THE DISC EMBEDDING THEOREM FOR 4-MANIFOLDS

JUDSON KUHRMAN

Abstract. In this expository paper, we intend to give an introduction to some of the basics of the topology of 4-manifolds. In particular, we will explain the disc embedding theorem, and the role which it plays in the work of Freedman on the study of 4-manifolds, and particularly its relevance to the 4-dimensional Poincaré conjecture. We assume familiarity with basic differential and algebraic topology but not much more.

Contents

1. Introduction

The main contents of this paper are all due, more or less, to Michael Freedman, whose work on the topology of 4-manifolds in the 1970s and 1980s brought about many interesting results. At the time, high-dimensional topology had been well-studied by mathematicians such as Stephen Smale and John Milnor, and powerful tools such as the h-cobordism theorem and surgery theory gave ways to classify smooth manifolds of high dimension up to homeomorphism. Such tools were not available to the study of topology in 3 and 4 dimensions. The work of Freedman was therefore largely centered around trying to find analogs between 4-dimensional topology and topology in higher dimensions. Ultimately, Freedman was able to prove deep results such as the classification of simply-connected 4-manifolds, which says that the homeomorphism type of a simply-connected 4-manifold is completely determined by its intersection form along with the Kirby-Siebenmann invariant, a cohomology class that detects the existence of a smooth structure, and conversely that every unimodular quadratic form on $\mathbb{Z}^2$ is represented by the intersection form of such a manifold. In this paper, we outline the ideas behind one major step in the proof of this theorem, the disc-embedding theorem.

Consider a smooth 4-dimensional manifold M and a closed, embedded curve γ in M. A fundamental question one might ask is, can we find an embedded disc with boundary γ? Moreover, can we find such a disc along with a framing, i.e. a tubular

Date: October 23, 2021.

neighborhood, that agrees with a fixed tubular neighborhood of γ? In 3 dimensions, the answer is certainly no, even in Euclidean space, since if γ is a knot bounding a disc D, then γ necessarily intersects D. On the other hand, in dimensions higher than 4, there are, roughly speaking, enough degrees of freedom that one can always find such a disc, at least if we assume that M is simply-connected.

The question is a bit weirder in dimension 4, but the answer turns out to be affirmative, at least topologically and under certain assumptions about the existence of a nice enough immersed disc bounded by γ (if γ does not extend to a map of a disc then the answer if obviously negative). More precisely, we intend to explain the following disc embedding theorem:

Theorem 1.1 (Disc embedding)**.** *Suppose M is simply-connected and suppose A is an immersed disc with embedded boundary in M and transverse sphere B, such that A and B have zero algebraic self-intersection. Then, there exists an embedded disc in M with the same framed boundary as A and with a transverse sphere.*

In fact, the hypothesis on $\pi_1(M)$ can be reduced from M being simply-connected to the Jordan-Hölder factors being finite or cyclic (in which case one has to worry a bit more about algebraic intersections, cf. [1] Section 5.1). Of course, there are some terms in this theorem statement which merit explaining. We will get to those in the course of the proof. Theorem 1.1, while appearing perhaps a bit technical, can be used to deduce deep and interesting results about 4-dimensional topology, such as the 4-dimensional Poincaré conjecture:

Theorem 1.2 (4-dimensional Poincaré conjecture)**.** *Let M be a closed smooth 4-manifold. If M has the homotopy type of a 4-sphere then M is homeomorphic to a 4-sphere.*

In the rest of this paper, we intend to explain Theorem 1.1, its proof, and the role that disc embeddings play in topology.

2. The High-Dimensional Case: Morse Theory and h-Cobordism

We begin by exploring the situation for high-dimensional manifolds (we will use "high-dimensional" to mean "of dimension ≥ 5"). Famously, one of the most important theorems in high-dimensional topology is Smale's h-cobordism theorem. For completeness, we state definitions first.

Definition 2.1. A (smooth, PL, or topological) *cobordism* is a triple $c = (W; V_0, V_1)$ of (smooth, PL, or topological) manifolds with W compact and $\partial W = V_0 \coprod V_1$. If the inclusions $V_0, V_1 \hookrightarrow W$ are homotopy equivalences then c is called an *h-cobordism*.

Then, the theorem states:

Theorem 2.2 (h-cobordism)**.** *Let $(W; V_0, V_1)$ be a smooth h-cobordism such that W, V_0, V_1 are simply-connected and $\dim W \geq 6$. Then, W is diffeomorphic rel. V_0 to the trivial cobordism $V_0 \times [0, 1]$.*

A particularly useful application of this theorem is in its ability to show that two high-dimensional smooth manifolds are diffeomorphic. Namely, to show that simply-connected manifolds V_0 and V_1 are diffeomorphic, where $\dim V_0 = \dim V_1 \geq 5$, it suffices to show that there exists a simply-connected h-cobordism $(W; V_0, V_1)$.

In particular, using the h-cobordism theorem one can relatively easily deduce the generalized Poincaré conjecture in high dimensions, that is

Corollary 2.3 (High-dimensional Poincaré conjecture)**.** *Let M be a closed smooth manifold of dimension $n \geq 5$. If M has the homotopy type of an n-sphere then M is homeomorphic to an n-sphere.*

To prove the generalized Poincaré conjecture, one first uses the h-cobordism theorem to deduce that any contractible smooth manifold with simply-connected boundary is diffeomorphic to a disc. Then, in the case $n \geq 6$, one considers an open disc $D \subseteq M$ and shows that $M \setminus D$ is contractible and hence M is the union of two discs glued by a homeomorphism of the boundaries, which can be shown to be homeomorphic to a sphere. The case $n = 5$ requires a bit of surgery theory to show that a homotopy 5-sphere bounds a smooth disc.

Now, what about in lower dimensions? Unfortunately, in dimensions 3 and 4, we do not have the h-cobordism theorem available to us. We do at least have the following weaker theorem in 4-dimensions, due to Freedman:

Theorem 2.4 (4-dimensional h-cobordism)**.** *Let $(W; V_0, V_1)$ be a smooth h-cobordism such that W, V_0, V_1 are simply-connected and $\dim W = 5$. Then, W is* homeomorphic *(not necessarily diffeomorphic) rel. V_0 to the trivial cobordism $V_0 \times [0, 1]$.*

This raises an obvious question: why is it that in high dimensions, the h-cobordism theorem holds *smoothly*, but in 4 dimensions it only holds *topologically*? To understand this question, we must first understand how the proof of the (high-dimensional) h-cobordism theorem works. Most of the nitty-gritty details are analytic in nature and will not be particularly relevant to our discussion, although the reader who has not previously seen the proof of Theorem 2.2 may certainly find it interesting and may be inclined to look at, say, Milnor's lecture notes on the subject [3].

Avoiding too much technicality, we outline the main ideas of the proof. For the rest of this section, let $(W; V_0, V_1)$ be a smooth cobordism with $\dim W = n$. The primary tool at our disposal in approaching the h-cobordism theorem is (as with many problems in differential toplogy) Morse theory, so we begin by recalling what, exactly, Morse theory allows us to do.

Morse theory begins with the study of Morse functions.

Definition 2.5. A critical point p of a smooth function $f : W \to \mathbb{R}$ is called *non-degenerate* if the Hessian $\mathrm{Hess}(f)_p$ is non-degenerate as a bilinear form. In this case, the *index* of f at p is the maximal dimension of a subspace on which $\mathrm{Hess}(f)_p$ is negative definite.

Definition 2.6. A *Morse function* on a smooth cobordism $(W; V_0, V_1)$ is a smooth function $f : W \to [a, b]$ such that

(i) All critical points of f are non-degenerate,
(ii) a, b are regular values of f, and
(iii) $f^{-1}(a) = V_0$ and $f^{-1}(b) = V_1$.

Given such a function, we use W^y to denote $f^{-1}((-\infty, y])$.

It should be noted that any Morse function can be be replaced with another with the same critical points and whatever critical values one pleases, so we assume that a Morse function f is *self-indexing*, that is the critical points of index k are exactly

the set $f^{-1}(k)$.

Morse functions allow us to study the topology of manifolds via their connection with handlebody decompositions.

Definition 2.7. A *k-handle* (of dimension n) is a pair $(H, \partial^+ H)$ diffeomorphic to $(D^k \times D^{n-k}, D^k \times S^{n-k-1})$. By *attaching* a k-handle, we mean gluing H to a manifold M by an embedding $\partial^+ H \hookrightarrow \partial M$.

Theorem 2.8. *Suppose there exists a Morse function $f : W \to [a, b]$ with no critical points. Then W is diffeomorphic rel. V_0 to $V_0 \times [0, 1]$.*

Theorem 2.9. *Let $f : W \to [a, b]$ be a Morse function. Suppose that $f^{-1}(k)$ contains exactly r many critical points, all of a given index k. Then (for small enough ϵ), $W^{k+\epsilon}$ is diffeomorphic to $W^{k-\epsilon}$ with r many k-handles attached.*

The idea behind these theorems is that one can use the flows of the (or more precisely a) gradient vector field associated to f to control topology. For Theorem 2.8, the gradient flow lines are actually the fibers $\{x\} \times [0, 1]$ in the product. The case of Theorem 2.9 is a bit more subtle, but one can use the gradient flow to deformation retract $M^{k+\epsilon}$ onto $M^{k-\epsilon}$ with k-cells attached, and a k-handle is just a k-cell times a disc.

Now, in light of Theorem 2.8 we would like a way to eliminate critical points of a Morse function. We have a few more definitions.

Definition 2.10. Let y be a regular value of a Morse function f on W, and p a critical point of f with index k. Let $V = f^{-1}(y)$.

(i) In the case $k < y$, we denote by $S_R(p)$ the set of points where the gradient flows of f through p intersect V. We call $S_R(p)$ the *right-hand* or *ascending sphere* of p in V.

(ii) In the case $k > y$, we define the *left-hand* or *descending sphere* $S_L(p)$ analogously.

It turns out that descending spheres of index k critical points are embedded copies of S^{k-1}, while ascending spheres are embedded copies of S^{n-k-1}. Their relevance is through the following fundamental theorem (5.4 in [3]):

Theorem 2.11. *Let p, q be critical points of a Morse function f with index $k, k+1$ respectively. Let $V = f^{-1}(y)$ for some $y \in (k, k+1)$ and suppose that $S_R(p)$ and $S_L(q)$ intersect transversely in exactly one point in V. Then, f can be modified near p and q so that p and q are no longer critical points.*

Now, it is good that we have a way of reducing the number of critical points of a Morse function, but the hypotheses for Theorem 2.11 appear fairly strong. Since the h-cobordism theorem assumes only a homotopy equivalence, we would like to somehow reduce the hypotheses to something algebraic-topological.

Recall that for oriented submanifolds $A, B \to V$ of complementary dimensions, we can define the homological intersection number $A \cdot B = \langle PD[A] \smile PD[B], [V] \rangle$, where $[A], [B], [V]$ are the homology classes associated to the orientations of A, B, V, and PD is the isomorphism $H_*(M) \to H^{n-*}(M)$ given by Poincaré duality. In the case that A, B are smoothly immersed and transverse, $A \cdot B$ equals the number of intersections of A and B counted with sign to account for orientation. We can then obtain a partial strengthening of the cancellation theorem.

Theorem 2.12. *Let p, q be critical points of a Morse function f with index $k, k+1$ respectively. If $k \geq 2$ and $k+1 \leq n-3$ and if $S_R(p) \cdot S_L(q) = \pm 1$, then f can be modified (near p and q) so that p and q are no longer critical points.*

The difficulty in proving Theorem 2.12 comes from the difficulty in promoting algebraic intersection numbers to bona-fide intersection numbers. That is, if we have some transverse embedded spheres $A, B \subseteq V$ of complementary dimensions, can we isotope A to some A' with $|A' \cap B| = |A \cdot B|$? With these hypotheses on the fundamental group (which are the reason why we need the additional bounds on k in the theorem), this can be accomplished by means of the so-called *Whitney trick*.

If we are given $A \cdot B = \pm 1$, then that means that the points of $A \cap B$ can be enumerated as $p_0, \ldots, p_r, q_1, \ldots, q_r$ where p_i and q_i have opposite signs for $1 \leq i \leq r$. Let γ be an embedded loop at p_r which first goes through A to q_r and then back to p_r through B. Then, the setup for the Whitney trick is the following: take a pair of curves γ_A and γ_B in the plane which together bound a disc, and let U be a neighborhood of said disc, as in Figure 1.

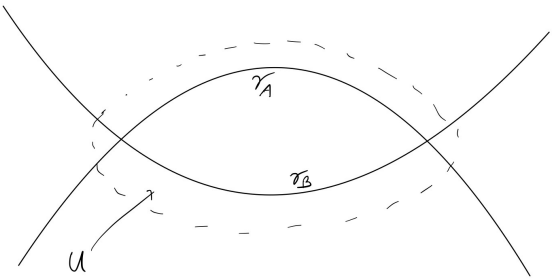

FIGURE 1. The model for a Whitney disc is a neighborhood of the region bounded by two parabolas

Under certain hypotheses on the dimensions of A and B, there exists an embedding $\phi : U \times \mathbb{R}^{\dim A - 1} \times \mathbb{R}^{\dim B - 1} \to V$ such that

(i) $\phi(\gamma_A \cup \gamma_B \times \{0\} \times \{0\}) = \gamma$,
(ii) $\phi^{-1}(A) = \gamma_A \times \mathbb{R}^{\dim A - 1} \times \{0\}$, and
(iii) $\phi^{-1}(B) = \gamma_B \times \{0\} \times \mathbb{R}^{\dim B - 1}$.

Now, we can perform an isotopy of $\gamma_A \times \mathbb{R}^{\dim A - 1} \times \{0\}$ with compact support to remove the intersection with $\gamma_B \times \{0\} \times \mathbb{R}^{\dim B - 1}$. Thus, embedding Whitney discs allows us to remove pairs of intersection points with opposite sign. Therefore, under certain necessary hypotheses, we can remove pairs of intersection points of transverse spheres, thereby promoting algebraic data (in the form of intersection numbers) to geometric data and allowing us to apply the weaker cancellation theorem.

Returning to the discussion of the h-cobordism theorem, by using some additional Morse theory tricks, one can modify a Morse function on a simply-connected h-cobordism so that all critical points lie within the range where Theorem 2.12 applies. At this point, one has to introduce Morse homology, a way of calculating $H_*(W, V_0)$ using the data from a Morse function, and the fact that $H_*(W, V_0) = 0$

allows one to show that the critical points can be paired in such a way that Theorem 2.12 can be used to cancel all critical points. Thus, we obtain a Morse function with no critical points.

The problem arises when we try to apply the Whitney trick in 4-dimensions. This is because in 4-dimensions, a generic 2-disc is only immersed, unlike in higher dimensions where 2-discs are generically embedded. Thus, we may not be able to find a framed embedding of a disc with boundary γ, but only an immersion. With this in mind, we need some way of actively removing intersections in immersed discs.

3. Manipulating Intersections

Take M to be a smooth 4-manifold. We want to understand, ultimately, how to remove intersections (and in particular self-intersections) between surfaces in M. For this, we first need to understand what geometry is like in four dimensions. We can think of 4-dimensional space (at least locally) as having three spacial dimensions and one time dimension, or equivalently as infinitely many 3-dimensional frames superimposed on each other, one for each real number.

Taking this point of view, most of the constructions we work with will have some 3-dimensional model, which we can make "nicer" by pushing certain parts "into the fourth dimension." This allows us to draw pictures of 4-dimensional space. As a pet example, consider the case of a knot K in $\mathbb{R}^3$. If we are allowed to pass in and out of the fourth dimension, we can change crossings of K by locally pushing different points either forwards or backwards in time, and thus there are no knots in $\mathbb{R}^4$. In Figure 2, we visualize this by using color to represent points which occur at different times.

Figure 2. Changing knot crossings.

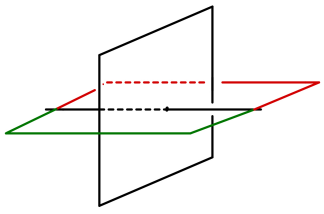

Figure 3. Intersecting surfaces.

Similarly, we can use an apparently "3-dimensional" picture to visualize surfaces intersecting in a single point, as in Figure 3, where green points "occur in the future"

and red points "occur in the past." By a surface in M, we mean an immersion $S \to M$ where S is a surface. It will often be enough to consider smooth immersions, but we should really be talking about topological framed immersions. Roughly speaking, this means maps which admit tubular neighborhoods. More precisely, we start with $S \times D^2$ as a model. Take two disjoint open sets $U_1, U_2 \subseteq S$, each with coordinate charts $\phi_i : U_i \to D^2$, and identify $U_1 \times D^2$ with $U_2 \times D^2$ by the relation $(x, y) \sim (\phi_2^{-1}(y), \phi_1(x))$, pictured in Figure 4. Such an identification is called a *plumbing* of the model. After applying some finite number of plumbings,

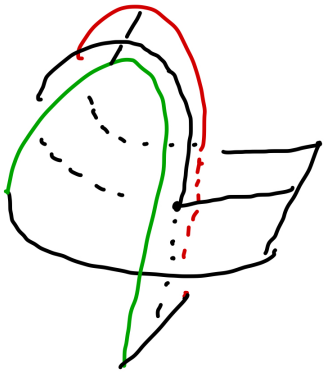

FIGURE 4. A plumbing of a disc: the disc has one double-point, where the normal plane of one preimage is the tangent plane of the other.

we obtain a quotient N of $S \times D^2$, where the image of $S \times \{0\}$ has finitely many self-intersections. An immersion of S means a topological embedding $N \to M$. We call the interior of N (i.e. the image of $S \times \operatorname{int} D^2$) a *regular neighborhood* of S.

We have already encountered one technique for manipulating intersections: the Whitney move, where we slide one surface (or more generally submanifold) over a disc to remove intersections with another surface. This move has a sort of inverse: the finger move, whose model is shown in Figure 5. This move pushes a "finger" of one surface over another, creating a pair of intersections with a Whitney disc.

Now, why would we do this? We are trying to get rid of intersections, after all, and it seems like we've just introduced new ones. For one thing, it should be noted that we have not introduced any *essential* intersections, since any intersections introduced by a finger move can be removed by a Whitney move, and intersections with Whitney discs are, at least morally speaking, almost as good as no intersections at all.

A more pragmatic and precise reason is that finger moves allow us to realize connected sums of surfaces in M. Recall that the abstract connected sum of two (equidimensional) manifolds A and B is the manifold $A\#B$ obtained by removing small balls from each and identifying their boundary (we use the word abstract to distinguish arbitrary manifolds from submanifolds of M). Generally, if A and B are surfaces in M, there is not exactly an *a priori* reason that $A\#B$ should also sit inside M, and this is one place where finger moves are useful.

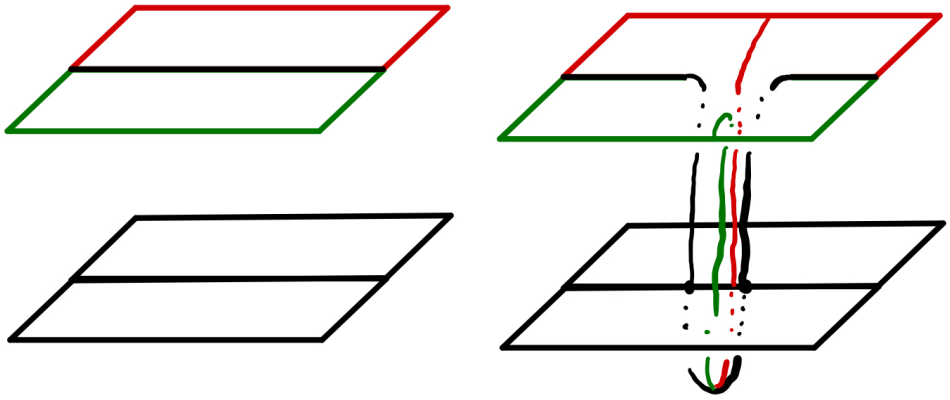

FIGURE 5. A finger move.

For example, if we take A and B to be the surfaces Figure 5, by pushing the tip of the finger of A back into the present, we obtain a circle where the two surfaces intersect. By ignoring the discs which these circles bound, we see that the $A\#B$ is just the union of A and B, minus some discs and plus part of a finger. More generally, if points $p \in A$ and $q \in B$ are connected by a path, then we can push a finger of A along this path and form the connected sum with B. We could also form the connected sum of A with itself, which raises the genus of A by 1 (in this situation, $A\#A \subseteq M$ is *not* actually a realization of the abstract connected sum of A with itself, but rather the result of attaching a cylinder to A with a pair of discs removed; abstractly this is connected sum of A with a torus).

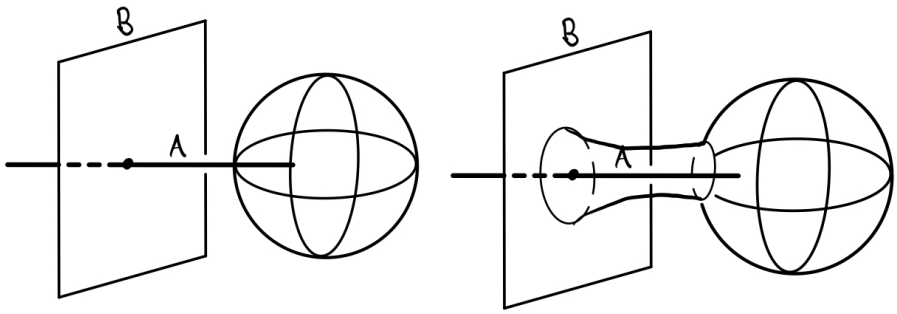

FIGURE 6. Removing intersections with transverse spheres. This gives a new surface which is regularly homotopic to B but which is disjoint from A.

Now, suppose that A and B intersect transversely, and suppose there exists a

(2-dimensional) sphere S in M which intersects A transversely at a single point. We call S a *transverse sphere* for A. In this situation, we can push a finger of B along A and form the connected sum with S to obtain a new surface B' with the same regular homotopy type as B and with no intersections with A, as in Figure 6 (*regular homotopy* here means a homotopy through immersions respecting the framed boundary of B). It should be noted that this may create new self-intersections of B, if S is not embedded.

With these tools established, we can now introduce capped surfaces, the first step towards the disc embedding theorem.

4. Capped Surfaces and Towers

We first start with an abstract model. Let S be a surface. Inside $S \times (-1, 1)$, we can form the connected sum of S with itself. This raises the genus of S by 1. Perform this operation some finite number of times to obtain a surface S'. Each time the operation is performed, two new generators are created in π_1. There are standard embedded discs, called *caps*, in $S \times (-1, 1)$ that kill these generators. The data of S' along with the caps constitute an *S-like capped surface*. An example is given in Figure 7. Now, we may define iterated capped surfaces.

Definition 4.1. A 0-stage S-like iterated capped surface is a copy of S. A 1-stage S-like iterated capped surface is an S-like iterated capped surface. An $(n+1)$-stage S-like capped surface is obtained from an n-stage S-like iterated capped surface by replacing some of the caps with disc-like capped surfaces. The *height* of an iterated capped surface is the minimal number of stages.

Remark 4.2. We note here as a warning that historically, such objects have been referred to as "gropes," such that the term is largely unavoidable in references. We will not otherwise use this term.

Such a construction can be done entirely inside $S \times (-1, 1)$, and hence inside $S \times D^2$, to give an abstract S-like iterated capped surface. Let E be an abstract S-like iterated capped surface. As with surfaces, an immersion of E means a map $E \to M$ given by the composition of some plumbings followed by an embedding. We say that such an immersion is *proper* if the plumbings occur only among the caps. That is, the body does not intersect itself or the caps (the body being the closure of $E \setminus \{\text{caps}\}$). Sometimes, we prefer that intersections occur between the caps and the body. In this case, we can "push down" intersections between the caps using finger moves which result in one of the caps intersecting a lower stage of the iterated capped surface. In this process, we gain 2 intersections with the body for every intersection with the caps.

In constructing a capped surface from a surface, we necessarily raise the surface's genus. However, this is more-or-less reversible via what is called *contraction*. This operation can be described as: cut out regular neighborhoods of each of the caps (in the abstract model) and fill in the new boundary components with parallel copies of the caps. Contraction of an S-like capped surface results in an immersion of S, whose self-intersections are exactly those coming from the self-intersections of the caps. Moreover, since contraction can be done in a regular neighborhood, we see that a regular neighborhood of an S-like capped surface is also a regular neighborhood of S. Inductively, contraction can be applied to an S-like iterated capped surface to return a copy of S.

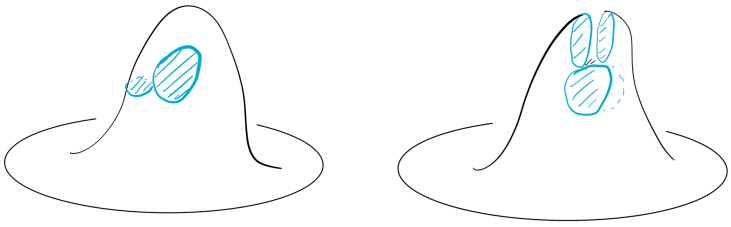

FIGURE 7. A disc-like capped surface and its contraction, with caps in blue. The cap on the left is "horizontal" and the cap on the right is "vertical."

Now, we are trying to prove something about discs, so to make any use of iterated capped surfaces we should hope that we can find such objects in place of discs. In fact this is true.

Lemma 4.3. Let A be as in the hypothesis of Theorem 1.1. Then, there exists a properly immersed iterated capped surface E in M of height at least 2 with the same framed boundary as A.

Proof. See [2], Proposition 3.11. The rough idea is that we can trade self-intersections for genus by forming the connected sum of A with itself, and then use S to generate caps to obtain a properly immersed capped surface. The tricky bit is showing that we can repeat the process for the caps. □

We have already seen the utility of transverse spheres for removing intersection. Similarly, for a surface A in M we can define a *transverse iterated capped surface* for A to be a properly immersed sphere-like iterated capped surface E whose body intersects A transversely in a single point. If A is an iterated capped surface rather than a surface, then we insist that the body of E and the body of A have a single transverse intersection, and neither body intersects the caps of the other. This intersection should happen in the lowest stage of both iterated capped surfaces.

Now, how do transverse iterated capped surfaces arise? There is one particular situation in which they arise and turn out to be quite useful. Given an iterated capped surface E, let E_* denote all of E except the initial stage. We can divide the components of E_* into two groups, denoted E_+ and E_-, based on whether they contract to a "vertical" or "horizontal" disc in the model (see Figure 7), where by components we mean connected components of the body along with their respective caps. Let F_+ and F_- be components of E_+ and E_-, respectively, which contract to discs whose boundaries touch each other. Take a cylinder $S^1 \times I$ formed by taking $\partial F_- \times I$ and perturbing it to remove the intersection with ∂F_-. Fill in the boundary components of this cylinder with parallel copies of F_-. Then, after pushing things into the fourth dimension to make intersections transverse, we obtain a transverse iterated capped surface E_+^t for E_+, which is otherwise disjoint from E except possibly for intersections with caps. With this construction, we can prove the first important theorem on iterated capped surfaces.

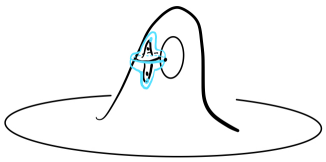

FIGURE 8. A transverse capped surface obtained from a cap by attaching parallel copies of the cap at the boundary circles of an annulus.

Theorem 4.4 (Height raising for iterated capped surfaces)**.** *Let E be a properly immersed iterated capped surface in M of height at least 2. Then the body of E can be extended to an iterated capped surface of any finite height.*

Proof. Suppose that E_+ has height m and E_- has height n, which we denote by saying that E has height (m, n). Without loss of generality we may assume that $n \geq 1$. The first step is to remove intersections among the caps. As described above, we can use E_- to form transverse iterated capped surfaces $E_1^t, E_2^t, \ldots, E_k^t$ for each of the components of E_+. Fully contracting these iterated capped surfaces gives spheres $S_1, \ldots, S_k$ such that for every intersection of a cap C of E and a cap of E_i^t, there are two intersections of C with S_i. These intersections come with immersed Whitney discs, which we may use to remove the intersections of C and S_i, possibly introducing new self-intersections of C. This makes the S_i into transverse spheres for the components. Pushing down intersections of the caps, the S_i become transverse spheres for the caps.

For each intersection of an E_+ cap C_1 with another cap C_2, we may make a parallel copy S of one of the transverse spheres S_i. Then, we may remove the intersection by forming the connected sum $C_2 \# S$. This may add self-intersections to C_2 but this is not an issue. Similarly, we may use transverse iterated capped surfaces to obtain transverse spheres and remove intersections between the E_- caps.

At this point, we are left with the situation that the only self-intersections of E are self-intersections between the caps (that is, distinct caps are disjoint). Let C_+ be a cap of E_+. If C_+ is embedded, then we can extend the iterated capped surface by embedding an abstract disc-like capped surface in place of C_+. Otherwise, if C_+ has self intersections, we may form connected sums with transverse iterated capped surfaces. Since these transverse iterated capped surfaces are formed from E_-, they have height at least 1, and so this operation raises the height of E_+. Thus, we are able to raise an iterated capped surface of height (m, n) to height $(m + n, n)$. Similarly, we can raise (m, n) to $(m, m + n)$. Therefore, if we start with at least one of m and n nonzero, we can achieve arbitrary height. □

Thus, we can extend iterated capped surfaces in M to arbitrary height. However, a finite-height iterated capped surface does not really carry that much useful information. We really want to extend an iterated capped surface to "infinite height" in a way that meaningfully converges. For this, we need to modify the construction slightly and introduce towers. Essentially, towers are like iterated capped surfaces where we interrupt the height raising every so often to change the caps, such that we gain more control over the behavior at infinity.

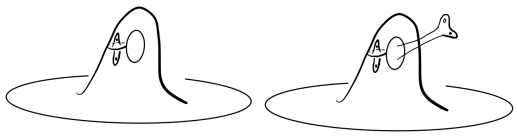

FIGURE 9. Extension from height (0,1) to height (1,1).

Consider the disc $D^2 \subseteq D^4$. By performing a finger move, we can introduce a pair of self-intersections in D^2, which come equipped with a Whitney disc, along with a second disc called the *accessory disc* such that the two discs kill all resulting essential curves.

Definition 4.5. An abstract *1-story tower* is obtained from an abstract iterated capped surface by performing some finite number of such finger moves to the iterated capped surface caps, for which the resulting Whitney and accessory discs are mutually disjoint. Then, the Whitney and accessory discs are called *tower caps*, and the rest of the tower (including the iterated capped surface caps) is called the *body*. An $(n+1)$-story tower is obtained from an n-story tower by replacing the tower caps with 1-story towers.

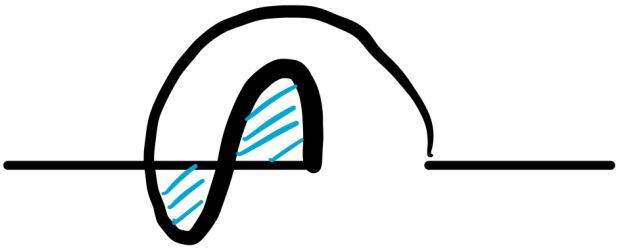

FIGURE 10. Whitney (left) and accessory (right) discs resulting from a finger move.

As with iterated capped surfaces, a *proper immersion* of a tower is an immersion such that the body intersects neither itself nor the tower caps. At this point, we introduce the concept of π_1-nullity.

Definition 4.6. An immersion $A \to M$ is called π_1-null if the induced map $\pi_1(\operatorname{im} A) \to \pi_1(M)$ is trivial.

This concept plays an important role in the convergence of towers, and in particular the reason that we need to introduce towers in the first place is that each story of a tower can be made π_1-null in the complement of the previous story. The general principle here is that the Whitney and accessory discs of a tower kill π_1 of the body, so if we can raise a tower to, say, $(n+1)$ stories, then contracting the top-story iterated capped surfaces gives a π_1-null tower of n stories.

First, to make any use of towers, we must be able to actually produce them. The following lemma tells us that we may obtain 1-story towers from nice enough properly immersed iterated capped surfaces.

Lemma 4.7. Let E be a properly immersed disc-like iterated capped surface in M. Suppose E is π_1-null and has a transverse sphere S. Then, the body of E extends to a 1-story capped tower with a transverse sphere.

Proof. As in the proof of iterated capped surface height raising, we can use transverse iterated capped surfaces to reduce to the case that the caps of E intersect only themselves, and that the caps have transverse spheres. Since E was initially π_1-null and this modification happens in a regular neighborhood of E, we can perform the reduction while maintaining π_1-nullity. Thus, we can find Whitney and accessory discs for the self-intersections of the caps killing any essential curves in E. To ensure that these discs do not intersect the body, we push intersections down and form connected sums with parallel copies of S.

At this point, we have effectively extended the body of E to a 1-story capped tower T. Repeating the transverse iterated capped surface contruction and contracting, we obtain transverse spheres for the caps of T, which we use to remove intersections with S. Thus, S gets modified to become a transverse sphere for T. □

Now, we would like a height raising theorem for towers, analogous to that for iterated capped surfaces. As for iterated capped surfaces, we need a notion of transverse iterated capped surfaces for towers. Let T be a tower and E its first-story iterated capped surface. We may form a transverse iterated capped surface E^t_+ for E_+. Now, the caps of E^t_+ may intersect the (iterated capped surface) caps of E_-, which is undesirable.

Such intersections come in opposite-sign pairs, so can be joined up by Whitney arcs (i.e. the arcs between intersection points which arise in the model for the Whitney move). Cutting out neighborhoods of these arcs in E^t_+, and replacing each with two parallel copies of the second story, we remove the intersections of E^t_+ with the first story, but possibly introduce intersections with the second-story iterated capped surface caps.

Repeating this process until E^t_+ intersects only the iterated capped surface caps of the top story and then contracting the top stage and pushing off the iterated capped surface caps of T, we obtain E^t_+ disjoint from T except for intersections among the caps. Notice that this may introduce new self-intersections of the iterated capped surface caps of T, but such self-intersections come with new Whitney and accessory discs. Moreover, essential loops in E^t_+ are killed by parallel copies of the tower caps of T, and hence E^t_+ is π_1-null. We now have the following height raising theorem:

Theorem 4.8 (Height raising for towers)**.** *Let T be a properly immersed n-story capped tower, whose first story has iterated capped surface height at least 3. Then, the body of T can be extended to an $(n+1)$-story properly immersed capped tower whose top story has arbitrary iterated capped surface height.*

Proof. Let E be the first-story iterated capped surface of T. Begin by constructing π_1-null transverse iterated capped surfaces E^t_+ and E^t_- as above. At this point, we can ignore the first stage of E. That is, take our ambient manifold to be M', the complement in M of a compact neighborhood of the first stage of E, so that E_+, E_- now constitute the first stage of our tower.

Suppose we want the top story of our final tower to have iterated capped surface height at least N. Then, we use Theorem 4.4 to raise the iterated capped surface

height of E^t_+ to $N+2$, and contract so that E^t_+ has height $N+1$ and is disjoint from T except at the transverse points. Since all this occurs in a regular neighborhood, π_1-nullity is maintained. Push self-intersections of the caps of T down and form connected sums with parallel copies of E^t_+.

This extends the body of T by replacing the tower caps with height $N+1$ iterated capped surfaces, with intersections among the iterated capped surface caps. Contracting these iterated capped surfaces and pushing them off each other makes them disjoint, and the new self-intersections among the iterated capped surface caps have Whitney and accessory discs – Whitney discs since they were formed by pushing off of a contraction and accessory discs by π_1-nullity. Thus, we have effectively extended the body of T_+ to $(n+1)$ stories, and the top story has height N. Any intersections of the tower caps with the body may then be pushed down and removed using transverse spheres constructed from E^t_+. Similarly, we can extend T_-. □

Now that we have height raising for towers, we can begin to control them. This is where towers really show their advantage over basic iterated capped surfaces.

Lemma 4.9 (Tower squeezing)**.** Let T be a properly immersed 1-story disc-like capped tower in M with iterated capped surface at least 3, and let $\epsilon > 0$. Then, the body of T extends to a 2-story capped tower whose top story has arbitrary iterated capped surface height and components with diameter less than ϵ.

Proof. Using Theorem 4.8, we can raise height so that T has 3 stories and the second story has arbitrary height. The third story of T then kills π_1 for the second story, and the tower caps kill π_1 for the third story (and hence for a regular neighborhood of the third story). Thus, if we contract the third story of T, we are left with a 2-story tower such that the second story is π_1-null in the complement of the first. We now want to use π_1-nullity to shrink the second story of T.

To do this, we approximate the top story of T by a graph. Let E denote the second-story iterated capped surface of T. Start with a vertex on the first stage of each component E. Connect these vertices by embedded arcs to the attaching circles of each of the caps of the first stage of E. Inductively, continue this process to obtain a tree ending in vertices for each of the tower caps, and add in loops bounding the tower caps. Call this graph K. Then, K is connected by cones to each iterated capped surface cap of the first story, a regular neighborhood of T is also a regular neighborhood of the union of K, these cones, and the first story.

Now, using the fact that the second story of T is π_1-null in the complement of the first story, we may homotope K so that the components end up in disjoint balls of radius less than ϵ. Then, this homotopy may be replaced with a homotopy through embeddings, which pulls the cones along and extends to an isotopy of a regular neighborhood of T, such that the components of the second story of T end up in the disjoint balls. □

Now, we are ready to consider infinite towers.

Definition 4.10. An abstract *infinite tower* T_∞ is the direct limit of a sequence of inclusions

$$T_1 \hookrightarrow T_2 \hookrightarrow T_3 \hookrightarrow \cdots$$

where T_n is an abstract n-story tower, and T_{n+1} is obtained by replacing the tower caps of T_n with 1-story towers.

As mentioned before, we want infinite towers in M to "converge" in some sense. In particular, they should be embedded (so no space-filling garbage), bounded (as we ultimately want the closure to be a 2-handle, which is compact) and the set of limit points should be nice (in general it will be a Cantor set). We consider convergence in the sense of the following theorem:

Theorem 4.11 (Tower convergence)**.** *Let E be a properly immersed iterated capped surface of height at least 3. Then, the body of E extends to an embedding of an infinite tower T_∞ such that there exists a neighborhood N of T_∞ whose closure is homeomorphic to the endpoint compactification of a regular neighborhood of T_∞.*

Proof Sketch. First, extend E to a 1-story capped tower T_1 with arbitrary iterated capped surface height in the first story. Let N_1 be a regular neighborhood of T_1. Inductively, suppose we have defined T_n and N_n. Then, extend T_n to an $(n+1)$-story capped tower T_{n+1}, and use Lemma 4.9 to squeeze the components of the top story into disjoint balls of radius less than $1/n$. Let N_{n+1} denote the interior of a regular neighborhood of the top story, such that N_{n+1} is contained in these disjoint balls. Take $T_\infty = \bigcup_n T_n$ and $N = \bigcup_n N_n$. Then, the desired properties hold. □

Such a neighborhood N is called a *pinched regular neighborhood* of T_∞. In the next section, we will see the punchline: for a good choice of tower T_∞ and neighborhood N, the pair $(\overline{N}, \partial^+\overline{N})$ is homeomorphic to a standard 2-handle, where $\partial^+\overline{N}$ is the portion of $\overline{N}$ comprising a closed regular neighborhood of the boundary ∂T_∞.

5. Obtaining a 2-Handle

Our goal in this section is to outline the proof that for N a pinched regular neighborhood of an embedded infinite tower T_∞, there exists a new tower $T'_\infty \subseteq T$ and neighborhood $N' \subseteq N$ such that we have a homeomorphism of pairs

$$(\overline{N}', \partial^+\overline{N}') \cong (D^2 \times D^2, D^2 \times S^1)$$

which on the boundary agrees with the standard identification of $\partial^+\overline{N}'$ and $D^2 \times S^1$ coming from the abstract model for an infinite tower. This homeomorphism then readily implies the disc embedding theorem: the results of Section 4 above imply that under the hypotheses of Theorem 1.1, an immersed disc can be replaced by a convergent infinite tower with the same boundary, and replacing a pinched regular neighborhood with a 2-handle gives the desired embedded disc as the co-core $\{0\} \times D^2$ of said handle. Greater detail on the following can be found in Chapter 4 of [1].

To this end, we begin by developing an alternate view of regular neighborhoods of towers which will lead to an alternate characterization of N. We start with the simpler case of a disc-like capped surface with a single pair of caps. Since everything we deal with is ultimately embedded, it suffices to work in the abstract model. Let $S \subseteq D^2 \times D^2$ be the model for an abstract disc-like capped surface. We can take discs A and B, each disjoint from the body of S and each other, and each intersecting one of the caps transversely in a single point. Then, a regular neighborhood of the body of S is obtained by deleting closed regular neighborhoods of A and B. The 3-dimensional situation is show in Figure 11. Such discs A and B, as well as the 2-handles $A \times D^2$ and $B \times D^2$, are called *dual to the caps*.

Similarly, for an n-story model capped tower T_n, a regular neighborhood of T_n is obtained by deleting a collection of 2-handles H_n dual to the tower caps. Now,

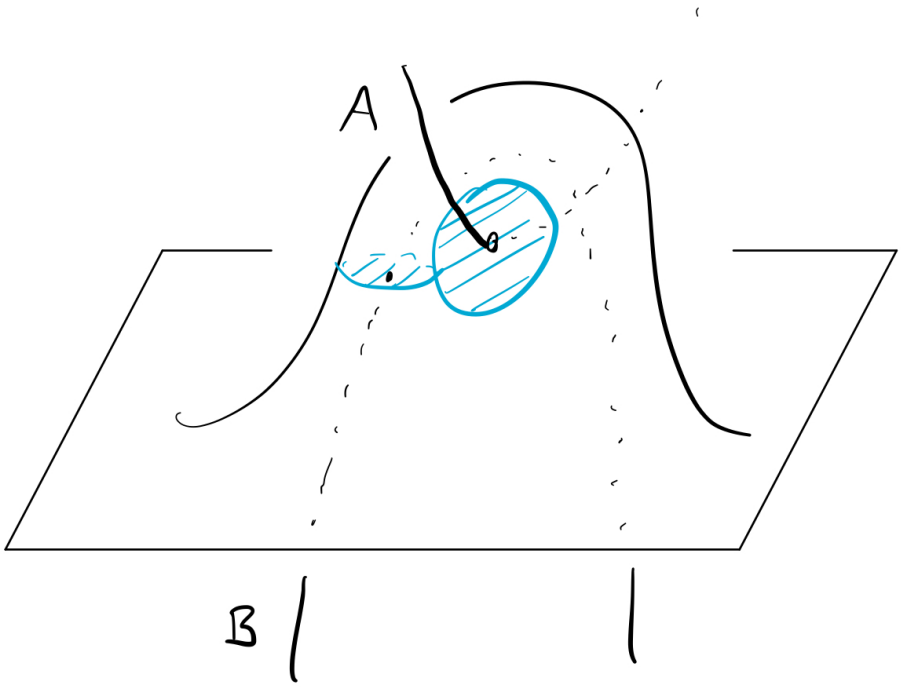

FIGURE 11. A capped surface with discs A and B dual to the caps.

suppose we add another story to obtain T_{n+1}. Then, we can arrange such that the top story is contained in the dual 2-handles, and that the dual handles H_{n+1} are contained in H_n, with $\partial^+ H_{n+1} \subseteq \partial^+ H_n$. A regular neighborhood of T_∞ is then the complement of $\bigcap_n H_n$.

Now, it turns out that in the boundary $\partial^+ H_n \cong \coprod D^2 \times S^1$, the inclusion $\partial^+ H_{n+1} \hookrightarrow \partial^+ H_n$ corresponds to the *Bing doubling* operation, which replaces a solid torus with two linked solid tori as in Figure 12, where the pairs of tori come from pairs of caps.

Using this fact, one can show that by isotopy, towers with sufficiently tall upper stories can be made such that the dual 2-handles lie in disjoint ϵ-balls balls centered at points of $S^1 \times \{0\} \subseteq \partial^+ N$. Then, one can prove the following:

Proposition 5.1. *There exists a pinched regular neighborhood $\overline{N}' \subseteq \overline{N}$ of an infinite tower with the same boundary as T_∞, such that $\overline{N}'$ is a quotient of $D^2 \times D^2$ by a collar neighborhood of $D^2 \times S^1$.*

One can then show that for $\overline{N}'$ as above, there exist a compact metric space Q and surjective maps $\alpha : D^2 \times D^2 \to Q, \beta : N' \to Q$ satisfying

(1) For both α and β, the collection of point inverses form a null decomposition of their respective domains, and the sets of points with nontrivial inverses are nowhere dense in Q.

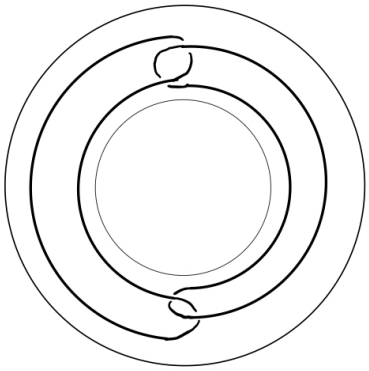

FIGURE 12. The Bing double of a solid torus

(2) For α, every nontrivial point inverse is either a ball or the union of an embedded D^2 and a regular neighborhood of its boundary circle.

Here, a decomposition of a space X means a collection $\mathcal{D}$ of disjoint subspaces whose union is X, and $\mathcal{D}$ is called null if for every $\epsilon > 0$, there are only finitely many elements of $\mathcal{D}$ with diameter greater than ϵ. Lastly, via some rather technical point-set topology, it can be shown that there exist homeomorphisms $\phi : D^2 \times D^2 \to Q$ and $\psi : N' \to Q$ agreeing with α, β respectively on the boundary. In the end, $\psi^{-1} \circ \phi$ gives the desired homeomorphism of N' with a standard 2-handle.

The general theory here is the theory of decompositions, and in particular there are conditions under which a quotient $X \to X/\mathcal{D}$ of a space by a decomposition can be shown to be *approximable by homeomorphisms*, a technical condition which, among other things, implies the existence of a homeomorphism. An outline of the proof using general techniques is contained in Sections 2 and 4 of [2].

To conclude, we explain a bit more about how the disc embedding theorem can be used to prove the 4-dimensional Poincaré conjecture. We have already seen the role of disc embeddings in proving the h-cobordism theorem. An important difference in 4 dimensions, however, is that since the disc embedding theorem only gives a topological embedding, we must do away with smooth techniques. In particular, if we want to use the disc embedding theorem then we no longer have access to Morse theory, and must instead do things in terms of (topological) handlebody decompositions. One then has to prove that the theory of handlebodies has cancellation theorems analogous to those in Morse theory (see [2] Section 1). The key point, however, is essentially the same: use Whitney discs to remove intersections of spheres (the attachinng spheres of handles), and we arrive at Theorem 2.4. Lastly, one has to show that two simply-connected, closed, homotopy equivalent 4-manifolds are actually h-cobordant (Theorem 4.9 in [2]), at which point the h-cobordism theorem readily implies the 4-dimensional Poincaré conjecture.

ACKNOWLEDGMENTS

I would like to thank Ao Sun, my mentor, for his enthusiasm and his help in this project. I would also like to thank Peter May for all the work he has done in making this REU happen.

References

[1] Michael H. Freedman and Frank Quinn. The Topology of 4-Manifolds. Princeton University Press. 1990.

[2] Danny Calegari. The 4-Dimensional Poincaré Conjecture. Online, 2019. https://math.uchicago.edu/~dannyc/courses/poincare_2018/4d_poincare_conjecture_notes.pdf

[3] Laurent Siebenmann and Jonathan Sondow. Lectures on the h-Cobordism Theorem by John Milnor. Princeton University Press. 1965.