is given by

$$\mathrm{d}g = \frac{\mathrm{d}\theta}{2\pi}.$$

Furthermore, since the irreducible representations of U(1) are one-dimensional (see the example), we already have the characters

$$\chi_n(\theta) = \rho_n(\theta) = \exp(in\theta).$$

Thus, by a needlessly complicated proof, we have that

$$\int_{\mathrm{U}(1)} \chi_n(g)\overline{\chi_m(g)} \, \mathrm{d}g = \frac{1}{2\pi} \int_0^{2\pi} e^{in\theta} e^{-im\theta} \, \mathrm{d}\theta = \delta_{nm}.$$

More interestingly, we also have a finite, integral version of Parseval's identity. Indeed, if f is an arbitrary finite character

$$f = \sum_{i=-N}^{N} n_i \chi_i,$$

then Corollary 3.13 tells us that

$$\frac{1}{2\pi} \int_0^{2\pi} |f(\theta)|^2 \, \mathrm{d}\theta = \sum_{i=-N}^{N} n_i^2.$$

$\lozenge$

4. $L^2(G)$ & the Peter–Weyl theorem

Having digressed enough on the subject of representations, it would be good to remind ourselves of the original goal of describing functions on G. In the case of finite groups, this could be achieved by considering the group algebra

$$\mathbb{C}[G] = \bigoplus_{g \in G} \mathbb{C}g,$$

which has a multiplication linearly extending that of G. This could be identified as the algebra of all functions $f : G \to \mathbb{C}$ with the convolution product by letting $f(g)$ be the coefficient of g in f.

Unfortunately, directly attempting to use the group algebra in the case of compact groups is a bit too general. Indeed, the space $\mathbb{C}[G] \sim \mathbb{C}^G$ contains plenty of unwieldy and uninteresting functions. More importantly, it also contains plenty of nonintegrable functions, which prevents us from using the tools we have developed. The easiest way to fix this issue is just to get rid of these problematic functions.

4.1. *The representation* $L^2(G)$.

Definition 4.1. Let G be a compact group. Let $L^2(G)$ be the Banach space of complex square-integrable functions, i.e., those functions $f : G \to \mathbb{C}$ such that

$$\int_G |f|^2 \, \mathrm{d}g \ \text{ exists and is } < \infty.$$

Then, G acts on $L^2(G)$ as

$$(gf)(x) = f(g^{-1}x).$$

Remark 4.2. Technically speaking, the space $L^2(G)$ is actually a quotient of the above definition by the equivalence of almost-everywhere equality, but we will ignore this complication as it is not essential. For more on L^p spaces, see [Axl19, Chap. 7–8].

Example 4.3. The matrix coefficients $C(G)_\rho$ are all continuous functions, and hence their squares are integrable over the compact set G. Thus, we have

$$C(G)_\rho \subseteq L^2(G) \quad \text{for all} \quad \rho.$$

◊

Example 4.4. For a finite group, we know that $L^2(G) \cong \mathbb{C}[G]$, since the integral is just summing over each group element. Thus, an element f looks like

$$f = \sum_{h \in G} f(h)h.$$

Therefore,

$$gf = \sum_{h \in G} (gf)(h)h = \sum_{h \in G} f(g^{-1}h)h = \sum_{h \in G} f(h)gh.$$

In other words, $L^2(G)$ is just the regular representation of G, which should not be too surprising given the analogy with the group algebra. ◊

Example 4.5. For the group $\mathrm{U}(1)$, $L^2(\mathrm{U}(1))$ can be identified as all square-integrable functions on the circle, since $\mathrm{U}(1) \cong S^1$, together with the translation action $f(x) \mapsto f(x - \theta)$. ◊

The reason to choose square integrability, rather than just normal integrability, is that it will allow us to promote $L^2(G)$ from just a Banach space to a

Hilbert space, i.e., a space with an inner product. Indeed, we can define

$$\langle f, g\rangle = \int_G f\overline{g}\,\mathrm{d}x\,,$$

which is guaranteed to exist by the complex polarization identities. However, before continuing with its representation theory, it is worth digressing to discuss the product structure of $L^2(G)$.

4.2. *Convolutions*. Without being too rigorous, we can think about an element $f \in L^2(G)$ as a "weighted integral" of elements of G

$$f = \int_G f(g) g\,\mathrm{d}g\,.$$

From this, we can calculate the product of two elements as

$$f_1 * f_2 = \int_G \int_G f_1(h) f_2(g) hg\,\mathrm{d}h\,\mathrm{d}g = \int_G \left(\int_G f_1(h) f_2(h^{-1}g)\,\mathrm{d}h\right) g\,\mathrm{d}g\,.$$

Looking at the coefficient of g in this expression thus motivates the following definition for the convolution.

Definition 4.6. Let G be a compact group. Then, for any $f_1, f_2 \in L^2(G)$, we define the **convolution**

$$(f_1 * f_2)(g) = \int_G f_1(h) f_2(h^{-1}g)\,\mathrm{d}h\,.$$

Remark 4.7. One should prove that this convolution does actually obey the regular associativity and distributivity laws of a product. This is a good exercise in analysis.

Also, note that if G is not discrete, $*$ does not technically have an identity element. However, as we will discuss, one can still approximate an identity element using $L^2(G)$ functions.

Note that the convolution is not, in general, abelian (which should not be a surprise, considering G need not be). As such, there are two natural operations we can extract from the convolution by fixing one of the two factors. Specifically, we will write

$$L_h(f) = h * f \quad \text{and} \quad R_h(f) = f * h \quad \text{for} \quad h, f \in L^2(G).$$

These operations are, in general, very well behaved. Specifically, we have the following collection of technical results from functional analysis, which are only partially reproduced as they are not the focus of this article.

PROPOSITION 4.8. *Let $h \in L^2(G)$, and define $\tilde{h}(x) = \overline{h(x^{-1})}$. Then we have that*

(1) *L_h and R_h are continuous compact operators*;

(2) $(L_h)^* = L_{\tilde{h}}$ *and* $(R_h)^* = R_{\tilde{h}}$. *In particular, if* $h = \tilde{h}$, *then* L_h *and* R_h *are self-adjoint.*

Proof.

(1) The continuity of both L_h and R_h follows easily enough by applying the Cauchy–Schwarz inequality to show that

$$\|L_h(f)(g)\| = \left\| \int_G h(x) f(x^{-1}g)\,\mathrm{d}x \right\| \leq \|h\| \|f\|,$$

and similarly for R_h. Compactness, on the other hand, is more technical, but can be done by noting that the convolution is an integral operator with a compactly supported kernel; the interested reader can find the full details in [Mor19, Chap V.4] or [vdB93, Sec. 8].

(2) We prove this for L_h, as the proof for R_h is nearly identical.

$$\begin{aligned}
\langle L_h f_1, f_2 \rangle &= \int_G h * f_1 \overline{f_2}\,\mathrm{d}x \\
&= \int_G \int_G h(y) f_1(y^{-1}x) \overline{f_2(x)}\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{G \times G} h(y^{-1}) f_1(x) \overline{f_2(y^{-1}x)}\,\mathrm{d}y\,\mathrm{d}x \quad (x \to yx \quad \text{and} \quad y \to y^{-1}) \\
&= \int_G f_1(x) \left[\int_G \overline{h(y^{-1})} f_2(y^{-1}x)\,\mathrm{d}y \right]^* \mathrm{d}x \\
&= \langle f_1, L_{\tilde{h}} f_2 \rangle.
\end{aligned}$$

Thus, if $h = \tilde{h}$, then $L_h = L_{\tilde{h}}$ is equal to its adjoint.

□

Despite being very nicely behaved, however, the convolution does have one major weakness: its lack of an identity element. This is rather annoying, as it means that $L^2(G)$ with $*$ as the product is a non-unital ring. However, as alluded to, we can still approximate an identity element as well as we need to.

Lemma 4.9. *Given any* $f \in L^2(G)$, *there is a sequence of functions* h_n *such that*

(1) $h_n = \tilde{h}_n$;
(2) $\|h_n\| = 1$;
(3) $f * h_n \to f$ *as* $n \to \infty$.

Proof. Denote by r_x right multiplication by x, i.e., $r_x f(y) = f(yx)$.

Now, let $\epsilon > 0$, and choose a neighborhood of the identity $U \subseteq G$ such that $U = U^{-1}$ and $\|r_x f - f\| < \epsilon$ for all $x \in U$, which is possible by continuity of the group multiplication. Define $h_\epsilon = \frac{1}{\mathrm{Vol}(U)} 1_U$. Then $h_\epsilon = \tilde{h}_\epsilon$ and $\|h_\epsilon\| = 1$

by definition. Furthermore, we have

$$\begin{aligned} f * h_\epsilon(g) - f(g) &= \frac{1}{\mathrm{Vol}(U)} \int_G f(x) 1_U(x^{-1}g)\,\mathrm{d}x - f(g) \\ &= \frac{1}{\mathrm{Vol}(U)} \int_G f(gx) 1_U(x^{-1})\,\mathrm{d}x - \frac{1}{\mathrm{Vol}(U)} \int_U f(g)\,\mathrm{d}x \\ &= \frac{1}{\mathrm{Vol}(U)} \int_U (r_x f)(g) - f(g)\,\mathrm{d}x\,. \end{aligned}$$

Thus, we can conclude that

$$\begin{aligned} \|f * h_\epsilon - f\|^2 &= \frac{1}{\mathrm{Vol}(U)^2} \int_{U\times U} ((r_x)f - f)\overline{((r_y)f - f)}\,\mathrm{d}x\,\mathrm{d}y \\ &\leq \frac{1}{\mathrm{Vol}(U)^2} \int_{U\times U} \|(r_x)f - f\|\overline{\|(r_x)f - f\|}\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{1}{\mathrm{Vol}(U)^2} \int_{U\times U} \epsilon^2\,\mathrm{d}x\,\mathrm{d}y \\ &= \epsilon^2. \end{aligned}$$

In particular, if we take the sequence $h_n := h_{2^{-n}}$, we get an approximation to the identity. □

4.3. *The Peter–Weyl theorem.* We can finally come to our first major result, the titular Peter–Weyl theorem. This, as mentioned earlier, is really just a generalization of the Artin–Wedderburn theorem to compact groups, giving us a decomposition of $L^2(G)$, the equivalent to $\mathbb{C}[G]$, into simpler spaces given by the irreducible representations of G.

We will still need one more theorem before proving Peter–Weyl: the spectral theorem for compact self-adjoint operators. However, we will just be stating this result, as it is purely a result from functional analysis.

Theorem 4.10 (Spectral theorem). *Let $T : V \to W$ be a compact self-adjoint operator between Hilbert spaces. Then V decomposes as an orthogonal direct sum*

$$V = \mathrm{Ker}(T) \bigoplus_{\lambda\in\Lambda} E_\lambda,$$

where $\Lambda \in \mathbb{R}^*$ is a discrete set of eigenvalues, and the E_λ are orthogonal, finite-dimensional eigenspaces.

Proof. See [Mor19, Chap. V.6] or [Axl19, Chap. 10D]. □

Having finally gone through all the preliminaries, we present the Peter–Weyl theorem.

THEOREM 4.11 (Peter & Weyl, 1927). *Let G be a compact group, and $\widehat{G}$ the set of finite irreducible representations of G. Then*

$$L^2(G) \cong \widehat{\bigoplus_{\rho \in \widehat{G}}} C(G)_\rho,$$

where $\widehat{\oplus}$ denotes the closure of the direct sum.

Proof. We will denote

$$\mathcal{R}(G) = \widehat{\bigoplus_{\rho \in \widehat{G}}} C(G)_\rho$$

for convenience. The proof will consist of two steps: showing that every finite subrepresentation of $L^2(G)$ occurs in $\mathcal{R}(G)$, and showing that this implies that the complement of $\mathcal{R}(G)$ is trivial.

For the first step, consider some arbitrary finite representation V of G. Without loss of generality, we may take V to be irreducible, since any finite representation is semisimple by Corollary 3.7. Our strategy will be to show that the image of any inclusion map $u : V \to L^2(G)$ commuting with the action of G is in fact contained in $\mathcal{R}(G)$, i.e., is in the span of all matrix coefficients. To do so, take some $v \in V$ and let $f \in L^2(G)$. We then have

$$\begin{aligned}
(u(v) * \tilde{f})(g) &= \int_G u(v)(h)\overline{f(g^{-1}h)}\,\mathrm{d}h \\
&= \int_G u(v)(gh)\overline{f(h)}\,\mathrm{d}h \\
&= \langle u(v) \circ g, f\rangle \\
&= \langle u(\rho(g^{-1})v), f\rangle \\
&= \langle \rho(g^{-1})v, u^*(f)\rangle.
\end{aligned}$$

This is a matrix coefficient for the dual representation of ρ, so it is in $\mathcal{R}(G)$. Now, if we take a sequence of $\tilde{f}_n$ approximating the identity, we can then conclude that

$$u(v) * \tilde{f}_n \to u(v) \in \mathcal{R}(G).$$

We are now ready to complete our proof of the theorem. To that end, consider an element $f \in \mathcal{R}(G)^\perp$. If we now consider any element $h \in L^2(G)$ such that $h = \tilde{h}$, we know that R_h is a self-adjoint compact operator. As such, $L^2(G)$ decomposes as

$$L^2(G) = \operatorname{Ker}(R_h) \bigoplus_i E_{\lambda_i},$$

where the E_{λ_i} are finite-dimensional. As such, they are all in $\mathcal{R}(G)$, so f is orthogonal to them. In particular, $f \in \operatorname{Ker}(R_h)$, i.e., $f * h = 0$. Again, taking now a sequence h_n that approximates the identity, we conclude that $f = 0$, completing the proof. □

5. Applications to S^1 and S^3

After all of that work, we are finally ready to discuss some concrete applications of all of this theory to Fourier-type decompositions on n-spheres. We will only handle the cases S^1, S^2, and S^3 in this article, as the general case needs more sophisticated tools. The cases of S^1 and S^3 are easiest to handle thanks to the fact that these two spheres actually have group structures; namely, we have $S^1 \cong U(1)$ and $S^3 \cong \mathrm{SU}(2)$ as discussed previously. As such, we will discuss them first.

5.1. U(1) *and* S^1. The case of S^1, though not particularly revolutionary in its conclusion, is still a wonderful and simple example of the Wedderburn-type decomposition we are trying to do. Moreover, it provides the framework with which we can approach more general cases.

THEOREM 5.1 (Fourier, 1807). *The space* $L^2(S^1)$ *decomposes as*

$$L^2(S^1) \cong \widehat{\bigoplus_{n\in\mathbb{Z}}} C(\mathrm{U}(1))_n \cong \widehat{\bigoplus_{n\in\mathbb{Z}}} \mathbb{C}e^{in\theta}.$$

More concretely, a function $f \in L^2(S^1)$ *can be written as*

$$f(\theta) = \sum_{n\in\mathbb{Z}} \widehat{f}(n)e^{in\theta} \quad \text{with} \quad \widehat{f}(n) = \frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-in\theta}\,\mathrm{d}\theta\,.$$

Proof. Note that $\mathrm{U}(1) \cong S^1$. Furthermore, by our classification of the irreducible representations, the span of matrix coefficients is clearly just

$$C(\mathrm{U}(1))_n \cong \mathbb{C}\exp(in\theta).$$

Thus, applying Theorem 4.11 gives us the first statement.

For the more concrete realization, note that we already know f decomposes as a sum:

$$f = \sum_{n\in\mathbb{Z}} a_n e^{in\theta}.$$

We can then use Theorem 3.10 to isolate what a_n is. Specifically, taking the inner product with the matrix coefficient $\overline{m_n} = e^{-in\theta} = m_{-n}$, and recalling that $\mathrm{d}g = \mathrm{d}\theta\,/(2\pi)$, gives us that

$$\frac{1}{2\pi}\int_0^{2\pi} f(\theta)e^{-in\theta}\,\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi}\sum_{k\in\mathbb{Z}} a_k m_k \overline{m_n}\,\mathrm{d}\theta = \sum_{k\in Z} a_k\delta_{kn} = a_n.$$

□

5.2. *Representation theory of* SU(2). As we noted previously, SU(2) can be viewed as the manifold S^3. Thus, to get a Fourier theory on S^3, it would be sufficient to determine the matrix coefficients of representations of SU(2). Before we can do that, however, we need to actually determine the irreducible representations themselves.

In order to find these irreducible representations, note that there is a natural action of SU(2) on $\mathbb{C}^2$ given by

$$g\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{bmatrix}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \alpha z_1 + \beta z_2 \\ -\overline{\beta} z_1 + \overline{\alpha} z_2 \end{bmatrix}.$$

A slight reframing of this involves considering $\mathbf{z} = (z_1, z_2)$ as variables for a 2-variable polynomial $p_1(\mathbf{z}) = az_1 + bz_2$. With this reframing, we get a representation

$$(gp_1)(\mathbf{z}) = p_1(g^{-1}\mathbf{z}),$$

the inverse being necessary to respect associativity. This can be generalized by considering higher-degree polynomials. Namely, if we let P_n be the space of all $\leq n$ degree complex polynomials in 2 variables, we get a representation

$$(gp_n)(\mathbf{z}) = p_n(g^{-1}\mathbf{z}) \quad \text{for} \quad p_n \in P_n.$$

This representation, unfortunately, is not irreducible. Indeed, consider the subspace $\mathcal{P}_n$ of all homogeneous degree n polynomials, i.e., the polynomials such that

$$p_n(\lambda\mathbf{z}) = \lambda^n p_n(\mathbf{z}).$$

Then this subspace is invariant under the SU(2) action, as

$$(gp_n)(\lambda\mathbf{z}) = p_n(g^{-1}\lambda\mathbf{z}) = p_n(\lambda g^{-1}\mathbf{z}) = \lambda^n (gp_n)(\mathbf{z}).$$

The natural question to ask is whether this new representation is irreducible. The answer, as we will prove, is yes.

PROPOSITION 5.2. *The* SU(2) *representation on* $\mathcal{P}_n$ *is irreducible for every* $n \geq 0$. *In particular, there is a representation* ρ_n *of dimension* $n+1$ *for every* $n \geq 0$.

Proof. Our proof will attempt to use Corollary 3.8 by showing that any endomorphism A of $\mathcal{P}_n$ commuting with the action of SU(2) is a scalar.

To start our proof, note that the polynomials $p_k = z_1^k z_2^{n-k}$ form a basis of $\mathcal{P}_n$ for $0 \leq k \leq n$. Now, consider the special elements

$$u_\theta = \begin{bmatrix} e^{-i\theta} & 0 \\ 0 & e^{i\theta} \end{bmatrix} \in U(1) \subseteq \mathrm{SU}(2).$$

These elements are of note, as

$$u_\theta p_k = (e^{i\theta} z_1)^k (e^{-i\theta} z_2)^{n-k} = e^{i\theta(2k-n)} p_k.$$

Namely, the p_k are eigenvectors of u_θ with respective eigenvalues $e^{i\theta(2k-n)}$. As such, in the basis of the p_k, we have

$$\rho_n(u_\theta) = \operatorname{diag}\big(e^{-i\theta n}, e^{-i\theta(n-2)}, \dots, e^{i\theta n}\big).$$

By choosing θ small enough, these eigenvalues are all distinct, so the p_k also generate all the eigenspaces of u_θ. Since A is assumed to commute with $\mathrm{SU}(2)$, it must map each of these eigenspaces to itself. Thus,

$$Ap_k = \lambda_k p_k \quad \text{for} \quad 0 \leq k \leq n.$$

We now want to show that $\lambda_k = \lambda_0$ for all k. To do so, consider the new elements

$$r_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \in U(1) \subseteq \mathrm{SU}(2).$$

We can then look at the action of r_θ and A on $p_0 = z_1^n$. Specifically, we have

$$\begin{aligned} Ar_\theta p_0 &= A(\cos\theta z_1 + \sin\theta z_2)^n \\ &= A\sum_{k=0}^{n} \binom{n}{k} (\cos\theta)^k (\sin\theta)^{n-k} z_1^k z_2^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (\cos\theta)^k (\sin\theta)^{n-k} Ap_k \\ &= \sum_{k=0}^{n} \lambda_k \binom{n}{k} (\cos\theta)^k (\sin\theta)^{n-k} p_k. \end{aligned}$$

On the other hand, since $Ar_\theta = r_\theta A$, we also get

$$r_\theta A p_0 = \lambda_0 r\theta p_0 = \sum_{k=0}^{n} \lambda_0 \binom{n}{k} (\cos\theta)^k (\sin\theta)^{n-k} p_k.$$

Comparing these two expressions, we can indeed conclude that $\lambda_0 = \lambda_k$ for all $0 \leq k \leq n$. Thus, $A = \lambda_0 I$ is a scalar, and we conclude that ρ_n is irreducible. □

Corollary 5.3. *For every* $n \geq 0$, $\mathrm{SU}(2)$ *has an irreducible character* χ_n *given by*

$$\chi_n(u_\theta) = \sum_{k=0}^{n} e^{i\theta(2k-n)}.$$

Proof. Note that, by the spectral theorem for finite vector spaces, any element of $\mathrm{SU}(2)$ is conjugate to a diagonal matrix of the form u_θ defined previously. Thus, it is sufficient to define the characters on this subspace.

Now, consider again the basis $p_k = z_1^k z_2^{n-k}$ of $\mathcal{P}_n$. We already saw that

$$\rho_n(u_\theta)p_k = e^{i\theta(2k-n)}p_k,$$

from which we can conclude that the trace of $\rho_n(u_\theta)$ is

$$\chi_n(u_\theta) = \sum_{k=0}^{n} e^{i\theta(2k-n)}.$$

□

The previous corollary tells us that the span of the characters of SU(2) is dense in the even periodic functions. Specifically, denoting $\chi_n(\theta) := \chi_n(u_\theta)$, we can express $\cos(n\theta)$ for $n \in \mathbb{Z}$ as

$$1 = \chi_0 \quad \text{and} \quad \cos(\theta) = \frac{1}{2}\chi_1(\theta) \quad \text{and} \quad \cos(n\theta) = \frac{1}{2}\Big(\chi_n(\theta) - \chi_{n-2}(\theta)\Big),$$

which are dense in the even periodic L^2 functions by Theorem 5.1. In fact, this observation allows us to conclude that the ρ_n we defined give all of the irreducible representations of SU(2).

PROPOSITION 5.4. *The ρ_n enumerate all irreducible representations of* SU(2)*.*

Proof. Let ρ be a representation with character χ. Note that χ is completely described by its restriction to the u_θ, since characters are invariant under conjugation and any SU(2) matrix can be diagonalized. Furthermore, since u_θ is conjugate to $u_{-\theta}$, we must have $\chi(-\theta) = \chi(\theta)$. In other words, χ is just an even function on the unit circle. Thus, by Theorem 5.1, χ decomposes as a sum of $\cos(n\theta)$ terms. However, we just saw that $\cos(n\theta)$ can be expressed in terms of the χ_n. Thus, χ can be expressed as a sum of the χ_n. In particular, χ contains at least one of the χ_n, so χ is either one of them or is reducible. □

5.3. SU(2) *and* S^3. Now that we have a concrete realization and understanding of all of the irreducible representations of SU(2), an application of Theorem 4.11 achieves our stated goal.

PROPOSITION 5.5. *The space* $L^2(S^3) \cong L^2(\mathrm{SU}(2))$ *decomposes as*

$$L^2(S^3) \cong \widehat{\bigoplus_{n\geq 0}} C(\mathrm{SU}(2))_n.$$

However, this is not really a satisfying result. Indeed, while this is certainly a valid decomposition of the functions on S^3 into smaller algebras, it is not clear at all what the spaces $C(\mathrm{SU}(2))_n$ look like, or how they even relate to functions on S^3. Thus, to get a better understanding, we need to put a bit more effort into studying the matrix coefficients of SU(2).

Recall that the link between SU(2) and S^3 we had was based on mapping

$$\begin{bmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{bmatrix} \mapsto (x_1, y_1, x_2, y_2) \in S^3 \quad \text{where} \quad \alpha = x_1 + iy_1,\ \beta = x_2 + iy_2.$$

As such, it would make sense to consider the matrix coefficients as functions of α and β. For example, we can consider the action

$$\begin{bmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{bmatrix} p_k = \Big(\alpha z_1 + \beta z_2\Big)^k \Big(-\overline{\beta} z_1 + \overline{\alpha} z_2\Big)^{n-k} =: F_k(\alpha, \beta)$$

as a polynomial in $\alpha, \beta, \overline{\alpha}$, and $\overline{\beta}$. Note that this is particularly convenient, as these 4 variables are linearly related to the variables x_1, y_1, x_2, and y_2.

There are now three insights that allow us to give a more concrete picture of $C(\mathrm{SU}(2))_n$. The first is that, since the p_k are a basis of $\mathcal{P}^n$, a basis for $C(\mathrm{SU}(2))_n$ is given by

$$F_k^m(\alpha, \beta) := \; z_1^{n-m} z_2^m \text{ coefficient of } \Big(\alpha z_1 + \beta z_2\Big)^k \Big(-\overline{\beta} z_1 + \overline{\alpha} z_2\Big)^{n-k},$$

where $0 \leq m, k \leq n$.

Example 5.6. Consider the space $\mathcal{P}_2$. We have

$$\begin{aligned} F_0 &= \Big(\alpha z_1 + \beta z_2\Big)^2 = \alpha^2 z_1^2 + 2\alpha\beta z_1 z_2 + \beta^2 z_2^2 \\ F_1 &= \Big(\alpha z_1 + \beta z_2\Big)\Big(-\overline{\beta} z_1 + \overline{\alpha} z_2\Big) = -\alpha\overline{\beta} z_1^2 + (\alpha\overline{\alpha} - \beta\overline{\beta}) z_1 z_2 + \overline{\alpha}\beta z_2^2 \\ F_2 &= \Big(-\overline{\beta} z_1 + \overline{\alpha} z_2\Big)^2 = \overline{\beta}^2 z_1^2 - 2\overline{\alpha}\overline{\beta} z_1 z_2 + \overline{\alpha}^2 z_2^2. \end{aligned}$$

Thus, the F_k^m, which form a basis for the space of matrix coefficients, are

F_k^m	0	1	2
0	α^2	$2\alpha\beta$	β^2
1	$-\alpha\overline{\beta}$	$\alpha\overline{\alpha} - \beta\overline{\beta}$	$\overline{\alpha}\beta$
2	$\overline{\beta}^2$	$-2\overline{\alpha}\overline{\beta}$	$\overline{\alpha}^2$

.

◊

The second observation is that F_k is still real-homogeneous of degree n, i.e.,

$$F_k(\lambda\alpha, \lambda\beta) = \Big(\lambda\alpha z_1 + \lambda\beta z_2\Big)^k \Big(-\overline{\lambda\beta} z_1 + \overline{\lambda\alpha} z_2\Big)^{n-k} = \lambda^n F_k(\alpha, \beta)$$

for $\lambda \in \mathbb{R}$. Thus, we can also interpret the matrix coefficients as some subspace of the homogeneous polynomials of degree n in 4 real variables, if we choose to write α, β, and their conjugates in terms of the x_i and y_i.

Example 5.7. Continuing the previous example, we get these degree 2 homogeneous polynomials:

$$\begin{aligned} \alpha^2 &= x_1^2 + 2ix_1y_1 - y_1^2 & \overline{\alpha}^2 &= x_1^2 - 2ix_1y_1 - y_1^2 \\ \beta^2 &= x_2^2 + 2ix_2y_2 - y_2^2 & \overline{\beta}^2 &= x_2^2 - 2ix_2y_2 - y_2^2 \\ 2\alpha\beta &= 2(x_1x_2 + y_1y_2) + 2i(x_1y_2 + x_2y_1) & -\alpha\overline{\beta} &= -x_1x_2 + y_1y_2 + i(x_1y_2 - x_2y_1) \\ \overline{\alpha}\beta &= x_1x_2 - y_1y_2 + i(x_1y_2 - x_2y_1) & -2\overline{\alpha}\overline{\beta} &= -2(x_1x_2 + y_1y_2) + 2i(x_1y_2 + x_2y_1) \end{aligned}$$

$$\alpha\overline{\alpha} - \beta\overline{\beta} = x_1^2 + y_1^2 - x_2^2 - y_2^2.$$

◊

The final observation is that, as a 4-variable real function, $F_k(x_1, y_1, x_2, y_2)$ is actually harmonic. If we write

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y_2^2} = 4\frac{\partial^2}{\partial\alpha\partial\overline{\alpha}} + 4\frac{\partial^2}{\partial\beta\partial\overline{\beta}},$$

the symmetry of the two terms defining F_k makes it easy to check that it is harmonic. Moreover, since F_k is harmonic, each of the F_k^m is too. We have therefore established that the matrix coefficients $C(\mathrm{SU}(2))_n$ are actually homogeneous harmonic polynomials of degree n on $\mathbb{R}^4$. These polynomials are so important, in fact, that it is worth giving them a special symbol.

Definition 5.8. Let

$$\mathfrak{H}_n^m = \{p \in \mathcal{P}_n(\mathbb{R}^m) | \Delta p = 0\},$$

i.e., the space of all harmonic homogeneous polynomials of degree n on $\mathbb{R}^m$.

Example 5.9. The space $\mathfrak{H}_2^4$ is simple enough that one can manually enumerate the possibilities. Doing so shows that $\mathfrak{H}_2^4$ is 9-dimensional, with basis

$$x_1^2 - y_1^2,\ x_2^2 - y_2^2,\ x_1^2 - x_2^2,\ x_1y_1,\ x_1x_2,\ x_1y_2,\ y_1x_2,\ y_1y_2,\ x_2y_2.$$

◊

Curiously, the previous examples show that $\mathfrak{H}_2^4$ and $C(\mathrm{SU}(2))_2$ actually have the same dimension and are thus the same space. It turns out this is a general phenomenon: $C(\mathrm{SU}(2))_n$ is not only a subspace of $\mathfrak{H}_n^4$, but is in fact equal to it. Proving this is most easily done by noting both of these spaces have dimension $(n+1)^2$. For $C(\mathrm{SU}(2))_n$, this follows immediately from the fact that the $n+1$ elements p_k form a basis of $\mathcal{P}_n$. On the other hand, to see that $\mathfrak{H}_n^4$ has dimension $(n+1)^2$, consult Appendix A. In any case, putting everything together, we finally get the proper hyperspherical decomposition on S^3.

THEOREM 5.10 (S^3 hyperspherical decomposition). *The space* $L^2(S^3) \cong L^2(\mathrm{SU}(2))$ *decomposes as*

$$L^2(S^3) \cong \widehat{\bigoplus_{n\geq 0}} \mathfrak{H}_n^4|_{S^3},$$

where $|_{S^3}$ *denotes restriction to* $S^3 \subseteq \mathbb{R}^4$. *Furthermore, the coefficients in this decomposition can be calculated as*

$$F_{ij}^n = \langle f, m_{ij}\rangle_{\mathrm{SU}(2)} = \int_{S^3} f\, \overline{m_{ij}^n}\, \mathrm{d}\mu \quad \text{for} \quad 0 \leq i, j \leq n.$$

Remark 5.11. The invariant metric μ on SU(2) is, unfortunately, rather complicated, so we will not be writing it down explicitly.

6. Applications to S^2

Our final task is to tackle spherical decompositions on S^2. This is hindered by the fact that S^2 has no obvious group structure; in fact, it can be shown that there is no way to give S^2 a group structure compatible with its geometry (see [Lee18]). For this reason, we will need to change our approach slightly.

The most important insight is that, while S^2 is not a group itself, it is certainly acted upon very naturally by many groups. In particular, the group SO(3) of three-dimensional rotations has a natural action on the 2-sphere. This action is transitive, i.e., the orbit of every point is all of S^2 but is not faithful. Indeed, the stabilizer of any point is a subgroup of SO(3) isomorphic to SO(2). This is easy to see geometrically: any rotation that fixes a particular point on the surface of the sphere must be a rotation through that point, and so these collectively form a group of two-dimensional rotations. Thus, by the orbit-stabilizer theorem, we have an identification

$$S^2 \cong \mathrm{SO}(3)/\mathrm{SO}(2).$$

This identification now gives us a useful way to think about $L^2(S^2)$. Namely, consider a function $f \in L^2(\mathrm{SO}(3))$ that is SO(2) invariant. Then, f can just be defined on $\mathrm{SO}(3)/\mathrm{SO}(2)$-cosets, which we just saw are isomorphic to S^2. On the other hand, any function $f \in L^2(S^2)$ can be lifted to a function $f \in L^2(\mathrm{SO}(3))$ that is SO(2) invariant, so we have established an isomorphism

$$L^2(S^2) \cong L^2(\mathrm{SO}(3))^{\mathrm{SO}(2)},$$

where the superscript $^{\mathrm{SO}(2)}$ denotes the subspace of SO(2)-invariant functions. But now note that we can understand $L^2(\mathrm{SO}(3))$ very well using Theorem 4.11, and taking an SO(2)-invariant subspace commutes nicely with the decomposition we had. Indeed, we have

$$L^2(\mathrm{SO}(3))^{\mathrm{SO}(2)} = \Big[\widehat{\bigoplus_{\rho\in\widehat{\mathrm{SO}(3)}}} C(\mathrm{SO}(3))_\rho \Big]^{\mathrm{SO}(2)} = \widehat{\bigoplus_{\rho\in\widehat{\mathrm{SO}(3)}}} \Big(C(\mathrm{SO}(3))_\rho \Big)^{\mathrm{SO}(2)}.$$

Because of this, we see that we should try to understand the irreducible representations of SO(3) in order to understand $L^2(S^2)$.

6.1. *From* SU(2) *to* SO(3) *representations.* To derive the irreducible representations of SO(3), we will use a classical result that SU(2) is the double-cover of SO(3). Intuitively, this means that there is a way of mapping SU(2) onto SO(3) such that a 2π rotation in SO(3) corresponds to the map $-I$ in SU(2). The precise proof and details of this result are not so important for us (see [vdB93, Sec. 20] for the details). All that matters is that there is a surjective homomorphism

$$\phi : \mathrm{SU}(2) \to \mathrm{SO}(3) \quad \text{with} \quad \ker \phi = \pm I.$$

The existence of this homomorphism allows us to use what we already know about SU(2) representations, namely Proposition 5.4, to completely characterize SO(3) representations.

In particular, consider an irreducible representation $\tilde{\rho}$ of SO(3). We can then lift this to a representation $\rho = \tilde{\rho} \circ \phi$ of SU(2) where $-I$ acts as the identity. Since $\tilde{\rho}$ is irreducible, ρ must be as well, so we can conclude that $\rho = \rho_n$ for some $n \geq 0$. But if n is odd, then

$$\rho(-I)p(\mathbf{z}) = p(-\mathbf{z}) = (-1)^n p(\mathbf{z}) = -p(\mathbf{z}),$$

so $-I$ does not act as the identity. Thus, $n = 2k$ is even, and we have that $\tilde{\rho}$ is given by projecting ρ_{2k} onto SO(3). We thus get the following proposition.

PROPOSITION 6.1. *The irreducible representations of* SO(3) *are given by*

$$\tilde{\rho}_k = \rho_{2k} \circ \phi^{-1} \quad \text{for some} \quad k \geq 0.$$

In particular, SO(3) *has an irreducible representation of dimension* $2k+1$ *for every* $k \geq 0$.

Remark 6.2. One should reasonably object that ϕ^{-1} is not actually defined, since ϕ is not injective. Instead, by ϕ^{-1}, we mean any right inverse of ϕ (i.e., a map such that $\phi \circ \phi^{-1} = \mathrm{id}_{\mathrm{SO}(3)}$, which exists since ϕ is surjective). That the $\tilde{\rho}_k$ do not depend on the choice of right inverse then follows from the fact that the only ambiguity in $\phi^{-1}(g)$ is a $\pm$ freedom, which is irrelevant since $\rho_{2k}(-g) = \rho_{2k}(g)$.

Having this description of the irreducible representations, we can also ask what the characters of SO(3) are. To do so, note that complexifying SO(3) and applying the complex spectral theorem shows that any element of SO(3) is conjugate to a matrix

$$R_\theta = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix},$$

which form an SO(2) subgroup. Furthermore, while we will not show it, the preimage of R_θ under ϕ is given by

$$u_{\theta/2} = \begin{bmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{bmatrix},$$

where we are treating the range of θ as $[0, 4\pi)$ to make this "inverse" continuous. Intuitively, this is just the fact that a full 2π rotation in SO(3) corresponds to the map $-I$ in SU(2). With this in hand, we can calculate the characters.

COROLLARY 6.3. *The character of* $\tilde{\rho}_n$ *is given by*

$$\tilde{\chi}_n(R_\theta) = \sum_{k=-n}^{n} e^{i\theta k}.$$

Proof. We have

$$\tilde{\chi}_n(R_\theta) = \chi_{2n}(u_{\theta/2}) = \sum_{k=0}^{2n} e^{i\theta(2k-2n)/2} = \sum_{k=0}^{2n} e^{i\theta(k-n)} = \sum_{k=-n}^{n} e^{i\theta k}.$$

□

6.2. *Harmonic polynomials and* SO(3) *representations*. While we now have both the characters and descriptions of the irreducible representations, it is still worth thinking about a more direct realization of them. Namely, our current scheme requires lifting elements of SO(3) to SU(2), and then acting on complex 2-variable polynomials, which is rather involved. Ideally, we would directly relate SO(3) representations to functions of 3 real variables.

To do so, we will draw some further inspiration from the case of SU(2) representations and consider the homogeneous harmonic polynomials $\mathfrak{H}_n^3$ on $\mathbb{R}^3$. Since the Laplacian is invariant under rotations, there is a natural action of SO(3) on these polynomials given by

$$gp(\mathbf{r}) = p(g^{-1}\mathbf{r}).$$

As shown in Appendix A, $\mathfrak{H}_n^3$ is $2n+1$ dimensional, the same as $\tilde{\rho}_n$. This begs asking if these two representations are, in fact, isomorphic. Indeed, they are.

PROPOSITION 6.4. *The representations* $(\tilde{V}_n, \tilde{\rho}_n)$ *from Proposition 6.1 and* $(\mathfrak{H}_n^3, \mathfrak{p}_n)$ *of* SO(3) *are isomorphic. In particular, the* $\mathfrak{H}_n^3$ *also exhaust the irreducible representations of* SO(3).

Proof. Since the $\tilde{V}_i$ exhaust all irreducible representations of SO(3), we must have

$$\mathfrak{H}_n^3 = \bigoplus_{i \in I} \tilde{V}_{m_i}$$

for some indexing set I. Comparing dimensions, we have that

$$2n + 1 = \sum_{i \in I} 2m_i + 1.$$

In particular, we just need to show that $m_i \geq n$ for some i and we are done. To do so, we can compare the characters

$$\chi(R_\theta) = \sum_{i \in I} \sum_{k=-m_i}^{m_i} e^{i\theta k}.$$

Now, notice that if we can show $\mathfrak{p}_n(R_\theta)$ has an eigenvalue $e^{in\theta}$, then the sum on the right must include an $e^{in\theta}$ term as well, since a character is just the sum of eigenvalues. This would in turn show that one of the m_i is greater than n, as that is the only way an $e^{in\theta}$ term could appear.

To show this, consider the polynomial $Y_n(\mathbf{r}) = (y+iz)^n$. This is a harmonic homogeneous polynomial of degree n. Indeed, it is holomorphic as a function of $y+iz$, and it is a standard result of complex analysis that the real and imaginary parts of holomorphic functions are harmonic when regarded as functions of two real variables. Furthermore, we have

$$\begin{aligned} \mathfrak{p}_n(R_\theta) Y_n &= \Big(y \cos\theta - z \sin\theta + i(y \sin\theta + z\cos\theta) \Big)^n \\ &= \Big(e^{i\theta} y + i e^{i\theta} z \Big)^n \\ &= e^{i\theta n} Y_n, \end{aligned}$$

completing the proof. □

6.3. SO(3)/ SO(2) *and* S^2. Now that we have a very concrete understanding of the representations of SO(3), the only thing stopping us from determining a decomposition of $L^2(S^2)$ is an understanding of the spaces

$$C(\mathrm{SO}(3))_n^{\mathrm{SO}(2)}.$$

To do so, let us first consider the general case of a space

$$C(G)_\rho^H \quad \text{for} \quad H \subseteq G.$$

This is, by definition, just the space of matrix coefficients invariant under the H-action. Letting V be the vector space of ρ, we can further say that the space of matrix coefficients is isomorphic to the endomorphisms of V, since they match dimension. As such, we can interpret $C(G)_\rho^H$ as the subspace of $\mathrm{End}(V)$ invariant under the H-action, i.e., the endomorphisms

$$A \in \mathrm{End}(V) \quad \text{such that} \quad Ah = A \quad \text{for} \quad h \in H.$$

Using now the unitary structure of V, we conclude that $A = 0$ on the orthogonal complement of V^H, the subspace of V fixed by H. Indeed, we could otherwise restrict to a subspace of $(V^H)^\perp$ where A is invertible and conclude that H acts as the identity, a contradiction. Thus, restriction to V^H now induces an isomorphism

$$C(G)_\rho^H \cong \mathrm{Hom}(V^H, V).$$

Example 6.5. If we take $H = \{1\}$ to be the trivial subgroup, we are just asserting a homomorphism

$$C(G)_\rho \cong C(G)_\rho^H \cong \mathrm{Hom}(V^H, V) \cong \mathrm{End}(V).$$

If we take ρ to be the trivial representation, on the other hand, and let H be any subgroup, we are asserting that

$$\mathbb{C} \cong C(G)_{\mathbf{1}}^H \cong \mathrm{Hom}(\mathbb{C}^H, \mathbb{C}) \cong \mathbb{C}.$$

$\Diamond$

Applying this to the case of $C(\mathrm{SO}(3))_n^{\mathrm{SO}(2)}$, we see that we really just need to study

$$\mathrm{Hom}(\tilde{V}_n^{\mathrm{SO}(2)}, \tilde{V}_n) \cong \mathrm{Hom}(\mathcal{P}_{2n}^{\mathrm{SO}(2)}, \mathcal{P}_{2n}).$$

However, recall from earlier that the preimage of $\mathrm{SO}(2)$ in $\mathrm{SU}(2)$ is just $\mathrm{U}(1)$. Furthermore, the $\mathrm{U}(1)$ action on $\mathcal{P}_{2n}$ was given by

$$u_\theta p_k = e^{i\theta(2k-2n)} p_k.$$

This action is trivial only when $k = n$, so the space $\mathcal{P}_{2n}^{\mathrm{SO}(2)} \cong \mathcal{P}_{2n}^{\mathrm{U}(1)}$ is actually just one-dimensional, i.e., is just $\mathbb{C}$. In particular, we have

$$C(\mathrm{SO}(3))_n^{\mathrm{SO}(2)} \cong \mathrm{Hom}(\mathbb{C}, \tilde{V}_n) \cong \tilde{V}_n \cong \mathfrak{H}_n^3.$$

Thus, we have finally proved our main result on the decomposition of $L^2(S^2)$.

THEOREM 6.6 (S^2 spherical decomposition). *The space* $L^2(S^2) \cong L^2(\mathrm{SO}(3))^{\mathrm{SO}(2)}$ *decomposes as*

$$L^2(S^2) \cong \widehat{\bigoplus_{n \geq 0}} \mathfrak{H}_n^3|_{S^2}.$$

Furthermore, the coefficients in this decomposition can be calculated as

$$F_i^n = \langle f, m_i^n \rangle_{S^2} = \int_{S^2} f \, \overline{m_i^n} \, \mathrm{d}\mu \quad \text{for} \quad |i| < n.$$

Remark 6.7. One might wonder why we are allowed to freely restrict $\mathfrak{H}_n^3$ to S^2. However, this restriction actually loses no information, as

$$p(\mathbf{r}) = p(r\hat{\mathbf{r}}) = r^n p(\hat{\mathbf{r}}) \quad \text{for} \quad p \in \mathfrak{H}_n^3.$$

In particular, p is already determined by its values on S^2.

Also, as it turns out, the relevant metric $\mathrm{d}\mu$ to use for the integration is indeed the standard metric on a sphere $\mathrm{d}\Omega$, though we will not prove this.

7. Conclusions

The results we obtain here are about as far as we can go with just the Peter–Weyl theorem. However, there are definitely many ways to extend these results. For starters, any practical decomposition of functions on S^2 would ideally involve spherical polar coordinates, as these are the most natural. Indeed, it is possible to derive an explicit formula giving a basis of $\mathfrak{H}_n^3$ in terms of polar coordinates. For a reference, see [vdB93, Sec. 31].

Generalizing the decompositions discussed here to even-higher-dimensional spheres is also certainly possible. Indeed, just from the results we obtained, one might already guess that a decomposition into harmonic homogeneous polynomials is always possible. This is indeed the case, though a full proof certainly requires much more work. One way to approach this generalization would be to realize that we can always consider S^n as a quotient

$$S^n \cong \mathrm{SO}(n+1)/\,\mathrm{SO}(n)$$

and apply similar techniques as we did in Section 6. This path of generalization actually has connections to very current mathematical research, such as the Langlands program (see [GR06]).

On the other hand, one could also consider the idea of further generalizing the Peter–Weyl theorem. Unfortunately, a full generalization to even all locally compact groups is much more difficult. However, the special case of abelian locally compact groups is well understood thanks to *Pontryagin duality*, which an interested reader can find more about in [Rud17, Chap. 1.7]. This route then gives, for example, the Fourier transform on $\mathbb{R}$, among other things.

Appendix A. Dimension of $\mathfrak{H}_m^n$

Letting $P(\mathbb{R}^n)$ be the space of all polynomials on $\mathbb{R}^n$, we will consider the subspaces

$$\mathcal{P}_m^n = \{p \in P(\mathbb{R}^n) | p(\lambda \mathbf{r}) = \lambda^m p(\mathbf{r})\} \quad \text{and} \quad \mathfrak{H}_m^n = \{p \in \mathcal{P}_m^n | \Delta p = 0\}.$$

To start, we calculate the dimension of $\mathcal{P}_m^n$.

Proposition A.1.

$$\dim \mathcal{P}_m^n = \binom{n+m-1}{n-1}.$$

Proof. A basis for $\mathcal{P}_m^n$ is given by the monomials

$$x_1^{\alpha_1} x_2^{\alpha_2} \dots x_n^{\alpha_n} \quad \text{with} \quad \alpha_1 + \alpha_2 + \cdots + \alpha_n = m.$$

In particular, the number of such monomials is just the number of ways to write m as the sum of an n-tuple of nonnegative integers. This is just the classical stars-and-bars problem from combinatorics, with solution $\binom{n+m-1}{n-1}$. □

Having determined the dimension of $\mathcal{P}_m^n$, we can now determine the dimension of $\mathfrak{H}_m^n$ by cleverly decomposing it into lower-dimensional homogeneous polynomial spaces.

Proposition A.2.

$$\dim \mathfrak{H}_m^n = \dim \mathcal{P}_m^{n-1} + \dim \mathcal{P}_{m-1}^{n-1}.$$

Proof. Consider some $p = p(x_1, x_2, \dots, x_n) \in \mathfrak{H}_m^n$. We can expand this as a sum around x_1, giving

$$p = \sum_{k=0}^{m} \frac{f_k(x_2, \dots, x_n)}{k!} x_1^k.$$

Note that f_k is a homogeneous polynomial, now of degree $m-k$; in other words, $f_k \in \mathcal{P}_{m-k}^{n-1}$. Taking the Laplacian, we get

$$\begin{aligned}\Delta p &= \sum_{k=2}^{m} \frac{f_k}{k!} k(k-1) x_1^{k-2} + \sum_{k=0}^{m} \frac{x_1^k}{k!} (\Delta' f_k) \\ &= \sum_{k=0}^{m-2} \frac{f_{k+2}}{k!} x_1^k + \sum_{k=0}^{m} \frac{x_1^k}{k!} (\Delta' f_k),\end{aligned}$$

where the Δ' Laplacian excludes the x_1 coordinate. Analyzing the second term a bit more, we see that if $k = m, m-1$, then f_k is a polynomial of degree 0 or 1, so the Laplacian must vanish. Thus, we get

$$\Delta p = \sum_{k=0}^{m-2} \frac{x_1^k}{k!} \Big(f_{k+2} + \Delta' f_k \Big).$$

In particular, if p is harmonic, we must have

$$f_{k+2} + \Delta' f_k = 0 \quad \text{for} \quad 0 \le k \le m-2.$$

Thus, specifying f_0 and f_1 determines p. Namely

$$\mathfrak{H}^n_m \cong \mathcal{P}^{n-1}_m \oplus \mathcal{P}^{n-1}_{m-1},$$

which proves the proposition. □

Corollary A.3.

$$\dim \mathfrak{H}^n_m = \binom{n+m-2}{n-2} + \binom{n+m-3}{n-2}.$$

Example A.4. If we take $n = 3$, we get

$$\dim \mathfrak{H}^3_m = \binom{m+1}{1} + \binom{m}{1} = 2m+1,$$

proving the claim that $\mathfrak{H}^3_m$ and $\tilde{V}_m$ have the same dimension. ◊

Example A.5. If we take $n = 4$, we get

$$\begin{aligned}\dim \mathfrak{H}^4_m &= \binom{m+2}{2} + \binom{m+1}{2} \\ &= \frac{(m+2)(m+1)}{2} + \frac{(m+1)m}{2} \\ &= \frac{(m+1)(2m+2)}{2} \\ &= (m+1)^2,\end{aligned}$$

proving the claim that $\mathfrak{H}^4_m$ and $C(\mathrm{SU}(2))_m$ have the same dimension. ◊

References

[Axl19] Sheldon Axler. *Measure, Integration & Real Analysis*. Springer Nature, 2019.

[GR06] Benedict Gross and Mark Reeder. From laplace to langlands via representations of orthogonal groups. *Bulletin of the American Mathematical Society*, 43:163–205, 2006.

[Lee18] Seewoo Lee. s^2 is not a lie group. https://web.archive.org/web/20240303215238/https://math.berkeley.edu/~seewoo5/S2notLiegrp.pdf, 2018.

[Mor19] Sophie Morel. Representation theory. https://web.archive.org/web/20221223074434/https://perso.ens-lyon.fr/sophie.morel/rep_theory_notes.pdf, 2019.

[Rud17] Walter Rudin. *Fourier Analysis on Groups*. Dover Publications, 2017.

[vdB93] E.P. van den Ban. Harmonic analysis. https://web.archive.org/web/20220815001533/https://webspace.science.uu.nl/~ban00101/lecnotes/mc93.pdf, 1993.

Department of Mathematics, Columbia University
E-mail: lmn2163@columbia.edu
https://journals.library.columbia.edu/index.php/cjum

Columbia Journal of Mathematics **1** (2024),
99–115

Elliptic bootstrapping and the nonlinear Cauchy–Riemann equation

By Jessica J. Zhang

Abstract

The goal of this paper is to deduce a nonlinear elliptic regularity result from a linear one. In particular, elliptic bootstrapping is a powerful method to determine the regularity of a solution to a partial differential equation. We apply elliptic bootstrapping and linear elliptic regularity to the nonlinear Cauchy–Riemann equation. In doing so, we generalize the fundamental analytic result that holomorphic functions are automatically smooth. In particular, we show that, under certain conditions, the same is true for so-called J-holomorphic functions. We conclude by discussing how this nonlinear regularity result relates to ideas in symplectic geometry.

Suppose we have a C^k (i.e., k-times continuously differential) function $F : \mathbb{R}^n \to \mathbb{R}^n$. Suppose furthermore that we have a C^1-solution to the nonlinear ordinary differential equation

$$\dot{x} = F(x).$$

Roughly speaking, we see that x should have "one more derivative" than $F(x)$ via the following argument: Notice that $F(x) \in C^1$, so $\dot{x} \in C^1$ too. But this implies that $x \in C^2$. Thus $F(x)$ is actually in C^2, so that $\dot{x} \in C^2$ too. This implies that $x \in C^3$, and so on. We may continue this until we get that $F(x) \in C^k$, so $x \in C^{k+1}$. After this, even though $x \in C^{k+1}$, we cannot conclude that $F(x) \in C^{k+1}$ since F is only C^k. Thus we see that x is differentiable at least one more time than F is.

This is the essence of elliptic bootstrapping, namely by using the regularity of the coefficients of some differential equation in order to improve the regularity of any solution to that differential equation. (By "regularity," we simply mean "smoothness," or "how many times the function can be differentiated.")

Following the presentations in McDuff–Salamon [MS12, Appendix B] and Wendl [Wen15, Section 2.11], we give a more nontrivial example of elliptic bootstrapping.

In particular, one astonishing fact of complex analysis is that holomorphic functions are automatically smooth. Another way to phrase this is that "solutions to the Cauchy–Riemann equation are smooth." It turns out that this rests on a certain property of the Cauchy–Riemann equation known as *ellipticity*. While the general theory of elliptic partial differential equations is beyond the scope of this article, we will explore a generalization of the Cauchy–Riemann equation and prove via elliptic bootstrapping that its solutions are also automatically smooth.

One way to generalize holomorphic functions is to define a so-called *complex manifold*. An n-dimensional complex manifold is simply a $2n$-dimensional smooth manifold whose transition functions are holomorphic. Much as how we may talk about smooth functions on a smooth manifold, we may also talk about holomorphic functions on a complex manifold. A holomorphic function on a complex manifold is smooth: Locally, a complex manifold is exactly $\mathbb{C}^n$. But smoothness is a local condition, so the question of smoothness of holomorphic functions on a complex manifold reduces to the question of smoothness of holomorphic functions on $\mathbb{C}^n$.

There is, however, a further generalization of holomorphic functions to spaces known as *almost complex manifolds*. These manifolds arise naturally out of symplectic geometry, and they come with their own notions of holomorphic curves, often called *J-holomorphic* or *pseudoholomorphic* curves. These curves are solutions to the *nonlinear Cauchy–Riemann equation*, which generalizes the typical Cauchy–Riemann equations in complex analysis. We will prove via elliptic bootstrapping that, under relatively relaxed conditions, any J-holomorphic curve is automatically smooth.

We will discuss almost complex manifolds in Section 1. We will spend Section 2 introducing the Sobolev spaces $W^{k,p}$, which can be thought of as spaces of functions "admitting $k-n/p$ derivatives." In the end, using a bootstrapping argument, we will prove in Theorem 3.1 that, if the almost complex structure on an almost complex manifold is smooth, then any associated holomorphic curve is also smooth. In particular, we deduce a nonlinear elliptic regularity result from a linear one, which we state without proof. Finally, in Section 4, we will briefly and informally discuss the importance of this result in the context of symplectic geometry. We assume some familiarity with manifolds, multivariable calculus, and L^p-spaces. It would be helpful also to have seen some facts about complex analysis and partial differential equations.

1. *J*-holomorphic curves

An **almost complex structure** is a vector bundle homomorphism $J : TX \to TX$ such that $J^2 = -\operatorname{id}$ on the tangent spaces. We denote the set of C^ℓ-almost complex structures on a manifold X by $\mathcal{J}^\ell(X)$; if $\ell = \infty$, we also write $\mathcal{J}(X) := \mathcal{J}^\infty(X)$.

Example 1.1. Let $X = \mathbb{C}^n$ with coordinates $z_j = s_j + it_j$, and consider the standard complex structure J_0 on $\mathbb{C}^n$, which is defined on each tangent space $T_p\mathbb{C}^n = \mathbb{C}^n$ as

$$J_0\left(\left.\frac{\partial}{\partial s_j}\right|_p\right) = \left.\frac{\partial}{\partial t_j}\right|_p, \qquad J_0\left(\left.\frac{\partial}{\partial t_j}\right|_p\right) = -\left.\frac{\partial}{\partial s_j}\right|_p.$$

(From now on, we omit the subscript $|_p$, which only serves to denote which tangent space J_0 is acting on.) In other words, we may write J_0 in matrix form as

$$J_0 = \begin{pmatrix} 0 & -\mathbb{1} \\ \mathbb{1} & 0 \end{pmatrix},$$

where $\mathbb{1}$ is the $n \times n$ identity matrix. ◊

Many manifolds do not admit any almost complex structure at all. Indeed, we have the following proposition.

Proposition 1.2. *Suppose that X is an almost complex manifold, i.e., that it is a smooth manifold equipped with an almost complex structure J. Then X is even-dimensional and orientable.*

Proof. Say $\dim X = n$. If $p \in X$, then $J_p : T_pX \to T_pX$ is a vector space isomorphism between n-dimensional vector spaces such that $J_p^2 = -\operatorname{id}$, which has determinant $(-1)^n$. Thus $(-1)^n = (\det J_p)^2 \geq 0$, and so $n = 2k$ is even.

To show orientability, consider an arbitrary Riemannian metric h on X. Define $g(v, w) := h(v, w) + h(Jv, Jw)$, so that

$$g(Jv, Jw) = h(Jv, Jw) + h(J^2v, J^2w) = h(Jv, Jw) + (-1)^2h(v, w) = g(v, w).$$

Then define the $\omega(v, w) := g(v, Jw)$. Note that this is skew-symmetric since

$$\omega(w, v) = g(w, Jv) = g(Jw, J^2v) = -g(Jw, v) = -\omega(v, w)$$

by symmetry of g. On the other hand, we know that $\omega(v, -Jv) = g(v, v) \geq 0$, with equality if and only if $v = 0$. Thus ω is a nondegenerate 2-form. Then the k-th wedge product ω^k is a nowhere vanishing $2k$-form. But a nowhere vanishing top form defines an orientation, so we are done. □

Consider a compact two-dimensional smooth manifold Σ equipped with an almost complex structure j. (It turns out, in fact, that for this low-dimensional case, such a manifold is necessarily a *complex* manifold, in the sense that it admits coordinate charts with holomorphic transition functions [Don11, Theorem 22]. In general, however, almost complex does not imply complex, though the opposite is true.)

Our main object of study will be so-called *J-holomorphic curves* from (Σ, j) to the almost complex manifold (X, J). In particular, if $u \in C^\infty(\Sigma, X)$ satisfies

$$du \circ j = J \circ du,$$

then we call it a **J-holomorphic curve**.

We may now define an operator

$$\begin{aligned} \overline{\partial}_J : C^\infty(\Sigma, X) &\to \Omega^{0,1}(\Sigma, u^*TX) \\ u &\mapsto \frac{1}{2}(du + J \circ du \circ j) \end{aligned}$$

taking a smooth map $u : \Sigma \to X$ to a complex antilinear 1-form on Σ with values in the pullback tangent bundle

$$u^*TX = \{(p, v) : p \in \Sigma, v \in T_{u(p)}X\}.$$

By *complex antilinear*, we mean that it anticommutes with the almost complex strucutres; that is, we say $\omega \in \Omega^{0,1}(\Sigma, u^*TX)$ if $J \circ \omega = -\omega \circ j$. This operator $\overline{\partial}_J$ is often called the **del bar operator**. Then we have the following equivalent characterization of J-holomorphic curves.

LEMMA 1.3. *A smooth map $u : (\Sigma, j) \to (X, J)$ is J-holomorphic if and only if $\overline{\partial}_J(u) = 0$.*

Proof. Recall that $J^2 = -\operatorname{id}_{TX}$. Furthermore, we know that u is J-holomorphic if and only if $du \circ j = J \circ du$, which is in turn true if and only if $J \circ du - du \circ j = 0$. Now $-J$ is an isomorphism with inverse J, so $J \circ du - du \circ j = 0$ if and only if

$$du + J \circ du \circ -j = -J\,(J \circ du - du \circ j) = 0,$$

i.e., if and only if $\overline{\partial}_J u = 0$. This proves equivalence of our two definitions of J-holomorphic curves. $\square$

Remark 1.4. To understand the $\overline{\partial}_J$ operator more explicitly, note that at each point $p \in \Sigma$, we have

$$\overline{\partial}_J(u)(p) = \frac{1}{2}\left(du_p + J_{u(p)} \circ du_p \circ j_p\right).$$

Now $du_p : T_p\Sigma \to T_{u(p)}X$; this codomain is exactly the fiber of u^*TX over the point $u(p) \in X$. On the other hand, we know that j_p is an endomorphism of $T_p\Sigma$, while $J_{u(p)}$ is an endomorphism of $T_{u(p)}X$. Thus $J_{u(p)} \circ du_p \circ j_p$ makes sense, and also maps from $T_p\Sigma$ to $T_{u(p)}X$. In particular, this is what it means to be a form "on Σ with values in u^*TX."

Now to verify that $\overline{\partial}_J$ does indeed take values in $\Omega^{0,1}(\Sigma, u^*TX)$, it suffices to show that

$$J \circ \overline{\partial}_J(u) = -\overline{\partial}_J(u) \circ j.$$

(This is, indeed, what it means to be a *complex antilinear form.*) But we see that

$$2J \circ \overline{\partial}_J(u) = J \circ du + J^2 \circ du \circ j = -J \circ du \circ j^2 - du \circ j = -2\overline{\partial}_J(u) \circ j,$$

where all we use is the fact that $J^2 = -\,\mathrm{id}_{TX}$ and $j^2 = -\,\mathrm{id}_{T\Sigma}$.

Example 1.5. Let $\{U_\alpha, \phi_\alpha\}$ be holomorphic coordinate charts on Σ. That is to say, the maps $\phi_\alpha : U_\alpha \to \mathbb{C}$ are diffeomorphisms such that $\phi_\alpha \circ \phi_\beta^{-1}$ are holomorphic maps of (open subsets of) $\mathbb{C}$. Recall that the almost complex structure on Σ is induced by these coordinate charts and the complex structure J_0 on $\mathbb{C}$. Then $u : (\Sigma, j) \to (X, J)$ is J-holomorphic if and only if each

$$u_\alpha := u \circ \phi_\alpha^{-1} : (\mathbb{C}, J_0) \supseteq (\phi_\alpha(U_\alpha), J_0) \to (X, J)$$

is J-holomorphic. Letting the coordinates of $\phi_\alpha(U_\alpha) \subseteq \mathbb{C}$ be $z = s + it$, we see that

$$\begin{aligned}\overline{\partial}_J u_\alpha &= \frac{1}{2}\left(du_\alpha + J \circ du_\alpha \circ J_0\right)\\ &= \frac{1}{2}\partial_s u_\alpha ds + \frac{1}{2}\partial_t u_\alpha dt - \frac{1}{2}J(u_\alpha)\partial_s u_\alpha ds \circ J_0 + \frac{1}{2}J(u_\alpha)\partial_t u_\alpha dt \circ J_0.\end{aligned}$$

Notice, however, that

$$(ds \circ J_0)\left(\frac{\partial}{\partial s}\right) = ds\left(\frac{\partial}{\partial t}\right) = 0, \qquad (ds \circ J_0)\left(\frac{\partial}{\partial t}\right) = ds\left(-\frac{\partial}{\partial s}\right) = -1.$$

Thus $ds \circ J_0 = -dt$. Similarly, we may check that $dt \circ J_0 = ds$. We find that

$$\overline{\partial}_J u_\alpha = \frac{1}{2}\left(\partial_s u_\alpha + J(u_\alpha)\partial_t u_\alpha\right) ds + \frac{1}{2}\left(\partial_t u_\alpha - J(u_\alpha)\partial_s u_\alpha\right) dt.$$

It follows that u_α is J-holomorphic if and only if $\partial_s u_\alpha + J(u_\alpha)\partial_t u_\alpha = 0$.

Now suppose that Σ and X are both simply $\mathbb{C}$ equipped with the standard holomorphic structure. Write $u = f + ig : \mathbb{C} \to \mathbb{C}$. Then the condition that u is J-holomorphic is exactly that

$$(\partial_s f + i\partial_s g) + J_0(\partial_t f + i\partial_t g) = (\partial_s f + i\partial_s g) + (i\partial_t f - \partial_t g) = 0.$$

In other words, a curve $u : \mathbb{C} \to \mathbb{C}$ is J_0-holomorphic exactly when it satisfies the Cauchy–Riemann equations

$$\partial_s f = \partial_t g, \qquad \partial_s g = -\partial_t f,$$

i.e., when it is holomorphic. Because of this, the operator $\overline{\partial}_J$ is often called the **nonlinear Cauchy–Riemann operator**. ◊

2. **Sobolev spaces and weak equivalence**

Our eventual goal is to have a statement of regularity for the nonlinear Cauchy–Riemann equation. However, this regularity result requires that we define a more general kind of space, known as a Sobolev space.

Loosely speaking, if $k \geq 0$ is an integer and $p \geq 1$ is a (possibly infinite) real number, then the **Sobolev space** $W^{k,p}(\Omega)$ on some open set $\Omega \subset \mathbb{R}^n$ is defined to be the set of L^p-functions u whose k-th derivatives exist and are also p-integrable. In this context, we often call Ω a **domain**.

While the above definition of $W^{k,p}(\Omega)$ is a helpful way of thinking about the space, it is not entirely accurate. In particular, we require only that a function in $W^{k,p}(\Omega)$ admit so-called *weak* derivatives, as opposed to the usual derivatives, which are accordingly known as *strong* derivatives.

Suppose $u : \Omega \to \mathbb{R}$ is a locally integrable function for a domain $\Omega \subset \mathbb{R}^n$. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a multi-index of nonnegative integers α_i. Then the **α-th weak derivative** $D^\alpha u$ of u is a locally integrable function satisfying

$$\int_\Omega u(x)\partial^\alpha(\phi(x))\,dx = (-1)^{\alpha_1+\cdots+\alpha_n}\int_\Omega D^\alpha u(x)\phi(x)\,dx$$

for every compactly supported smooth function $\phi \in C_0^\infty(\Omega)$. Integration by parts implies that the above equation is always satisfied if u_α is the usual derivative. As such, this definition effectively asks that a weak derivative behave like the usual derivative under integration. Indeed, because integrals ignore what happens on a measure zero set, one may think of a weak derivative as a function which is the derivative almost everywhere.

Example 2.1. Let $\Omega = \mathbb{R}$, and define $u(x) = |x|$. This is locally integrable and admits the weak derivative

$$Du(x) = \begin{cases} -1 & \text{if } x < 0, \\ r & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Here r can be any real number. (In fact, any function which differs from the above formula for Du at a measure zero set is a weak derivative for u.) Furthermore, this weak derivative itself is p-integrable for any p, so that $u(x) \in W^{1,p}(\Omega)$. In fact, because u has further weak derivatives (namely functions which are 0 for $x \neq 0$), we actually have $u(x) \in W^{\infty,p}(\Omega)$. ◊

At this point, it is natural to wonder why we introduce the relatively complicated Sobolev spaces, rather than using C^k spaces, for example. The primary advantage is that Sobolev spaces are *complete*, which implies many theorems including Theorem 2.2 below. Indeed, we may define the **Sobolev**

norm

$$\|u\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \le k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p},$$

at least when $p \neq \infty$. Here the sum over $|\alpha| \le k$ indicates that we are summing over all multi-indices of length at most k. (When $p = \infty$, we may take the norm to be the maximum of the L^∞-norms of $D^\alpha u$, where α again ranges over all multi-indices of length at most k.) It turns out that this gives another way to define the Sobolev space $W^{k,p}(\Omega)$, namely as the completion of $C^\infty(\Omega)$ under the Sobolev norm $\|\cdot\|_{W^{k,p}(\Omega)}$, at least when $k \neq \infty$.

This definition of a Sobolev space generalizes to spaces of maps between manifolds, so that we may also define, for example, the space $W^{k,p}(\Sigma, X)$ to be the completion of $C^\infty(\Sigma, X)$ under the $W^{k,p}$-norm. For more information, one may look at [Wen15, pp. 126–128], for example.

To prove our regularity result, we will use a couple facts about Sobolev spaces.

THEOREM 2.2 (Sobolev embedding theorem). *Suppose* $\Omega \subset \mathbb{R}^n$ *is a bounded* C^1 *domain. If* $kp > n$, *then there is a continuous inclusion* $W^{k,p}(\Omega) \hookrightarrow C^0(\Omega)$. *If* $kp < n$, *then there is a continuous inclusion* $W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$, *where* $q = np/(n - kp)$.

The proof of this is rather difficult, but we will simply take it for granted here. An interested reader may find it as Theorem 6 in [Eva10, Section 5.6.3]. As a note, it is actually enough to have Ω be a bounded Lipschitz domain; since we will mostly be working with balls, however, we may restrict our attention to C^1 domains. Furthermore, the Sobolev embedding theorem actually says more than what we have mentioned here. In particular, it shows that, for certain k and p, this is actually a *compact* inclusion. We will not require that fact, however.

Because Σ is two-dimensional, we will primarily work with domains in $\mathbb{C} = \mathbb{R}^2$; thus, when we apply the Sobolev embedding theorem, we will generally have $n = 2$. In this $n = 2$ case, we have the following corollary of Hölder's inequality, which gives us our first use of the Sobolev embedding theorem and will be used in the proof of Theorem 3.3.

LEMMA 2.3 ([MS12, Lemma B.4.5]). *If* $p > 2$ *and* $1 < r \le p$ *and* $\Omega \subset \mathbb{R}^2$ *is any open set, then* $f \in W^{1,p}(\Omega)$ *and* $g \in W^{1,r}(\Omega)$ *together imply that* $fg \in W^{1,r}(\Omega)$.

Proof. It is enough to show that $D(fg) = f(Dg) + (Df)g \in L^r$ given that $Df \in L^p$ and $Dg \in L^r$. The Sobolev embedding theorem implies that that

$W^{1,p}(\mathbb{R}^2) \hookrightarrow C^0(\mathbb{R}^2)$, so that $f(Dg) \in C^0 \cdot L^r \subset L^r$. Thus it is sufficient to show that $(Df)g \in L^r$.

Since $1 < r \leq p$, there exists $q = pr/(p-r) \in (0, \infty]$ so that $1/p + 1/q = 1/r$. Now consider the following generalization of Hölder's inequality:

$$\|uv\|_{L^r} \leq \|u\|_{L^p} \|v\|_{L^q} .$$

In particular, it follows that

$$\|(Df)g\|_{L^r} \leq \|Df\|_{L^p} \|g\|_{L^q} .$$

Notice that $g, Dg \in W^{1,r}$ implies that $g, Dg \in W^{1,r'}$ for any $r' \leq r$. Thus without loss of generality $r < 2$, and so $g \in L^q$ by the Sobolev embedding theorem. Now since $p > 2$, we know that $q = pr/(p-r) < 2r/(2-r)$, and so $W^{1,r} \subset L^q$, proving the lemma. □

Before turning to the statement and proof of our elliptic regularity result for J-holomorphic curves, we return to and generalize the notion of a weak derivative. Recall that we asked a weak derivative to behave the same way as the usual derivative under integration. In general, we may call two functions **weakly equivalent** if they behave the same way under integration. That is to say, if $f, g \in L^1(\Omega)$ satisfy

$$\int_\Omega u\phi = \int_\Omega v\phi$$

for every compactly supported smooth function $\phi \in C_0^\infty(\Omega)$, then we call f and g weakly equivalent.

Finally, we take a moment here to standardize certain notation. In general, we will always use C_0^∞ to refer to compactly supported smooth functions, rather than simply functions which vanish near infinity. We sometimes call an element $\phi \in C_0^\infty$ a **test function**. Furthermore, when we say that Ω is a domain, we will always assume that Ω is a C^1 bounded open set in $\mathbb{R}^2$.

3. **Elliptic regularity**

With these results in mind, we are now ready to state and deduce the following elliptic regularity result for the nonlinear Cauchy–Riemann equation. We closely follow [MS12, Appendix B.4] here.

THEOREM 3.1 (Elliptic regularity, [MS12, Theorem B.4.1]). *Suppose $k \geq 2$ is an integer, and $p > 2$ is a real number. If $j \in \mathcal{J}(\Sigma)$, $J \in \mathcal{J}^k(X)$, and $u \in W^{1,p}(\Sigma, X)$ is J-holomorphic, then $u \in W^{k+1,p}(\Sigma, X)$. In particular, if J is smooth, then so too is any J-holomorphic curve u.*

Note that it is enough to prove this result locally, since being $W^{k+1,p}$ is a local condition. Thus, in this local setting, we may rephrase the theorem as follows: Suppose $\Omega \subseteq \mathbb{C}$ is open. Let J be a C^k-almost complex structure on $\mathbb{R}^{2n}$. (This J is obtained by pushing forward the original C^k-almost complex structure on X by a smooth local coordinate map.) Suppose furthermore that $u \in W^{1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ satisfies

$$\partial_s u + J(u)\partial_t u = 0.$$

Then $u \in W^{k+1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$. (Notice that we use local integrability here, since u need not satisfy any particular constraints at the boundary of Ω.)

Note that J is a C^k-almost complex structure on $\mathbb{R}^{2n}$, where $k \geq 2$. Thus $J \circ u$ is a $W^{1,p}_{\text{loc}}$-almost complex structure on the domain of u, namely Ω. In particular, we have $J \circ u \in W^{1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n \times 2n})$. Note now that u is a $(J \circ u)$-holomorphic map. If we can use this to show that u was actually in $W^{2,p}_{\text{loc}}$, then we would have that $J \circ u$ is actually a $W^{2,p}_{\text{loc}}$-almost complex structure, and so on. We would be able to continue this process on until $W^{k,p}_{\text{loc}}$. This argument is known as **elliptic bootstrapping**, and is used often to improve the regularity of solutions to elliptic partial differential equations.

In particular, it would be enough to prove the following.

THEOREM 3.2. *Suppose $\Omega \subseteq \mathbb{C}$ is an open, bounded, C^1 domain and $J \in W^{k,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n \times 2n})$ satisfies $J^2 = -\mathbb{1}$. If $\partial_s u + J\partial_t u = 0$ then $u \in W^{k+1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$.*

To prove this local version of elliptic regularity, we must first weaken our hypotheses somewhat. First, instead of requiring that $\partial_s u + J(u)\partial_t u = 0$, we must allow $\partial_s u + J(u)\partial_t u = \eta$ for some suitably regular $\eta : \Omega \to \mathbb{R}^{2n}$. Furthermore, we will actually want to consider $u \in L^q$ for some q, but the expression $\partial_s u$ is not well-defined in this case, since u is only integrable. Indeed, we want the notion, discussed in Section 2, of taking weak derivatives. For clarity we will explicitly state what weak equivalence means in this context.

If u had had first derivatives, then we would know by integration by parts that

$$(*) \qquad \int_\Omega \left\langle \partial_s \phi + J^T \partial_t \phi, u \right\rangle = -\int_\Omega \langle \phi, \partial_s u + \partial_t (Ju) \rangle = -\int_\Omega \langle \phi, \eta + (\partial_t J) u \rangle$$

for every test function $\phi \in C_0^\infty(\Omega, \mathbb{R}^{2n})$, where J^T denotes the transpose of $J : \Omega \to \text{End}(T\mathbb{R}^{2n}) = \mathbb{R}^{2n \times 2n}$. To see this equality, we use the fact that

$$\int_\Omega \langle \partial_s \phi, u \rangle = -\int_\Omega \langle \phi, \partial_s \rangle$$

by integration by parts, and that

$$\int_\Omega \left\langle J^T \partial_t \phi, u \right\rangle = \int_\Omega \langle \partial_t \phi, Ju \rangle$$

by definition of the transpose. Thus we will say that $\partial_s u + J\partial_t u = \eta$ **weakly** when Equation $(*)$ is satisfied.

We will prove the following proposition, which only assumes our weakened hypotheses.

PROPOSITION 3.3 ([MS12, Proposition B.4.9]). *Consider a bounded C^1 domain $\Omega \subset \mathbb{C}$. Let $J \in W^{\ell,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n\times 2n})$ satisfy $J^2 = -\mathbb{1}$, where ℓ is a positive integer and $p > 2$ is a real number. Suppose $u \in L^p_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ and $\eta \in W^{\ell,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ satisfy $\partial_s u + J\partial_t u = \eta$ weakly, i.e., satisfy Equation ($*$). Then $u \in W^{\ell+1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$, and $\partial_s u + J\partial_t u = \eta$ almost everywhere.*

This proposition proves Theorem 3.2, which in turn, as discussed earlier, proves our global statement of elliptic regularity in Theorem 3.1.

Proof of Theorem 3.2. Suppose $\partial_s u + J\partial_t u = 0$, where $J \in W^{k,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n\times 2n})$ satisfies $J^2 = -\mathbb{1}$. Notice that $\eta = 0$ is, in particular, an element of $W^{k,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ for every k. Now we apply Theorem 3.3 with $\eta = 0$ and $\ell = k$. Since $u \in W^{1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ also belongs to L^p, it follows that $u \in W^{k+1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$, as desired. □

Before we can prove Theorem 3.3, however, we must prove the following statement. Its main purpose is that, when combined with the second part of the Sobolev embedding theorem, this theorem "upgrades" regularity (for certain q): Theorem 3.4 says an L^q function is actually $W^{1,r}$, while the Sobolev embedding theorem says that, under certain conditions, this $W^{1,r}$ function is actually $L^{q'}$ for some $q' > q$.

PROPOSITION 3.4 ([MS12, Proposition B.4.6]). *Let $\Omega \subset \mathbb{C}$ be a bounded C^1 domain. Suppose $p, q, r \in \mathbb{R}_+ \cup \{\infty\}$ such that*

$$2 < p, \qquad 1 < r < \infty, \qquad \frac{1}{p} + \frac{1}{q} = \frac{1}{r}.$$

Suppose further that $J \in W^{1,p}_{\text{loc}}(\Omega, \mathbb{R}^{2n\times 2n})$ satisfies $J^2 = -\mathbb{1}$. Let $u \in L^q_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ and $\eta \in L^r_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ satisfy

$$\int_\Omega \left\langle \partial_s \phi + J^T \partial_t \phi, u \right\rangle = \int_\Omega \langle \phi, \eta + (\partial_t J)u \rangle$$

for every $\phi \in C_0^\infty(\Omega, \mathbb{R}^{2n})$. Then $u \in W^{1,r}_{\text{loc}}(\Omega, \mathbb{R}^{2n})$ and $\partial_s u + J\partial_t u = \eta$ almost everywhere.

To prove this, we require one highly nontrivial fact, known as *linear* elliptic regularity. In particular, recall that the Laplacian Δu is simply $\partial^2 u/\partial s^2 + \partial^2 u/\partial t^2$. Then we have the following theorem.

THEOREM 3.5 (Linear elliptic regularity, [MS12, Theorem B.3.1]). *If Δu is weakly equivalent to $\partial_s f + \partial_t g$ for some $f, g \in L^r$, then $u \in W^{1,r}$.*

In some ways, this is the key fact which allows our proof of Theorem 3.1 to go through. We do not describe a proof here, but for some intuition for this fact, notice that $\partial_s f, \partial_t g$ have one fewer derivative than f and g do; of course, since $f, g \in L^r$, we think of them as having "zero derivatives," so we can roughly think of $\partial_s f, \partial_t g$ as elements of $W^{-1,r}$. Then $\Delta u \in W^{-1,r}$, and so u, which has two more derivatives than Δu, should belong to $W^{1,r}$.

Proof of Theorem 3.4. Let $\psi \in C_0^\infty(\Omega, \mathbb{R}^{2n})$ be arbitrary. Then set

$$\phi := \partial_s \psi - J^T \partial_t \psi \in W^{1,p}(\Omega, \mathbb{R}^{2n}).$$

This belongs to $W^{1,p}$ since J, hence its transpose J^T, does. Notice that Equation ($*$) is satisfied for $W^{1,p}$ functions, too, since smooth functions are dense in Sobolev spaces. In particular, recall our alternate definition of Sobolev spaces as completions of C^∞ under the Sobolev norm. As such, since Equation ($*$) behaves well under limits, we may consider $W^{1,p}$ functions as well.

In particular, Equation ($*$) is satisfied for this particular value of ϕ, even though ϕ is not actually smooth. We may compute that

$$\begin{aligned}\partial_s \phi + J^T \partial_t \phi &= \partial_s^2 \psi - \partial_s \left(J^T \partial_t \psi \right) + J^T \partial_t \partial_s \psi - J^T \partial_t \left(J^T \partial_t \psi \right) \\ &= \partial_s^2 \psi - (\partial_s J^T)(\partial_t \psi) - J^T \partial_s \partial_t \psi + J_T \partial_t \partial_s \psi \\ &\quad - (J^T)^2 \partial_t^2 \psi - J^T (\partial_t J^T)(\partial_t \psi).\end{aligned}$$

Notice that $(J^T)^2 = (J^2)^T = -\mathbb{1}$. Furthermore, because J^2 is constant, we know that

$$0 = \partial_t (J^2) = J \partial_t J + (\partial_t J) J.$$

The same holds when we take transposes, and so we conclude that

$$\begin{aligned}\partial_s \phi + J^T \partial_t \phi &= \partial_s^2 \psi - (\partial_s J)^T (\partial_t \psi) + \partial_t^2 \psi + (\partial_s J)^T J^T (\partial_t \psi) \\ &= \Delta \psi - (\partial_s J)^T (\partial_t \psi) + (\partial_t J)^T J^T (\partial_t \psi).\end{aligned}$$

In particular, we find that

$$\int_\Omega \langle \Delta \psi, u \rangle = \int_\Omega \left\langle \partial_s \phi + J^T \partial_t \phi, u \right\rangle - \int_\Omega \left\langle (\partial_t J)^T J^T \partial_t \psi, u \right\rangle + \int_\Omega \left\langle (\partial_s J)^T (\partial_t \psi), u \right\rangle .$$

Using the fact that u and η satisfy Equation ($*$), we know that this first integral is equal to

$$- \int_\Omega \langle \phi, \eta + (\partial_t J) u \rangle = - \int_\Omega \left\langle \partial_s \psi - J^T \partial_t \psi, \eta + (\partial_t J) u \right\rangle .$$

Rearranging so that the left-hand terms in each of the inner products is either $\partial_s\psi$ or $\partial_t\psi$, we see that

$$\int_\Omega \langle \Delta\psi, u\rangle = -\int_\Omega \langle \partial_s\psi, \eta + (\partial_t J)u\rangle + \int_\Omega \langle \partial_t\psi, J\eta + J\left((\partial_t J)u\right)\rangle - \int_\Omega \langle \partial_t\psi, J\left((\partial_t J)u\right)\rangle + \int_\Omega \langle \partial_t\psi, (\partial_s J)u\rangle .$$

Setting $f := \eta + (\partial_t J)u$ and $g := -J\eta - (\partial_s J)u$, we now see that

$$\int_\Omega \langle \Delta\psi, u\rangle = -\int_\Omega \langle \partial_s\psi, f\rangle - \int_\Omega \langle \partial_t\psi, g\rangle .$$

At this point, we would like to use Theorem 3.5. By integration by parts, the above equation tells us that $\Delta u = \partial_s f + \partial_t g$ weakly, as they behave the same under integration. Now $\eta \in L^r_{\mathrm{loc}}$ by hypothesis. Furthermore, since $J \in W^{1,p}_{\mathrm{loc}}$, we know that $\partial_t J \in L^p_{\mathrm{loc}}$. Since $u \in L^q_{\mathrm{loc}}$, and $1/p + 1/q = 1/r$, we know that $(\partial_t J)u \in L^p_{\mathrm{loc}} \cdot L^q_{\mathrm{loc}} \subseteq L^r_{\mathrm{loc}}$. Hence $f \in L^r_{\mathrm{loc}}$. Similarly we may check that $g \in L^r_{\mathrm{loc}}$.

Thus Δu is weakly equivalent to $\partial_s f + \partial_t g$ for $f, g \in L^r_{\mathrm{loc}}$. Theorem 3.5 implies that $u \in W^{1,r}_{\mathrm{loc}}$, as desired. ☐

We now prove that Theorem 3.4 implies Theorem 3.3.

Proof of Theorem 3.3. We break this proof into three steps: First, assuming $k = 1$, we will prove that $u \in W^{1,p}$. Second, we will prove that $u \in W^{2,p}$, which completes the $k = 1$ case. Finally, we will prove the general case.

Step 1. $J \in W^{1,p}_{\mathrm{loc}}$, $\eta \in W^{1,p}_{\mathrm{loc}}$, $u \in L^p_{\mathrm{loc}}$ *implies* $u \in W^{1,p}_{\mathrm{loc}}$.

It is possible to find finite sequences $\{q_0, \ldots, q_m\}$ and $\{r_0, \ldots, r_m\}$ such that the following four conditions hold:

$$\frac{p}{p-1} < q_0 \leq p, \qquad q_{m-1} < \frac{2p}{p-2} < q_m,$$
$$q_{j+1} := \frac{2r_j}{2-r_j}, \qquad r_j := \frac{pq_j}{p+q_j}.$$

In particular, we define r_j so that $1/p + 1/q_j = 1/r_j$; furthermore, the conditions in the first line guarantee that $r_j \neq 1, \infty$, so that Theorem 3.4 can be applied. Furthermore, we define q^{j+1} so that Theorem 2.2 holds. Finally, we may verify that the endpoints q_m, r_m are defined so that $r_m > 2$ and $r_0, \ldots, r_{m-1} < 2$.

In particular, notice that $1 < \frac{p}{p-1} < q_0 \leq p$. Because $u \in L^p_{\mathrm{loc}}$, it follows that $u \in L^{q_0}_{\mathrm{loc}}$. Now by Theorem 3.4, we know that $u \in W^{1,r_0}_{\mathrm{loc}}$. Because $r_0 < 2$, it follows by Theorem 2.2 that $u \in L^{q_1}_{\mathrm{loc}}$ now.

Continuing in this fashion, we see that $u \in L^{q_m}_{\mathrm{loc}}$, and so $u \in W^{1,r_m}_{\mathrm{loc}}$. But now $r_m > 2$. By Theorem 2.2 again, we now have that $u \in C^0$, i.e., u is continuous.

But $C^0 \subset L^\infty_{\mathrm{loc}}$, and so we now have that $u \in L^\infty_{\mathrm{loc}}$. Furthermore, recall that $\eta \in W^{1,p}_{\mathrm{loc}}$, so it certainly belongs to L^p_{loc} as well. Applying Theorem 3.4 with $q = \infty$ and $r = p$ now implies that $u \in W^{1,p}_{\mathrm{loc}}(\Omega, \mathbb{R}^{2n})$, as desired. (Recall that q may be ∞; only r must be finite.)

STEP 2. *$J \in W^{1,p}_{\mathrm{loc}}$, $\eta \in W^{1,p}_{\mathrm{loc}}$, $u \in W^{1,p}_{\mathrm{loc}}$ implies $u \in W^{2,p}_{\mathrm{loc}}$.*

We will begin by showing the following lemma.

LEMMA 3.6. *If $q, r > 1$ such that $1/p + 1/q = 1/r$, and if $u \in W^{1,q}_{\mathrm{loc}}$, then $u \in W^{2,r}_{\mathrm{loc}}$.*

Proof. Fix such q and r. Then set

$$\widetilde{u} := \partial_s u \in L^q_{\mathrm{loc}}, \qquad \widetilde{\eta} := \partial_s \eta - (\partial_s J)\partial_t u.$$

Notice that $\partial_s \eta \in L^p_{\mathrm{loc}} \subset L^r_{\mathrm{loc}}$, where we use the fact that $p > r$. Furthermore, since $1/p + 1/q = 1/r$, we know that $(\partial_s J)\partial_t u \in L^p \cdot L^q_{\mathrm{loc}} \subseteq L^r_{\mathrm{loc}}$. Thus $\widetilde{\eta} \in L^r_{\mathrm{loc}}$.

It turns out that $\widetilde{u}$ and $\widetilde{\eta}$ satisfy Equation ($*$), in the sense that

$$\int_\Omega \left\langle \partial_s \phi + J^T \partial_t \phi, \widetilde{u} \right\rangle = -\int_\Omega \langle \phi, \partial_s \widetilde{u} + \partial_t (J\widetilde{u}) \rangle = -\int_\Omega \langle \phi, \widetilde{\eta} + (\partial_t J)\widetilde{u} \rangle$$

for all smooth test functions ϕ. To see this, observe that the first equality above follows directly from integration by parts. Thus it suffices to prove that

$$\partial_s \widetilde{u} + \partial_t (J\widetilde{u}) = \widetilde{\eta} + (\partial_t J)\widetilde{u}$$

weakly. But the left-hand side is exactly equal to

$$\partial_s \widetilde{u} + (\partial_t J)\widetilde{u} + J\partial_t \widetilde{u} = \partial_s^2 u + (\partial_t J)\partial_s u + J\partial_t \partial_s u.$$

On the other hand, using the definition for $\widetilde{\eta}$ and the hypothesis that $\partial_s u + J\partial_t u = \eta$ weakly, it follows that the right-hand side is given by

$$\partial_s (\partial_s u + J\partial_t u) - (\partial_s J)\partial_t u = \partial_s^2 u + (\partial_s J)\partial_t u + J\partial_s \partial_t u - (\partial_s J)\partial_t u = \partial_s^2 u + J\partial_s \partial_t u.$$

These two expressions are equal, since $\partial_s \partial_t u = \partial_t \partial_s u$. Thus $\widetilde{u}$ and $\widetilde{\eta}$ satisfy Equation ($*$), as desired.

But now we may apply Theorem 3.4 to conclude that $\partial_s u = \widetilde{u} \in W^{1,r}$. If we could show that $\partial_t u \in W^{1,r}_{\mathrm{loc}}$ as well, then we would have $u \in W^{2,r}_{\mathrm{loc}}$, proving the fact. But notice that

$$\partial_t u = J(\partial_s u - \eta) \in W^{1,p}_{\mathrm{loc}} \cdot W^{1,r}_{\mathrm{loc}} \subseteq W^{1,r}_{\mathrm{loc}},$$

where we use Theorem 2.3. This proves Theorem 3.6. ∎

Now the same argument using q_j and r_j from Step 1 holds. In particular, we eventually get that $u \in W^{2,r_j}_{\text{loc}}$ for each j; since $r_m > 2$, it follows that u is continuously differentiable, and hence belongs to $W^{1,\infty}_{\text{loc}}$. But now applying Theorem 3.6 with $q = \infty$ and $r = p$ implies that $u \in W^{2,p}$, as desired.

STEP 3. $J \in W^{k,p}_{\text{loc}}$, $\eta \in W^{k,p}_{\text{loc}}$, $u \in L^p_{\text{loc}}$ *implies* $u \in W^{k+1,p}_{\text{loc}}$.

We prove this inductively. In particular, suppose we have proven this step for some $k - 1 \geq 1$. Set $\widetilde{u}$ and $\widetilde{\eta}$ as before, so that they satisfy Equation ($*$) again. Then we find that $\partial_s u = \widetilde{u}$ and $\partial_t u$ are both in $W^{k-1,p}_{\text{loc}}$, so that $u \in W^{k,p}_{\text{loc}}$. This completes the induction. □

We showed earlier that Theorem 3.3 implies Theorem 3.2. As discussed toward the beginning of this section, Theorem 3.2 is a local statement of, and thus implies, our main regularity statement.

4. The moduli space of J-holomorphic curves

In this section, we discuss J-holomorphic curves in the context of symplectic geometry. This will be a relatively informal section; a small amount of algebraic topology (namely the notion of a fundamental class of a surface in homology) will be useful. We also briefly mention the first Chern class of a vector bundle, though it is only tangential to the larger story here.

A **symplectic form** ω on a smooth manifold X is a closed, nondegenerate 2-form. Being *closed* means that $d\omega = 0$, while being *nondegenerate* means that, for every nonzero tangent vector $v \in T_pX$, there exists $w \in T_pX$ so that $\omega_p(v, w) \neq 0$. If ω is a symplectic form on X, then we call (X, ω) a **symplectic manifold**. It turns out that any symplectic manifold has dimension $2n$, and ω^n is a nonvanishing top form, i.e., a volume form, on X. Hence X is orientable too.

Example 4.1. Consider the manifold $\mathbb{R}^{2n}$ (or $\mathbb{C}^n$). Define $\omega_{\text{std}} := dx_1 \wedge dy_1 + \cdots + dx_n \wedge dy_n$. Recall that $d^2 = 0$, so

$$d\omega_{\text{std}} = \sum_{i=1}^{n} (ddx_i \wedge dy_i - dx_i \wedge ddy_i) = 0.$$

Thus ω_{std} is closed. On the other hand, it is nondegenerate because

$$\omega_{\text{std}}(p) \left(\frac{\partial}{\partial x_i}\bigg|_p, \frac{\partial}{\partial y_i}\bigg|_p \right) = 1.$$

This is called the **standard symplectic structure**. In fact, Darboux's theorem says that every $2n$-dimensional symplectic manifold (X, ω) may be covered by coordinate charts in which the symplectic form may be written as $\omega = dx_1 \wedge dy_1 + \cdots + dx_n \wedge dy_n$. In particular, every symplectic manifold (X, ω)

is locally **symplectomorphic** to the standard symplectic manifold $(\mathbb{R}^{2n}, \omega_{\text{std}})$, in the sense that there are local diffeomorphisms ϕ between open sets of $\mathbb{R}^{2n}$ and X such that $\phi^*\omega = \omega_{\text{std}}$. ◊

Suppose now that J is an almost complex structure on X, i.e., is a map $J : TX \to TX$ with $J^2 = -\mathbb{1}$. If $\omega(v, Jv) > 0$ for every nonzero vector v and $\omega(v, w) = \omega(Jv, Jw)$ for every point $p \in X$ and every pair of vectors $v, w \in T_pX$, then we say that J is **ω-compatible**. The set of ω-compatible, C^ℓ-almost complex structures is written $\mathcal{J}^\ell(X, \omega)$. Furthermore, if $\ell = \infty$, then we omit the superscript.

Example 4.2. Recall the almost complex structure J_0 for $\mathbb{C}^n$ from Theorem 1.1. If $v = \sum_{i=1}^n \left(a_i \frac{\partial}{\partial x_i} + b_i \frac{\partial}{\partial y_i}\right)$ is a nonzero vector in $T_p\mathbb{R}^{2n}$, then we may compute

$$\begin{aligned}\omega(v, J_0v) &= \omega\left(\sum_{i=1}^n \left(a_i \frac{\partial}{\partial x_i} + b_i \frac{\partial}{\partial y_i}\right), \sum_{i=1}^n \left(a_i \frac{\partial}{\partial y_i} - b_i \frac{\partial}{\partial x_i}\right)\right)\\ &= \sum_{i=1}^n \left(a_i^2 + b_i^2\right) > 0.\end{aligned}$$

(Another way to show $\omega(v, J_0v) > 0$ for all nonzero v is to compute $\omega(v, J_0v) = 1$ for all basis vectors v.) A similar computation shows that

$$\omega(v, w) = \omega(J_0v, J_0w),$$

and so J_0 is ω_{std}-compatible. ◊

Let (X, ω) be a symplectic manifold with compatible smooth almost complex structure $J \in \mathcal{J}(X, \omega)$. Let (Σ, j) be a compact two-dimensional almost complex manifold. For every homology class $A \in H_2(X; \mathbb{Z})$, define the space

$$\mathcal{M}(A, \Sigma; J) := \{u \in C^\infty(\Sigma, X) : [u] = A \text{ and } \overline{\partial}_J u = 0\}.$$

Here $[u]$ is simply the pushforward $u_*[\Sigma]$ of the fundamental class of Σ. We call this space the **moduli space** of J-holomorphic curves representing A. (The phrase "moduli space" simply means that this is a space whose points correspond to certain geometric objects—which, in this case, are J-holomorphic curves.)

We will, however, focus on a slightly simpler moduli space, namely the moduli space of all J-holomorphic maps representing A which are *simple*. In particular, say (Σ', j') is another compact two-dimensional almost complex manifold, and say $u' : (\Sigma', j') \to (X, J)$ is J-holomorphic. Suppose furthermore that there is a holomorphic branched covering $\phi : \Sigma \to \Sigma'$ so that $u' \circ \phi = u$. If, in this setting, we always have $\deg \phi = 1$, then we call u **simple**. A more geometric way to think about simple J-holomorphic maps is as maps which do not "cover their image multiple times." Then

$$\mathcal{M}^*(A,\Sigma;J) := \{u \in C^\infty(\Sigma, X) : [u] = A, \overline{\partial}_J u = 0, \text{ and } u \text{ is simple}\}$$

is the subset of $\mathcal{M}(A,\Sigma;J)$ consisting of simple J-holomorphic curves.

A priori, this moduli space has no manifold structure. Even if it were clearly a manifold, it is not clear that it would be finite-dimensional. It turns out, however, that we have the following theorem.

THEOREM 4.3 ([MS12, Theorem 3.1.6]). *For "generic" $J \in \mathcal{J}(X,\omega)$, the moduli space $\mathcal{M}^*(A,\Sigma;J)$ is a manifold of finite dimension.*

Remark 4.4. By *generic*, we mean that J belongs to a set $\mathcal{J}_{\text{reg}}(X,\omega) \subset \mathcal{J}(X,\omega)$ which contains an intersection of countably many open and dense subsets of $\mathcal{J}(X,\omega)$. Such a set is called **residual**. It is worth noting that, often, the "natural" choice of J is not actually generic, and work must be done in order to perturb J to be in this set $\mathcal{J}_{\text{reg}}(X,\omega)$. Certain regularity criteria are presented in [MS12, Section 3.3].

Remark 4.5. The theorem in [MS12] actually gives an exact formula for the dimension of this moduli space, namely $n(2-2g) + 2\langle c_1(TX), A\rangle$. Here g is the genus of Σ and $c_1(TX) \in H^2(X;\mathbb{Z})$ is the first Chern class. The inner product is the standard pairing between cohomology and homology.

The proof of this theorem turns out to depend somewhat heavily on Theorem 3.1. In particular, the theorem implies that, if $J \in \mathcal{J}^\ell$, then the space of $W^{k,p}$ J-holomorphic curves is independent of k, so long as $k \leq \ell + 1$. In particular, the space of J-holomorphic curves of class $W^{k,p}$ is independent of k whenever J is a smooth almost complex structure. This lets us work in $W^{k,p}$-neighborhoods when necessary; combined with completeness, this will allow us to show that $\mathcal{M}^*(A,\Sigma;J)$ is a finite-dimensional smooth submanifold of the space $W^{k,p}(\Sigma,X)$ of J-holomorphic curves $u : \Sigma \to X$ of class $W^{k,p}$.

References

[Don11] Simon K. Donaldson. *Riemann Surfaces*. 1st ed. Vol. 22. Oxford Graduate Texts in Mathematics. Oxford University Press, 2011.

[Eva10] Lawrence C. Evans. *Partial Differential Equations*. 2nd ed. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, 2010.

[MS12] Dusa McDuff and Dietmar Salamon. *J-holomorphic Curves and Symplectic Topology*. 2nd ed. Vol. 52. Colloquium Publications. American Mathematical Society, 2012.

[Wen15] Chris Wendl. "Lectures on Holomorphic Curves in Symplectic and Contact Geometry". May 2015. URL: https://web.archive.org/web/20230522053304/https://www.mathematik.hu-berlin.de/~wendl/pub/jhol_bookv33.pdf (visited on 05/22/2023).

Department of Mathematics, Harvard University
E-mail: jjzhang@college.harvard.edu
https://journals.library.columbia.edu/index.php/cjum