

# Computer-Assisted Proofs in Geometry and Physics

by

Gregory T. Minton

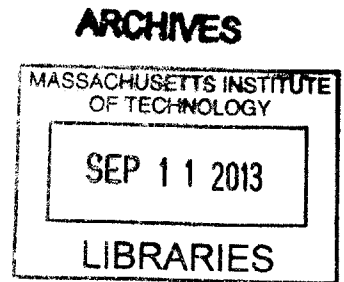
Submitted to the Department of Mathematics  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

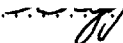
MASSACHUSETTS INSTITUTE OF TECHNOLOGY


September 2013



© Gregory T. Minton, MMXIII. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly  
paper and electronic copies of this thesis document in whole or in part in any medium  
now known or hereafter created.

Author .....  .....  
Department of Mathematics  
August 16, 2013

Certified by .....  .....  
Abhinav Kumar  
Associate Professor of Mathematics  
Thesis Supervisor

Accepted by .....  
Paul Seidel  
Co-Chairman, Department Graduate Committee



# Computer-Assisted Proofs in Geometry and Physics

by  
Gregory T. Minton

Submitted to the Department of Mathematics  
on August 16, 2013, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

## Abstract

In this dissertation we apply computer-assisted proof techniques to two problems, one in discrete geometry and one in celestial mechanics. Our main tool is an effective inverse function theorem which shows that, in favorable conditions, the existence of an approximate solution to a system of equations implies the existence of an exact solution nearby. This allows us to leverage approximate computational techniques for finding solutions into rigorous computational techniques for proving the existence of solutions.

Our first application is to tight codes in compact spaces, i.e., optimal codes whose optimality follows from linear programming bounds. In particular, we show the existence of many hitherto unknown tight regular simplices in quaternionic projective spaces and in the octonionic projective plane. We also consider regular simplices in real Grassmannians.

The second application is to gravitational choreographies, i.e., periodic trajectories of point particles under Newtonian gravity such that all of the particles follow the same curve. Many numerical examples of choreographies, but few existence proofs, were previously known. We present a method for computer-assisted proof of existence and demonstrate its effectiveness by applying it to a wide-ranging set of choreographies.

Thesis Supervisor: Abhinav Kumar

Title: Associate Professor of Mathematics



## Acknowledgments

First and foremost, clichéd though it may be, I would like to thank my parents: they were always around, through every up and every down. I also thank my MIT advisor, Abhinav Kumar, and my Microsoft Research mentor, Henry Cohn. They encouraged and enabled me to pursue complete awesomeness at all times. Along the way I have benefited greatly from the friendship of my officemates and fellow graduate students — especially Peter, John, and Sam, who probably could have written two theses each if I had stopped distracting them — as well as my friends outside the department, especially Scott, Nora, Jason, and Lisa; thank you all. I also want to acknowledge Sara, who had a big influence on this thesis.

Finally, this is for James, who would have done it all better.

---

I also thank the Fannie and John Hertz Foundation for, together with a National Science Foundation Graduate Research Fellowship, supporting my graduate studies.



# Contents

<b>I</b>	<b>Introduction</b>	<b>9</b>
1	Computer-Assisted Proof . . . . .	9
2	An Effective Existence Theorem . . . . .	9
2.1	Related existence theorems . . . . .	13
3	Computational Considerations . . . . .	13
3.1	Choice of norm . . . . .	13
3.2	Interval arithmetic . . . . .	14
<b>II</b>	<b>Optimal simplices and codes in projective spaces</b>	<b>17</b>
1	Introduction . . . . .	17
2	Codes in Projective Spaces . . . . .	18
2.1	Projective spaces over $\mathbb{R}$ , $\mathbb{C}$ , $\mathbb{H}$ and $\mathbb{O}$ . . . . .	18
2.2	Tight simplices . . . . .	20
2.3	Linear programming bounds . . . . .	21
2.4	Tight codes in $\mathbb{R}\mathbb{P}^{m-1}$ . . . . .	25
2.5	Tight codes in $\mathbb{C}\mathbb{P}^{m-1}$ . . . . .	26
2.6	Tight codes in $\mathbb{H}\mathbb{P}^{d-1}$ and $\mathbb{O}\mathbb{P}^2$ . . . . .	27
2.7	Gale duality . . . . .	27
3	Simplices in Quaternionic Projective Spaces . . . . .	29
3.1	Generic case . . . . .	29
3.2	12- and 13-point simplices . . . . .	33
3.3	15-point simplices . . . . .	35
4	Simplices in $\mathbb{O}\mathbb{P}^2$ . . . . .	38
4.1	Generic case . . . . .	38
4.2	24- and 25-point simplices . . . . .	40
4.3	27-point simplices . . . . .	41
5	Simplices in Grassmannians $G(m, n, \mathbb{R})$ . . . . .	42
5.1	Miscellaneous special cases in Grassmannians . . . . .	45
6	Algorithms and Computational Methods . . . . .	48
6.1	Rigorous proof . . . . .	48
6.2	Finding approximate solutions . . . . .	49
6.3	Finding stabilizers . . . . .	50
6.4	Real algebraic numbers . . . . .	51
6.5	Estimating dimensions . . . . .	51
7	Explicit Constructions . . . . .	52
7.1	Two universal optima in $\text{SO}(4)$ . . . . .	52
7.2	39 points in $\mathbb{O}\mathbb{P}^2$ . . . . .	54

<b>III</b>	<b>Gravitational Choreographies</b>	<b>61</b>
1	Background . . . . .	61
1.1	Periodic orbits . . . . .	62
1.2	Choreographies . . . . .	62
1.3	Figure-eight . . . . .	63
1.4	Variational proofs . . . . .	64
1.5	More choreographies . . . . .	67
1.6	Non-variational approaches . . . . .	69
2	Formal Problem Statement . . . . .	71
2.1	Physics . . . . .	71
2.2	Conserved quantities . . . . .	72
2.3	Normalization . . . . .	72
2.4	Fourier series representation . . . . .	73
2.5	The action and its gradient . . . . .	75
3	Our Results . . . . .	75
3.1	Proving existence of choreographies . . . . .	76
3.2	Finding choreographies numerically . . . . .	77
3.3	Saddle points of the action . . . . .	79
3.4	Identifying real orbits . . . . .	80
3.5	Symmetry . . . . .	81
3.6	Stability . . . . .	84
4	Our Proof Technique . . . . .	85
4.1	Action principle . . . . .	86
4.2	Expressions for the gradient and Hessian . . . . .	90
4.3	Computing bounds on functions and their Fourier coefficients . . . . .	92
4.4	Step 1: bounding the gradient . . . . .	95
4.5	Step 2: accounting for symmetries . . . . .	95
4.6	Step 3: bounding the Hessian . . . . .	96
4.7	Step 4: bounding the change in the Hessian . . . . .	97
4.8	Final bounds . . . . .	99
5	Implementation Details . . . . .	100
5.1	Optimizations in Hessian computations . . . . .	100
5.2	Dropping variables . . . . .	101
5.3	Computing bounds on Fourier coefficients . . . . .	103
5.4	Faster matrix multiplication . . . . .	104
6	Comparison With Previous Work . . . . .	105
6.1	Kapela et al.'s computer-assisted proofs . . . . .	105
6.2	Arioli et al.'s computer-assisted proofs . . . . .	106
6.3	Treatment of symmetries . . . . .	107
7	Further Work . . . . .	107
8	Gravitational Gallery . . . . .	111



# Chapter I

## Introduction

### 1 Computer-Assisted Proof

The role of computational results in mathematics is well-established, but its role in rigorous proof is relatively young. The first proof of the four color theorem, given in 1976 by Appel and Haken [2], is the break-out example of a proof involving computations so extensive that they cannot be checked by a human. Since then there have been many more theorems proven using extensive and essential computer calculations; two particularly noteworthy examples are the Kepler conjecture, now Hales' theorem [6], and the existence of the Lorenz attractor [19]. Even the solution of checkers [17] can be thought of as a theorem of this sort.

These are examples of *computer-assisted proofs*. Our work gives two applications of computer-assisted proof techniques, one in geometry (Chapter II) and one in physics (Chapter III). Of the four examples just given, our applications have more in common with the Kepler conjecture and the Lorenz attractor, as opposed to the four color theorem and the solution of checkers. For the latter pair, the problem is essentially *discrete*; once made finite, it is clear that a computer could address the problem, at least in principle. By contrast, the former pair, and our work, are *continuous* problems. It is perhaps less intuitive that computational results can resolve such problems.

Note that we distinguish computer-*verified* proofs from computer-*assisted* proofs, with our distinction being that the former involves the formal verification of proofs, whereas the latter is concerned with proving new theorems. The focus here is on new theorems, and in particular on theorems for which we know of no plausible approach avoiding the invocation of electronic computers.

Both of our applications concern the existence of objects with certain properties. Moreover, in both settings there is no particular reason to suppose that these objects have a concise, explicit representation (say, in terms of low-degree algebraic numbers). But, as we shall formalize later, there is no particular reason why they should not exist; they can be viewed as solutions to systems with at least as many variables as constraints. Based on these observations, we propose that rather than existing for a “nice” reason, which could be succinctly analyzed by hand, they simply exist because ... why not?

### 2 An Effective Existence Theorem

The main tool for all of our computer-assisted proofs is Theorem 2.1 below. It shows that, under suitable conditions, wherever there is an approximate solution to a system of equations

there must be a true solution nearby. This result fits into a body of related work (see §2.1). It is possible that this specific formulation and proof may not have previously appeared in the literature, but we make no assertions of originality.

The theorem is stated for general Banach spaces. We use standard notation:  $|\cdot|$  denotes the norm on a Banach space,  $\|\cdot\|$  denotes the operator norm,  $Df(x)$  is the Fréchet derivative of  $f$  at  $x$ ,  $B(x_0, \varepsilon)$  is the open ball around  $x_0$  with radius  $\varepsilon$ , and  $\text{id}_W$  denotes the identity operator on  $W$ .

**Theorem 2.1.** *Let  $V$  and  $W$  be real Banach spaces. Given  $x_0 \in V$  and  $\varepsilon > 0$ , suppose that  $f: B(x_0, \varepsilon) \rightarrow W$  is Fréchet differentiable. Suppose also that  $T: W \rightarrow V$  is a bounded linear operator such that*

$$\|Df(x) \circ T - \text{id}_W\| < 1 - \frac{\|T\| \cdot |f(x_0)|}{\varepsilon} \quad (2.1)$$

for all  $x \in B(x_0, \varepsilon)$ . Then there exists  $x_* \in B(x_0, \varepsilon)$  such that  $f(x_*) = 0$ .

We will obtain this result as a straightforward generalization of the following effective inverse function theorem. For notational brevity we use  $B(r)$  and  $\overline{B}(r)$  to denote the open ball and closed ball, respectively, around  $0 \in W$  and of radius  $r$ .

**Proposition 2.2.** *Let  $W$  be a Banach space, fix  $r > 0$ , and suppose  $g: B(r) \rightarrow W$  is differentiable and satisfies  $g(0) = 0$ . Suppose also that, for some  $\gamma < 1$ ,*

$$\|Dg(x) - \text{id}_W\| \leq \gamma \quad \text{for all } x \in B(r). \quad (2.2)$$

Then  $g$  extends continuously to  $\overline{B}(r)$ , and there exists  $h: \overline{B}(r(1 - \gamma)) \rightarrow \overline{B}(r)$  such that  $g(h(y)) = y$  for all  $y \in \overline{B}(r(1 - \gamma))$ . Moreover, if the inequality in (2.2) is strict, then the image of  $h$  lies in  $B(r)$ .

*Proof.* Define  $c: B(r) \rightarrow W$  by

$$c(x) = g(x) - x.$$

The map  $c$  is differentiable and the operator norm of its derivative is everywhere bounded by  $\gamma$ . Thus the mean value inequality implies that, for any  $x_1, x_2 \in B(r)$ ,

$$|c(x_1) - c(x_2)| \leq \gamma \cdot |x_1 - x_2|. \quad (2.3)$$

In particular,  $c$  is Lipschitz, so uniformly continuous. The same is true of  $g(x) = c(x) + x$ . As the codomain  $W$  is complete,  $c$  and  $g$  extend continuously to  $\overline{B}(r)$ . Note that the Lipschitz bound (2.3) is obeyed on the entire closed ball.

Let  $y \in \overline{B}(r(1 - \gamma))$  be arbitrary and define the function

$$u(x) = u_y(x) = x + (y - g(x)) = y - c(x)$$

on  $\overline{B}(r)$ . Applying (2.3) and the triangle inequality,

$$|u(x)| \leq |y| + |c(x)| \leq |y| + \gamma \cdot |x| \leq r(1 - \gamma) + \gamma r = r,$$

so in fact  $u$  maps  $\overline{B}(r)$  to itself. Moreover, for any  $x_1, x_2 \in \overline{B}(r)$ ,

$$|u(x_1) - u(x_2)| = |c(x_1) - c(x_2)| \leq \gamma \cdot |x_1 - x_2|$$

by (2.3). Thus  $u$  is a contraction.

By the Banach contraction mapping principle,  $u$  has a (unique) fixed point; but, by definition of  $u_y$ , a fixed point is exactly a preimage of  $y$  under  $g$ . Thus, letting  $h(y)$  be the fixed point of  $u = u_y$ , we have the desired inverse function.

Now suppose that the inequality in (2.2) is (everywhere) strict. Then  $|c(x)| < \gamma \cdot |x|$  for any  $x \neq 0$ , so the argument we gave to show  $u(\overline{B}(r)) \subset \overline{B}(r)$  actually proves  $u(\overline{B}(r)) \subset B(r)$ . The fixed point of  $u$  is in its image, whence the final claim follows.  $\square$

The proof just given is cribbed from a standard proof of the inverse function theorem. We note in passing that the function  $h$  in the proposition statement is actually unique, continuous, and differentiable on  $B(r)$ .

*Proof of Theorem 2.1.* Let  $g(x) = f(x_0 + Tx) - f(x_0)$ , define  $\gamma = 1 - \|T\| \cdot |f(x_0)|/\varepsilon$ , and set  $r = \varepsilon/\|T\|$ . We have  $\gamma \leq 1$ , and (2.1) implies  $\gamma > 0$ . The map  $T$  cannot be zero because  $\|\text{id}_W\| = 1$ ; thus  $\gamma = 1$  iff  $f(x_0) = 0$ , but in this case the theorem statement is trivial. If instead  $\gamma \in (0, 1)$ , then apply Proposition 2.2 and take  $x_* = x_0 + T(h(-f(x_0)))$ .  $\square$

In this thesis we will apply Theorem 2.1 in two different settings.

In Chapter II we apply it to finite-dimensional normed vector spaces. In this setting  $Df(x)$  is simply the Jacobian of  $f$  at  $x$ , and (given suitable smoothness conditions) a successful application of the theorem also yields the dimension of the solution set. Moreover, a weaker assumption suffices.

**Proposition 2.3.** *Let  $V$  and  $W$  be real Banach spaces. Given  $x_0 \in V$  and  $\varepsilon > 0$ , suppose that  $f: B(x_0, \varepsilon) \rightarrow W$  is  $C^1$ . Suppose also that  $T: W \rightarrow V$  is a bounded linear operator such that, for all  $x \in B(x_0, \varepsilon)$ ,*

$$\|Df(x) \circ T - \text{id}_W\| < 1. \quad (2.4)$$

*If the cokernel of  $T$  is finite-dimensional, then the zero locus  $f^{-1}(0) \subset B(x_0, \varepsilon)$  is a  $C^1$  manifold of dimension  $\dim \text{coker } T$ . In particular, if  $V$  and  $W$  are finite-dimensional, then  $f^{-1}(0)$  is a  $C^1$  manifold of dimension  $\dim V - \dim W$ .*

*Proof.* This is basically a corollary of the preimage theorem in differential geometry, but that theorem is not usually stated in Banach spaces and so we shall give a few details. Note that (2.4) implies that, for all  $x \in B(x_0, \varepsilon)$ ,  $Df(x) \circ T$  is invertible; this is because the power series  $\sum_{k \geq 0} (\text{id}_W - Df(x) \circ T)^k$  converges, and that series yields the inverse. Fix a finite-dimensional subspace  $F$  of  $V$  lifting  $\text{coker } T = V/T(W)$ . Let  $x_1 \in f^{-1}(0) \subset B(x_0, \varepsilon)$  be arbitrary and write  $x_1 = T(y_1) + z_1$  with  $y_1 \in W$  and  $z_1 \in F$ . Choose neighborhoods  $U_1$  of  $y_1 \in W$  and  $U_2$  of  $z_1 \in F$  such that  $T(U_1) + U_2 \subset B(x_0, \varepsilon)$ . Then the function  $q: U_1 \times U_2 \rightarrow W$  defined by  $q(y, z) = f(T(y) + z)$  is  $C^1$  and the restriction of  $Dq(y_1, z_1)$  to  $W$  is  $Df(x_1) \circ T$ , which is invertible. Thus the implicit function theorem applies to  $q$ . This identifies a neighborhood of  $x_1 \in f^{-1}(0)$  with a neighborhood of  $z_1 \in F \cong \mathbb{R}^{\dim \text{coker } T}$ , which proves the first statement. For the second statement, note that invertibility of  $Df(x_0) \circ T$  implies that  $T$  is injective, and in the finite-dimensional case that means that  $\dim \text{coker } T = \dim V - \dim W$ .  $\square$

If  $f$  satisfies stronger smoothness conditions, then the manifold  $f^{-1}(0)$  inherits the same properties. In Chapter II we will actually always use  $C^\infty$  functions, so the zero sets will be smooth manifolds.

In Chapter III, by contrast, we apply Theorem 2.1 in a case where we expect  $f$  to be a local isomorphism. In this case it is of interest to have an effective uniqueness result. We will not actually apply this result in the present document, but we enunciate it for the record.

**Proposition 2.4.** *Let  $V$  and  $W$  be real Banach spaces and let  $B \subset V$  be an open, convex set. Let  $f: B \rightarrow W$  be a  $C^1$  function, and let  $T: W \rightarrow V$  be a bounded linear operator. If*

$$\|T \circ Df(x) - \text{id}_V\| < 1 \tag{2.5}$$

*for all  $x \in B$ , then  $f$  is injective on  $B$ .*

*Proof.* Let  $x \neq y \in B$  be arbitrary. Applied to the function  $T \circ f - \text{id}_V$ , the mean value inequality asserts that

$$\|(T \circ f)(x) - (T \circ f)(y) - (x - y)\| \leq \|T \circ Df(z) - \text{id}_V\| \cdot |x - y|$$

for some  $z \in B$  (in fact, for some  $z$  on the line segment between  $x$  and  $y$ ). Thus

$$\|(T \circ f)(y) - (T \circ f)(x)\| \geq (1 - \|T \circ Df(x) - \text{id}_V\|) \cdot |x - y| > 0.$$

This proves that  $T \circ f$  is injective, so  $f$  is injective *a fortiori*. □

In the finite-dimensional setting, it is straightforward to apply Theorem 2.1. Our task is to compute an *approximate* right inverse  $T$  of  $Df(x_0)$  and bound  $\|Df(x) \circ T - \text{id}_W\|$  for all  $x \in B(x_0, \varepsilon)$ . The first step is no obstacle (see the next paragraph), and the second step can be done simply and elegantly using interval arithmetic (see §3.2). The difficulty in applying the theorem lies in identifying which functions  $f$  to use; that is the main challenge we face in Chapter II. By contrast, in infinite-dimensional spaces, even the computational application of Theorem 2.1 is more delicate (see §III.4).

As we shall see, the freedom to choose any matrix  $T$  which is suitably close to an approximate right inverse is immensely useful. It lets us compute  $T$  via non-rigorous methods, which are much faster than, e.g., interval arithmetic calculations (see §II.6.1 and §III.5.1).

*Remark.* While the conclusion of Theorem 2.1 strengthens as  $\varepsilon$  decreases, the hypotheses do not change monotonically. The bound we need to prove for  $\|Df(x) \circ T - \text{id}_W\|$  becomes stronger, but the domain on which we need to prove it shrinks. Thus it can, and sometimes does, happen that we can verify the hypotheses of Theorem 2.1 only by choosing a smaller  $\varepsilon$ .

*Aside.* We proved Theorem 2.1 using a fixed-point theorem, and indeed in the literature existence results of this form tend to be proven this way. Fixed-point theorems are powerful and general. There is an alternative perspective, though, which we think is more intuitive and explanatory: in the context of Theorem 2.1, if  $x(t)$  is a solution of the initial-value problem

$$x'(t) = -T(Df(x(t)) \circ T)^{-1}f(x(t)), \quad x(0) = x_0,$$

then  $x(1)$  satisfies  $f(x(1)) = 0$ . This differential equation is a time-rescaled continuous analog of Newton's method, and its application to existence theorems is known [14]. If  $f$  is  $C^1$  and  $W$  is finite-dimensional, then one can use the Peano existence theorem to find a solution to the differential equation. In the infinite-dimensional case, if  $f$  is  $C^1$  and its derivative is locally Lipschitz, then the Picard existence theorem gives a solution. However, our proof avoids the need for these smoothness assumptions.

## 2.1 Related existence theorems

Theorem 2.1 is effectively an effective inverse function theorem, since it can be viewed as giving quantitative bounds on a neighborhood in which  $f$  has a (right) inverse. Such results date back to the Newton-Kantorovich theorem [8], which gives quantitative bounds on the convergence of Newton’s method; see Ortega’s short note [15] for an English-language proof and Moore and Cloud’s textbook [12] for a development from the computational perspective. Kantorovich’s result is close in spirit to our theorem. The main difference is that he analyzed Newton’s method itself, whereas we are interested only in existence and so accept a degree of abstractness in exchange for a cleaner statement and proof.

Kantorovich’s theorem shares with Theorem 2.1 the property of immediate generalization to Banach spaces. It lacks, though, the freedom we have to use an approximate right inverse  $T$ . Instead Kantorovich’s theorem restricts to invertible functions and uses the actual inverse of  $Df$ . It also requires that the derivative be Lipschitz. While this is often a natural condition to check computationally (as in §III.4.7), sometimes it is more convenient to instead verify (2.1) directly (as in §II.6.1).

There are other existence theorems in the literature that are closer to ours in terms of concrete computations. One such is the Krawczyk-Moore theorem [9, 11]. (The literature tends to refer to the theorem by one of those two names, but not both. Krawczyk presented an analysis of convergence but assumed the existence of a solution, while Moore noted that existence could be deduced rather than assumed.) This theorem is very similar to ours, and in fact it is stronger than ours when restricted to the finite-dimensional setting with the supremum norm. Again, the primary difference is that we have exchanged neatness for generality. The Krawczyk-Moore theorem shares with Theorem 2.1 the flexibility to choose an approximate inverse. It also generalizes to infinite-dimensional settings, although not as easily [5]. Other results in a similar vein include the existence theorems of Miranda [10] and Borsuk [3] and Smale’s analysis of Newton’s method [18].

## 3 Computational Considerations

### 3.1 Choice of norm

Theorem 2.1 applies to any Banach space, so (especially in the finite-dimensional setting) we are free to choose the norm best suited to our problem. We could even choose substantially different norms on the domain and codomain, although in our work we have not exploited this particular freedom.

In Chapter II we use the  $\ell^\infty$  norm and in Chapter III we use the  $\ell^1$  norm. These are both convenient choices computationally because, in addition to being easy to compute the norm of a vector (which is true of any  $\ell^p$  norm), it is easy to compute  $\ell^p \rightarrow \ell^p$  operator norms when  $p = 1, \infty$ . In particular, the  $\ell^\infty \rightarrow \ell^\infty$  operator norm of a matrix is the maximum of the  $\ell^1$  norms of its rows, and the  $\ell^1 \rightarrow \ell^1$  operator norm is the maximum of the  $\ell^1$  norms of its columns. The first statement is basically obvious and the second follows from convexity (or duality).

Using the  $\ell^2$  norm would also be an acceptable choice, as the  $\ell^2 \rightarrow \ell^2$  operator norm can be approximated efficiently. (It is the largest singular value of the matrix.) However, for any other choice of  $\ell^p$ , even approximating the operator norm is NP-hard [7]. Indeed, computing the  $\ell^\infty \rightarrow \ell^1$  operator norm of a matrix is already a difficult problem [1].

## 3.2 Interval arithmetic

Interval arithmetic is a standard tool in numerical analysis to control the errors inherent in floating-point computations [13]. The principle is simple: instead of rounding the results of arithmetic operations to a number that can be exactly represented in a computer, we work with intervals of representable numbers that are guaranteed to contain the correct value. This lets us offload any concerns about numerical round-off error onto the computer; it automatically tracks the errors for us.

For instance, consider a hypothetical computer capable of storing 4 decimal digits of precision. Using floating-point arithmetic,  $\pi$  would best be represented as 3.142. Using this, if we computed  $2 \cdot \pi$  then we would get 6.284, which is obviously not correct. By contrast, interval arithmetic on the same computer would represent  $\pi$  as the interval  $[3.141, 3.142]$ . Then  $2 \cdot \pi$  would be represented by the interval  $[6.282, 6.284]$ , which does contain the exact value.

In our software we use the MPFI library to provide support for interval arithmetic [16]. That in turn relies on MPFR, a library for multiple-precision floating-point arithmetic [4]. One of the main problems with interval arithmetic is that the size of the intervals can grow exponentially with the number of arithmetic operations; this problem can be ameliorated by increasing the precision of the underlying floating-point numbers. Thus it is a major advantage of MPFI and MPFR that the precision can be increased arbitrarily.

This did not turn out to be a significant issue in our applications, though. We used 128-bit floating-point numbers for the intervals in our rigorous proofs, and that proved to be sufficient in all cases save three (see Figure III.3.1). We did not explore whether less precision would have sufficed; presumably such optimization could yield some (possibly significant) improvements in runtime.

# Bibliography

- [1] Noga Alon and Assaf Naor, *Approximating the cut-norm via Grothendieck's inequality*, SIAM J. Comput. **35** (2006), no. 4, 787–803 (electronic). MR 2203567 (2006k:68176)
- [2] Kenneth Appel and Wolfgang Haken, *Every planar map is four colorable*, Contemporary Mathematics, vol. 98, American Mathematical Society, Providence, RI, 1989, With the collaboration of J. Koch. MR 1025335 (91m:05079)
- [3] Karol Borsuk, *Drei sätze über die  $n$ -dimensionale euklidische Sphäre*, Fund. Math. **20** (1933), no. 1, 177–190 (German).
- [4] Laurent Fousse, Guillaume Hanrot, Vincent Lefèvre, Patrick Pélicier, and Paul Zimmermann, *MPFR: A multiple-precision binary floating-point library with correct rounding*, ACM Trans. Math. Softw. **33** (2007), no. 2.
- [5] Zbigniew Galias and Piotr Zgliczyński, *Infinite-dimensional Krawczyk operator for finding periodic orbits of discrete dynamical systems*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **17** (2007), no. 12, 4261–4272. MR 2394226 (2009i:37222)
- [6] Thomas C. Hales, *A proof of the Kepler conjecture*, Ann. of Math. (2) **162** (2005), no. 3, 1065–1185. MR 2179728 (2006g:52029)
- [7] Julien M. Hendrickx and Alex Olshevsky, *Matrix  $p$ -norms are NP-hard to approximate if  $p \neq 1, 2, \infty$* , SIAM J. Matrix Anal. Appl. **31** (2010), no. 5, 2802–2812. MR 2740634 (2011j:15035)
- [8] Leonid Vitaliyevich Kantorovich, *On Newton's method for functional equations*, Doklady Akad. Nauk SSSR **59** (1948), 1237–1240 (Russian).
- [9] R. Krawczyk, *Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken.*, Computing (Arch. Elektron. Rechnen) **4** (1969), 187–201. MR 0255046 (40 #8253)
- [10] Carlo Miranda, *Un'osservazione su un teorema di Brouwer*, Boll. Un. Mat. Ital. (2) **3** (1940), 5–7. MR 0004775 (3,60b)
- [11] Ramon E. Moore, *A test for existence of solutions to nonlinear systems*, SIAM J. Numer. Anal. **14** (1977), no. 4, 611–615. MR 0657002 (58 #31801)
- [12] Ramon E. Moore and Michael J. Cloud, *Computational functional analysis*, 2nd ed., Ellis Horwood Series: Mathematics and its Applications, Woodhead Publishing, 2007.
- [13] Ramon E. Moore, R. Baker Kearfott, and Michael J. Cloud, *Introduction to interval analysis*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009. MR 2482682 (2010d:65106)
- [14] J. W. Neuberger, *The continuous Newton's method, inverse functions, and Nash-Moser*, Amer. Math. Monthly **114** (2007), no. 5, 432–437. MR 2309983 (2008a:49037)

- [15] James M. Ortega, *The Newton-Kantorovich theorem*, Amer. Math. Monthly **75** (1968), 658–660. MR 0231218 (37 #6773)
- [16] Nathalie Revol and Fabrice Rouillier, *Motivations for an arbitrary precision interval arithmetic and the MPFI library*, Reliab. Comput. **11** (2005), no. 4, 275–290. MR 2158338 (2006e:65078)
- [17] Jonathan Schaeffer, *One jump ahead: computer perfection at checkers*, 2nd ed., Springer, 2008.
- [18] Steve Smale, *Newton's method estimates from data at one point*, The merging of disciplines: new directions in pure, applied, and computational mathematics (Laramie, Wyo., 1985), Springer, New York, 1986, pp. 185–196. MR 870648 (88e:65076)
- [19] Warwick Tucker, *A rigorous ODE solver and Smale's 14th problem*, Found. Comput. Math. **2** (2002), no. 1, 53–117. MR 1870856 (2003b:37055)



# Chapter II

## Optimal simplices and codes in projective spaces

*This chapter presents joint work with Henry Cohn and Abhinav Kumar. Most of it appears verbatim in a joint paper prepared for publication [15].*

### 1 Introduction

The study of codes in spaces such as spheres, projective spaces, and Grassmannians has been the focus of much interest recently, with an interplay of methods from many aspects of mathematics, physics, and computer science. Given a compact metric space  $X$ , the problem is how to arrange  $N$  points in  $X$  so as to maximize the minimal distance between them. A point configuration is called a *code*, and an *optimal code*  $\mathcal{C}$  maximizes the minimal distance between its points given its size  $|\mathcal{C}|$ . Finding optimal codes is a central problem in coding theory. Even when  $X$  is finite (for example, the cube  $\{0, 1\}^n$  under Hamming distance), this optimization problem is generally intractable, and it is even more difficult when  $X$  is infinite.

Most of the known optimality theorems have been proved using linear programming bounds, and we are especially interested in codes for which these bounds are sharp. We call them *tight* codes.<sup>1</sup> These cases include many of the most remarkable codes known, such as the icosahedron or the  $E_8$  root system.

In this paper, we explore the landscape of tight codes in projective spaces. We are especially interested in simplices of  $N$  points in  $d$ -dimensional projective space (i.e., collections of  $N$  equidistant points). Tight simplices correspond to tight equiangular frames [39], which have applications in signal processing and sparse approximation, and they also capture interesting invariants of their ambient spaces.

In real and complex projective spaces, tight simplices occur only sporadically. All known constructions are based on geometric, group-theoretic, or combinatorial properties that depend delicately on the size  $N$  and dimension  $d$ . By contrast, we find a surprising new phenomenon in quaternionic and octonionic spaces: in each dimension, there are substantial intervals of sizes for which tight simplices always seem to exist.

---

<sup>1</sup>The word “tight” is used for a related but more restrictive concept in the theory of designs. We use the same word here for lack of a good substitute. This makes “tight” a noncompositional adjective, much like “optimal”: codes and designs are both just sets of points, so every code is a design and vice versa, but a tight code is not necessarily a tight design. (However, one can show that every tight design is a tight code.)

This behavior cannot plausibly be explained using the sorts of constructions that work in real or complex spaces. Instead, the new tight simplices seem not to exist for any special reason, but rather simply because of parameter counting. Specifically, they can be characterized using more variables than constraints, in a way that suggests they should exist but does not prove it. We do not know how to prove that they exist in general, but we prove existence in many hitherto unknown cases. We also extend our methods to handle some exceptional cases that are more subtle.

Our results settle several open problems dating back to the early 1980s. We show the existence of a 15-point simplex in  $\mathbb{H}\mathbb{P}^2$  and a 27-point simplex in  $\mathbb{O}\mathbb{P}^2$ , which are not only optimal codes, but also the largest possible simplices in these spaces. (For comparison, the six diagonals of an icosahedron form a maximal simplex in  $\mathbb{R}\mathbb{P}^2$ , and the largest simplex in  $\mathbb{C}\mathbb{P}^2$  has size nine.) Furthermore, these simplices are tight 2-designs. We also construct a set of 13 mutually unbiased bases in  $\mathbb{O}\mathbb{P}^2$ . The mutually unbiased bases had been conjectured to exist [24, p. 35], but no construction was known, and the tight simplices were conjectured not to exist [23, p. 251]. It would be interesting to determine whether these simplices could lead to minimal triangulations of  $\mathbb{H}\mathbb{P}^2$  and  $\mathbb{O}\mathbb{P}^2$ , which would have the same number of vertices (see [10]).

We also rigorously show the existence of many tight simplices in real Grassmannians, which were conjectured to exist in [17] based on numerical evidence. This case is similar to quaternionic and octonionic projective spaces, in that parameter counts

In contrast to the usual algebraic methods for constructing tight codes, we take a rather different approach to show the existence of families of simplices. We use our general effective implicit function theorem (Theorem I.2.1), which allows us to show the existence of a real solution to a system of polynomial equations near an approximate solution. Furthermore, when this theorem applies, it also shows that the space of solutions is a smooth manifold near the approximate solution and tells us its dimension (see Proposition I.2.3). This allows us to establish many new results for which algebraic constructions seem out of reach.

In §2 we describe linear programming bounds and recall what is known about tight codes in projective spaces over  $\mathbb{R}$ ,  $\mathbb{C}$ ,  $\mathbb{H}$ , and  $\mathbb{O}$ . Our results concerning existence of new families of simplices, proved using our main existence theorem (Theorem I.2.1), are described in §3 and §4. In §5 we use our methods to produce several positive-dimensional families of simplices in real Grassmannians. We then give a discussion of the algorithms and computer programs used for these computer-assisted proofs in §6. Finally, we conclude in §7 with a few explicit constructions of universally optimal codes, the most notable of which is a maximal system of mutually unbiased bases in  $\mathbb{O}\mathbb{P}^2$ .

## 2 Codes in Projective Spaces

### 2.1 Projective spaces over $\mathbb{R}$ , $\mathbb{C}$ , $\mathbb{H}$ and $\mathbb{O}$

If  $K = \mathbb{R}$ ,  $\mathbb{C}$ , or  $\mathbb{H}$ , we let  $K\mathbb{P}^{d-1} := (K^d \setminus \{0\})/K^\times$  be the set of lines in  $K^d$ . That is, we identify  $x$  and  $x\alpha$  for  $x \in K^d \setminus \{0\}$  and  $\alpha \in K^\times$ . Note the convention that  $K^\times$  acts on the right; this is important for the noncommutative algebra  $\mathbb{H}$ .

We equip  $K^d$  with the Hermitian inner product  $\langle x_1, x_2 \rangle = x_1^\dagger x_2$ , where  $\dagger$  denotes the conjugate transpose. We may represent an element of the projective space by a unit-length vector  $x \in K^d$ , and we often abuse notation by treating the element itself as such a vector.

Under this identification, the *chordal distance* between two points of  $K\mathbb{P}^{d-1}$  is

$$\rho(x_1, x_2) = \sqrt{1 - |\langle x_1, x_2 \rangle|^2}.$$

It is not difficult to check that this formula defines a metric, which is equivalent to the Fubini-Study metric. Specifically, if  $\vartheta(x_1, x_2)$  is the geodesic distance on  $K\mathbb{P}^{d-1}$  under the Fubini-Study metric, normalized so that the greatest distance between two points is  $\pi$ , then

$$\cos \vartheta(x_1, x_2) = 2|\langle x_1, x_2 \rangle|^2 - 1$$

and

$$\rho(x_1, x_2) = \sin \left( \frac{\vartheta(x_1, x_2)}{2} \right).$$

Alternatively, elements  $x \in K\mathbb{P}^{d-1}$  correspond to projection matrices  $\Pi = xx^\dagger$ , which are Hermitian matrices with  $\Pi^2 = \Pi$  and  $\text{Tr } \Pi = 1$ . The space  $\mathcal{H}(K^d)$  of Hermitian matrices is a real vector space endowed with a positive-definite inner product

$$\langle A, B \rangle = \text{Tr} \frac{1}{2}(AB + BA) = \text{Re Tr } AB.$$

Because  $\text{Re } ab = \text{Re } ba$  for  $a, b \in K$ , it follows that  $\text{Re Tr}(ABC) = \text{Re Tr}(CAB)$  for  $A, B, C \in K^{d \times d}$ ; in other words, the operator  $\text{Re Tr}$  is *cyclic invariant*. Hence, for any  $x_1, x_2 \in K\mathbb{P}^{d-1}$  with associated projection matrices  $\Pi_1, \Pi_2 \in \mathcal{H}(K^d)$ , we have

$$\begin{aligned} \langle \Pi_1, \Pi_2 \rangle &= \text{Re Tr } x_1 x_1^\dagger x_2 x_2^\dagger \\ &= \text{Re Tr } x_2^\dagger x_1 x_1^\dagger x_2 \\ &= \text{Re } \langle x_2, x_1 \rangle \langle x_1, x_2 \rangle \\ &= |\langle x_1, x_2 \rangle|^2. \end{aligned} \tag{2.1}$$

Thus the metric on  $K\mathbb{P}^{d-1}$  can also be defined by  $\rho(x_1, x_2) = \sqrt{1 - \langle \Pi_1, \Pi_2 \rangle}$ . Equivalently, it equals  $\|\Pi_1 - \Pi_2\|/\sqrt{2}$ , where  $\|\cdot\|$  denotes the Frobenius norm:  $\|A\| = (\sum_{i,j} |A_{ij}|^2)^{1/2}$  for a matrix whose  $i, j$  entry is  $A_{ij}$ .

The one remaining projective space is the octonionic projective plane  $\mathbb{O}\mathbb{P}^2$ . (See [5] for an account of why  $\mathbb{O}\mathbb{P}^d$  cannot exist for  $d > 2$ ; one can construct  $\mathbb{O}\mathbb{P}^1$ , but we will ignore it as it is simply  $S^8$ .) Due to the failure of associativity, the construction of  $\mathbb{O}\mathbb{P}^2$  is more complicated than that of the other projective spaces; we cannot simply view it as the space of lines in  $\mathbb{O}^3$ . However, there is a construction analogous to that in the previous paragraph. The result is an exceptionally beautiful space that has been called the panda of geometry [8, p. 155]. The points of  $\mathbb{O}\mathbb{P}^2$  are  $3 \times 3$  projection matrices over  $\mathbb{O}$ , i.e.,  $3 \times 3$  Hermitian matrices  $\Pi$  satisfying  $\Pi^2 = \Pi$  and  $\text{Tr } \Pi = 1$ . The (chordal) metric in  $\mathbb{O}\mathbb{P}^2$  is given by

$$\rho(\Pi_1, \Pi_2) = \frac{1}{\sqrt{2}} \|\Pi_1 - \Pi_2\| = \sqrt{1 - \langle \Pi_1, \Pi_2 \rangle}.$$

Each projection matrix  $\Pi$  is of the form

$$\Pi = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \begin{pmatrix} \bar{a} & \bar{b} & \bar{c} \end{pmatrix},$$

where  $a, b, c \in \mathbb{O}$  satisfy  $|a|^2 + |b|^2 + |c|^2 = 1$  and  $(ab)c = a(bc)$ . We can cover  $\mathbb{O}\mathbb{P}^2$  by three affine charts as follows. Any point may be represented by a triple  $(a, b, c) \in \mathbb{O}^3$  with  $|a|^2 + |b|^2 + |c|^2 = 1$ , and for the three charts we assume  $a, b$ , or  $c$  are in  $\mathbb{R}_+$ , respectively. In practice, for computations with generic configurations we can simply work in the first chart and refer to a projection matrix by its associated point  $(a, b, c) \in \mathbb{R}_+ \times \mathbb{O}^2$ .

## 2.2 Tight simplices

Projective spaces can be embedded into Euclidean space by identifying a point with its associated projection matrix; using this embedding, the standard bounds on the size and distance of regular simplices in Euclidean space imply bounds on simplices in projective space. The resulting bounds for regular simplices in real projective space were first proven by Lemmens and Seidel [30]. They are also known in information theory as Welch bounds [41].

As above, let  $K$  be  $\mathbb{R}, \mathbb{C}, \mathbb{H}$ , or  $\mathbb{O}$ . We consider regular simplices in  $K\mathbb{P}^{d-1}$ , with the understanding that  $d = 2$  when  $K = \mathbb{O}$ .

**Definition 2.1.** A *regular simplex* in a metric space  $(X, \rho)$  is a collection of distinct points  $x_1, \dots, x_N$  of  $X$  with the distances  $\rho(x_i, x_j)$  all equal for  $i \neq j$ .

We often drop the adjective ‘‘regular,’’ as by a simplex we always mean a set of equidistant points.

**Proposition 2.2.** Consider a regular simplex consisting of  $N$  points  $x_1, \dots, x_N$  with associated projection matrices  $\Pi_1, \dots, \Pi_N$ , and let  $\alpha = \langle \Pi_i, \Pi_j \rangle$  be the common inner product for  $i \neq j$ . Then

$$N \leq d + \frac{(d^2 - d) \dim_{\mathbb{R}} K}{2}$$

and, for any such value of  $N$ ,

$$\alpha \geq \frac{N - d}{d(N - 1)}.$$

*Proof.* The Gram matrix  $G$  associated to  $\Pi_1, \dots, \Pi_N$  has unit diagonal and  $\alpha$  in each off-diagonal entry. Since  $G$  is  $(1 - \alpha)I_N$  plus a rank one matrix, an easy computation shows  $\det G = (1 - \alpha)^{N-1}(1 + (N - 1)\alpha)$ , which is nonzero because  $\alpha \in [0, 1)$ . Thus  $G$  is nonsingular, and the elements  $\Pi_1, \dots, \Pi_N \in \mathcal{H}(K^d)$  are linearly independent, implying  $N \leq \dim_{\mathbb{R}} \mathcal{H}(K^d) = d + (d^2 - d)(\dim_{\mathbb{R}} K)/2$ . Now note that  $\langle \Pi_i, I_d \rangle = |x_i|^2 = 1$  for each  $i = 1, \dots, N$ . Using this we compute

$$\left\langle \left( \sum_{i=1}^N \Pi_i \right) - \frac{N}{d} I_d, \left( \sum_{i=1}^N \Pi_i \right) - \frac{N}{d} I_d \right\rangle = N - \frac{N^2}{d} + N(N - 1)\alpha.$$

Nonnegativity of this expression gives the desired bound on  $\alpha$ . □

**Definition 2.3.** We refer to a regular simplex with

$$\alpha = \frac{N - d}{d(N - 1)}$$

as a *tight simplex*. That is, it is a simplex with the maximum possible distance allowed by Proposition 2.2.

Note that this definition is independent of the coordinate algebra  $K$ . In other words, the embeddings  $\mathbb{R}\mathbb{P}^{d-1} \subset \mathbb{C}\mathbb{P}^{d-1} \subset \mathbb{H}\mathbb{P}^{d-1}$  preserve tight simplices.

**Proposition 2.4.** *Every tight simplex is an optimal code.*

More generally, the bound on  $\alpha$  in Proposition 2.2 applies to the minimal distance of any code, not just a simplex.

*Proof.* Let  $\Pi_1, \dots, \Pi_N$  be the projection matrices corresponding to any  $N$ -point code in  $K\mathbb{P}^{d-1}$ . As in the proof of Proposition 2.2,

$$N - \frac{N^2}{d} + \sum_{\substack{i,j=1 \\ i \neq j}}^N \langle \Pi_i, \Pi_j \rangle = \left\langle \left( \sum_{i=1}^N \Pi_i \right) - \frac{N}{d} I_d, \left( \sum_{i=1}^N \Pi_i \right) - \frac{N}{d} I_d \right\rangle \geq 0.$$

Thus, the average of  $\langle \Pi_i, \Pi_j \rangle$  over all  $i \neq j$  satisfies

$$\frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^N \langle \Pi_i, \Pi_j \rangle \geq \frac{N^2/d - N}{N(N-1)} = \frac{N-d}{d(N-1)}.$$

In particular, the greatest value of  $\langle \Pi_i, \Pi_j \rangle$  for  $i \neq j$  must be at least this large.  $\square$

A regular simplex of  $N \leq d$  points in  $K\mathbb{P}^{d-1}$  is optimal if and only if the points are orthogonal (i.e.,  $\alpha = 0$ ). Such simplices always exist. We only consider them to be tight when  $N = d$ , as the  $N < d$  cases are degenerate. There also always exists a tight simplex with  $N = d + 1$  points, obtained by projecting the regular simplex on the sphere  $S^{d-1}$  into  $\mathbb{R}\mathbb{P}^{d-1}$ . Therefore in what follows we will generally assume  $N \geq d + 2$ .

It follows immediately from the proof of Proposition 2.2 that a simplex  $\{x_1, \dots, x_N\}$  is tight if and only if

$$\sum_{i=1}^N x_i x_i^\dagger = \frac{N}{d} I_d.$$

This condition can be reformulated in the language of projective designs [20, 34] (see also [23] for a detailed account of the relevant computations in projective space). Specifically, it says that the configuration is a 1-design. We will make no serious use of the theory of designs in this paper, and for our purposes we could simply regard  $\sum_{i=1}^N x_i x_i^\dagger = \frac{N}{d} I_d$  as the definition of a 1-design, but we will briefly recall the general concept in our discussion of linear programming bounds.

### 2.3 Linear programming bounds

Linear programming bounds [26, 20] use harmonic analysis on a space  $X$  to prove bounds on codes in  $X$ . These bounds and their extensions [4] are one of the only known ways to prove systematic bounds on codes, and they are sharp in a number of important cases. Later in this section we will summarize the sharp cases that are known in projective spaces (see also Table 1 in [14] for a corresponding list for spheres), but first we will give a brief review of how linear programming bounds work.

The simplest setting for linear programming bounds is a compact two-point homogeneous space. We will focus on the connected examples, namely spheres and projective spaces, but

discrete two-point homogeneous spaces such as the Hamming cube are also important in coding theory.

Let  $X$  be a sphere or projective space, and let  $G$  be its isometry group under the geodesic metric  $\vartheta$  (normalized so that the greatest distance is  $\pi$ ). Then  $L^2(X)$  is a unitary representation of  $G$ , and we can decompose it as a completed direct sum

$$L^2(X) = \widehat{\bigoplus_{k \geq 0} V_k}$$

of irreducible representations  $V_k$ . There is a corresponding sequence of *zonal spherical functions*  $C_0, C_1, \dots$ , one attached to each representation  $V_k$ . The zonal spherical functions are most easily obtained as reproducing kernels; for a brief review of the theory, see Sections 2.2 and 8 of [14]. We can represent them as orthogonal polynomials with respect to a measure on  $[-1, 1]$ , which depends on the space  $X$ , and we index them so that  $C_k$  has degree  $k$ .

For our purposes, the most important property of zonal spherical functions is that they are *positive definite kernels*: for all  $N \in \mathbb{N}$  and  $x_1, \dots, x_N \in X$ , the  $N \times N$  matrix  $(C_i(\cos \vartheta(x_i, x_j)))_{1 \leq i, j \leq N}$  is positive semidefinite. In fact, the zonal spherical functions span the cone of all such functions.

For projective spaces  $K\mathbb{P}^{d-1}$ , the polynomials  $C_k$  may be taken to be the Jacobi polynomials  $P_k^{(\alpha, \beta)}$ , where  $\alpha = (d-1)(\dim_{\mathbb{R}} K)/2 - 1$  (i.e.  $\alpha = (\dim_{\mathbb{R}} K\mathbb{P}^{d-1})/2 - 1$ ) and  $\beta = (\dim_{\mathbb{R}} K)/2 - 1$ . We will normalize  $C_0$  to be 1.

Linear programming bounds for codes amount to the following proposition:

**Proposition 2.5.** *Let  $\theta \in [0, \pi]$ , and suppose the polynomial*

$$f(z) = \sum_{k=0}^n f_k C_k(z)$$

*satisfies  $f_0 > 0$ ,  $f_k \geq 0$  for  $1 \leq k \leq n$ , and  $f(z) \leq 0$  for  $-1 \leq z \leq \cos \theta$ . Then every code in  $X$  with minimal geodesic distance at least  $\theta$  has size at most  $f(1)/f_0$ .*

*Proof.* Let  $\mathcal{C}$  be such a code. Then

$$\sum_{x, y \in \mathcal{C}} f(\cos \vartheta(x, y)) \geq f_0 |\mathcal{C}|^2,$$

because each zonal spherical function  $C_k$  is positive definite and hence satisfies

$$\sum_{x, y \in \mathcal{C}} C_k(\cos \vartheta(x, y)) \geq 0.$$

On the other hand,  $f(\cos \vartheta(x, y)) \leq 0$  whenever  $\vartheta(x, y) \geq \theta$ , and hence

$$\sum_{x, y \in \mathcal{C}} f(\cos \vartheta(x, y)) \leq |\mathcal{C}| f(1)$$

because only the diagonal terms contribute positively. It follows that  $f_0 |\mathcal{C}|^2 \leq f(1) |\mathcal{C}|$ , as desired.  $\square$

We say this bound is *sharp* if there is a code  $\mathcal{C}$  with minimal distance at least  $\theta$  and  $|\mathcal{C}| = f(1)/f_0$ . Note that we require exact equality, rather than just  $|\mathcal{C}| = \lfloor f(1)/f_0 \rfloor$ .

**Definition 2.6.** A *tight code* is one for which linear programming bounds are sharp.

Examining the proof of Proposition 2.5 yields the following characterization of tight codes:

**Lemma 2.7.** A code  $\mathcal{C}$  with minimal geodesic distance  $\theta$  is tight iff there is a polynomial  $f(z) = \sum_{k=0}^n f_k C_k(z)$  satisfying  $f_0 > 0$ ,  $f_k \geq 0$  for  $1 \leq k \leq n$ ,  $f(z) \leq 0$  for  $-1 \leq z \leq \cos \theta$ ,

$$\sum_{x,y \in \mathcal{C}} C_k(\cos \vartheta(x,y)) = 0$$

whenever  $f_k > 0$  and  $k \neq 0$ , and  $f(\cos \vartheta(x,y)) = 0$  for  $x, y \in \mathcal{C}$  with  $x \neq y$ . In fact, these conditions must hold for every polynomial  $f$  satisfying both  $f(1)/f_0 = |\mathcal{C}|$  and the hypotheses of Proposition 2.5.

By Proposition 2.5, every tight code is as large as possible given its minimal distance, but it is less obvious that such a code maximizes minimal distance given its size.

**Proposition 2.8.** Every tight code is optimal.

*Proof.* Suppose  $f$  satisfies the hypotheses of Proposition 2.5, and  $\mathcal{C}$  is a code of size  $f(1)/f_0$  with minimal geodesic distance at least  $\theta$ . We wish to show that its minimal distance is exactly  $\theta$ .

By Lemma 2.7

$$\sum_{x,y \in \mathcal{C}} (f(\cos \vartheta(x,y)) - f_0) = 0$$

and  $f(\cos \vartheta(x,y)) = 0$  for  $x, y \in \mathcal{C}$  with  $x \neq y$

Now suppose  $\mathcal{C}$  had minimal geodesic distance strictly greater than  $\theta$ , and consider a small perturbation  $\mathcal{C}'$  of  $\mathcal{C}$ . It must satisfy

$$\sum_{x,y \in \mathcal{C}'} (f(\cos \vartheta(x,y)) - f_0) \geq 0,$$

by positive definiteness. On the other hand,

$$\sum_{x,y \in \mathcal{C}'} (f(\cos \vartheta(x,y)) - f_0) = |\mathcal{C}'|f(1) - |\mathcal{C}'|^2 f_0 + \sum_{\substack{x,y \in \mathcal{C}' \\ x \neq y}} f(\cos \vartheta(x,y)).$$

We have  $|\mathcal{C}'|f(1) - |\mathcal{C}'|^2 f_0 = 0$  since  $|\mathcal{C}'| = |\mathcal{C}| = f(1)/f_0$ . Thus,

$$\sum_{\substack{x,y \in \mathcal{C}' \\ x \neq y}} f(\cos \vartheta(x,y)) \geq 0.$$

If the perturbation is small enough, then  $f(\cos \vartheta(x,y)) \leq 0$  for distinct  $x, y \in \mathcal{C}'$ . In that case, we must have  $f(\cos \vartheta(x,y)) = 0$  for distinct  $x, y \in \mathcal{C}'$ . However, this fails for some perturbations, for example if we move two points slightly closer together. It follows that every code of size  $f(1)/f_0$  and minimal geodesic distance at least  $\theta$  has minimal distance exactly  $\theta$ , so these codes are all optimal.  $\square$

**Lemma 2.9.** *Tight simplices in projective space are tight codes.*

*Proof.* Up to scaling, the first-degree zonal spherical function  $C_1$  on  $K\mathbb{P}^{d-1}$  is  $z + \frac{d-2}{d}$ . Now let

$$f(z) = 1 + \frac{(N-1)d}{2(d-1)} \left( z + \frac{d-2}{d} \right).$$

It satisfies  $f(z) \leq 0$  for  $z \in [-1, 2\alpha - 1]$ , where

$$\alpha = \frac{N-d}{d(N-1)},$$

and  $f(1)/f_0 = N$ , as desired.  $\square$

Note that in this proof  $C_1$  depends only on  $d$ , and not on  $K$ . By contrast, higher-degree zonal spherical functions for  $K\mathbb{P}^{d-1}$  depend on both  $d$  and  $K$ .

A  $t$ -design in  $X$  is a code  $\mathcal{C} \subset X$  such that for every  $f \in V_k$  with  $0 < k \leq t$ ,

$$\sum_{x \in \mathcal{C}} f(x) = 0.$$

In other words, every element of  $V_0 \oplus \cdots \oplus V_t$  has the same average over  $\mathcal{C}$  as over the entire space  $X$ . (Note that all functions in  $V_k$  for  $k > 0$  have average zero, since they are orthogonal to the constant functions in  $V_0$ .) Using the reproducing kernel property, this can be shown to be equivalent to

$$\sum_{x, y \in \mathcal{C}} C_k(\cos \vartheta(x, y)) = 0$$

for  $0 < k \leq t$ .

In  $K\mathbb{P}^{d-1}$ , one can check that

$$\sum_{i=1}^N x_i x_i^\dagger = \frac{N}{d} I_d$$

holds if and only if  $\{x_1, \dots, x_N\}$  is a 1-design.

A code is *diametrical* in  $X$  if it contains two points at maximal distance in  $X$ , and it is *m-distance set* if exactly  $m$  distances occur between distinct points.

**Definition 2.10.** A *tight design* is an  $m$ -distance set that is a  $(2m - \varepsilon)$ -design, where  $\varepsilon$  is 1 if the set is diametrical and 0 otherwise.

For example, an  $N$ -point tight simplex in  $K\mathbb{P}^{d-1}$  with  $N = d + (d^2 - d)(\dim_{\mathbb{R}} K)/2$  (the largest possible value of  $N$ ) is a tight 2-design. See [7] for further examples.

Every tight  $t$ -design is the smallest possible  $t$ -design in its ambient space. This was first proved for spheres in [20]; see Propositions 1.1 and 1.2 in [6] for the general case. The converse is false: the smallest  $t$ -design is generally not tight.

A theorem of Levenshtein [32] says that every  $m$ -distance set that is a  $(2m - 1 - \varepsilon)$ -design is a tight code, where as above  $\varepsilon$  is 1 if the set is diametrical and 0 otherwise. For example, all tight designs are tight codes. In [14], it was also shown that under these conditions,  $\mathcal{C}$  is *universally optimal* for potential energy: it minimizes energy for every completely monotonic



Table 2.1: Known universal optima of  $N$  points in real projective spaces  $\mathbb{RP}^{d-1}$ . The tight simplices are indicated by an asterisk and have maximal squared inner product  $(N - d)/(d(N - 1))$ ; for simplicity we omit the Gale duals of the tight simplices.

$d$	$N$	$\max  \langle x, y \rangle ^2$	Name/origin
$d$	$N \leq d + 1$	*	Euclidean simplex
$d$	$2d$	*	symm. conf. matrix of order $2d$
$d$	$d(d + 2)/2$	$1/d$	MUB with $d = 4^k$
2	$N$	$\cos^2(\pi/N)$	regular polygon
4	60	$(\sqrt{5} - 1)/4$	regular 600-cell
6	16	*	Clebsch
6	36	$1/4$	$E_6$ root system
7	28	*	equiangular lines
7	63	$1/4$	$E_7$ root system
8	120	$1/4$	$E_8$ root system
23	276	*	equiangular lines
23	2300	$1/9$	kissing configuration of next line
24	98280	$1/4$	Leech lattice minimal vectors
35	120	*	embedding of $E_8$ from [9]

function of squared chordal distance. (See also [13] for context.) This applies in particular to simplices, so all tight simplices are universally optimal.

In fact every known tight code is universally optimal. Moreover, except for the regular 600-cell in  $S^3$  and its image in  $\mathbb{RP}^3$ , they all satisfy the design condition just mentioned. For lack of a counterexample, we conjecture that tight codes are always universally optimal. (But see [16] for perspective on why the simplest reason why this might hold fails.)

## 2.4 Tight codes in $\mathbb{RP}^{n-1}$

We now describe what is known about tight codes in real projective spaces. Table 2.1 provides a summary of the current state of knowledge.

Euclidean simplices give the simplest infinite family of tight codes.

Another infinite family of tight simplices comes from conference matrices [40] (see [17, p. 156]): if a symmetric conference matrix of order  $2d$  exists, then there is a tight simplex of size  $2d$  in  $\mathbb{R}^d$ . In particular, we get a tight simplex in  $\mathbb{R}^d$  whenever  $2d - 1$  is a prime power congruent to 1 modulo 4. One can also construct such codes through the Weil representation of the group  $G = \text{PSL}_2(\mathbb{F}_q)$ . Note that the icosahedron arises as the special case  $q = 5$ , which is why it is not listed separately in Table 2.1.

Levenshtein [31] described a family of tight codes in  $\mathbb{RP}^{n-1}$  for  $n$  a power of 4, based on a construction using Kerdock codes. These codes meet the orthoplex bound (Corollary 5.3 in [17]) and give rise to mutually unbiased bases in their dimensions. The regular 24-cell is the special case with  $n = 4$ .

A trivial systematic family of tight codes is formed by the diameters of the regular polygons in the plane. The remaining configurations in Table 2.1 correspond to exceptional geometric structures.

Table 2.2: Known universal optima of  $N$  points in complex projective spaces  $\mathbb{C}\mathbb{P}^{d-1}$ . The tight simplices are indicated by an asterisk and have maximal squared inner product  $(N-d)/(d(N-1))$ ; for simplicity we omit the Gale duals of the tight simplices as well as the tight simplices from  $\mathbb{R}\mathbb{P}^{d-1}$ .

$d$	$N$	$\max  \langle x, y \rangle ^2$	Name/origin
$d$	$2d$	*	skew-symm. conf. matrix of order $2d$
$d$	$d^2$	*	SIC-POVMs
$d$	$d(d+1)$	$1/d$	MUB with $d = p^k$ and $p$ a prime
$d$	$2d+1$	*	skew-Hadamard matrix of order $2d+2$ ( $d$ odd)
$d$	$2d-1$	*	skew-Hadamard matrix of order $2d$ ( $d$ even)
4	40	$1/3$	Eisenstein structure on $E_8$
5	45	$1/4$	kissing configuration of next line
6	126	$1/4$	Eisenstein structure on $K_{12}$
28	4060	$1/16$	Rudvalis group
$ S $	$ G $	*	difference set $S$ in abelian group $G$

We also observe the phenomenon of Gale duality: tight simplices of size  $N$  in  $K\mathbb{P}^{d-1}$  correspond to tight simplices of size  $N$  in  $K\mathbb{P}^{N-d-1}$ . For instance, the Gale dual of the Clebsch configuration gives a tight simplex of 16 points in  $\mathbb{R}\mathbb{P}^9$ . See §2.7 for more details.

## 2.5 Tight codes in $\mathbb{C}\mathbb{P}^{n-1}$

Table 2.2 lists the tight codes we are aware of in complex projective spaces. For a detailed survey of tight simplices, we refer the reader to Chapter 4 of [28].

Here, we observe a few more infinite families. In particular, if a conference matrix of order  $2d$  exists, then there is a tight code of  $2d$  lines in  $\mathbb{C}\mathbb{P}^{d-1}$  [44, p. 66]. For prime powers  $q \equiv 3 \pmod{4}$ , this gives a construction of a tight  $(q+1)$ -point code in  $\mathbb{C}\mathbb{P}^{(q-1)/2}$ . As mentioned before, such codes may also be constructed using the Weil representation of  $\text{PSL}_2(\mathbb{F}_q)$ . Another family of codes of  $d(d+1)$  points in  $\mathbb{C}\mathbb{P}^{d-1}$ , for  $d$  an odd prime power, was constructed by Levenshtein [31] using dual BCH codes. These codes meet the orthoplex bound and give rise to mutually unbiased bases in their dimensions. They were rediscovered by Wootters and Fields [42], with an extension to even characteristic and applications to physics. A third infinite family is obtained from skew-Hadamard matrices (see [36] for a construction using explicit families of skew-Hadamard matrices and Theorem 4.14 in [28] for the general case).

The most mysterious tight simplices are the awkwardly named SIC-POVMs (symmetric, informationally complete positive operator valued measures). SIC-POVMs are simplices of size  $d^2$  in  $\mathbb{C}\mathbb{P}^{d-1}$ , i.e., simplices of the greatest size allowed by Proposition 2.2. These configurations play an important role in quantum information theory, which leads to their name. Numerical experiments suggest they exist in all dimensions, and that they can even be taken to be orbits of the Weyl-Heisenberg group [44, 37]. Exact SIC-POVMs are known for  $d \leq 15$ , as well as  $d = 19, 24, 35$ , and  $48$ , and numerical approximations are known for all  $d \leq 67$  (see [38]).

The last line of the table refers to a construction based on difference sets [43] (see also

Table 2.3: Previously known universal optima of  $N$  points in quaternionic and octonionic projective spaces. For simplicity we omit the tight simplices from  $\mathbb{R}\mathbb{P}^{d-1}$  and  $\mathbb{C}\mathbb{P}^{d-1}$ .

Space	$N$	$\max  \langle x, y \rangle ^2$	Name/origin
$\mathbb{H}\mathbb{P}^{d-1}$	$d(2d+1)$	$1/d$	MUB with $d = 4^k$
$\mathbb{H}\mathbb{P}^4$	165	$1/4$	quaternionic reflection group
$\mathbb{O}\mathbb{P}^2$	819	$1/2$	generalized hexagon of order (2, 8)

[29]). Let  $G$  be an abelian group of order  $N$ ,  $S$  a subset of  $G$  of order  $d$ , and  $\lambda$  a natural number such that every nonzero element of  $G$  is a difference of exactly  $\lambda$  pairs of elements of  $S$ . It follows that  $d(d-1) = \lambda(N-1)$ , and that the vectors

$$v_\chi = (\chi(s))_{s \in S}$$

give rise to a tight simplex of  $N$  points in  $\mathbb{P}^{d-1}$  as  $\chi$  ranges over all characters of  $G$ . As particular cases of this construction, one can obtain a tight simplex of  $n^2 + n + 1$  points in  $\mathbb{C}\mathbb{P}^n$ , when there is a projective plane of order  $n$ . A generalization of this example was given in [43], using Singer difference sets, to produce  $(q^{d+1} - 1)/(q - 1)$  points in  $\mathbb{C}\mathbb{P}^{d-1}$ , with  $d = (q^d - 1)/(q - 1)$ . Similarly, if  $q$  is a prime power congruent to 3 modulo 4, then the quadratic residues give a difference set, yielding a tight simplex of  $q$  points in  $\mathbb{C}\mathbb{P}^{(q-3)/2}$ . As another example, there is a difference set of 6 points in  $\mathbb{Z}/31\mathbb{Z}$  (namely,  $\{0, 1, 4, 6, 13, 21\}$ ), which gives rise to a tight simplex of 31 points in  $\mathbb{C}\mathbb{P}^5$ .

## 2.6 Tight codes in $\mathbb{H}\mathbb{P}^{d-1}$ and $\mathbb{O}\mathbb{P}^2$

Relatively little is known about tight codes in quaternionic or octonionic projective spaces, aside from the real and complex tight simplices they automatically contain. There is a construction of mutually unbiased bases due to Kantor [27], and two exceptional codes are known.

The 165 points in  $\mathbb{H}\mathbb{P}^4$  from Table 2.3 are constructed using a quaternionic reflection group (Example 9 in [23]). The 819-point universal optimum is a remarkable code in the octonionic projective plane [12]; see also [21] for another construction. It can be thought of as the 196560 Leech lattice minimal vectors modulo the action of the 240 roots of  $E_8$  (viewed as units in the integral octonions), although this does not yield an actual construction: there is no such action because the multiplication is not associative.

## 2.7 Gale duality

Gale duality is a fundamental symmetry of tight simplices. It goes by several names in the literature, such as coherent duality, Naimark complements, or the theory of eutactic stars. We call it Gale duality because it is a metric version of Gale duality from the theory of polytopes.

Let  $K$  be  $\mathbb{R}$ ,  $\mathbb{C}$ , or  $\mathbb{H}$ . Note that Gale duality does not apply to  $\mathbb{O}\mathbb{P}^2$ .

**Proposition 2.11** (Hadwiger [22]). *Let  $v_1, \dots, v_N$  span a  $d$ -dimensional vector space  $V$  over  $K$ , and suppose they have the same norm  $|v_i|^2 = d/N$ . Then their images in  $K\mathbb{P}^{d-1}$*

form a 1-design if and only if there is an  $N$ -dimensional vector space  $U$  containing  $V$  and an orthonormal basis  $u_1, \dots, u_N$  of  $U$  such that  $v_i$  is the orthogonal projection of  $u_i$  to  $V$ .

*Proof.* Let  $M$  be the  $d \times N$  matrix whose  $i^{\text{th}}$  column is  $v_i$ . The existence of  $U$  and  $u_1, \dots, u_N$  is equivalent to an extension of  $M$  to a unitary matrix by adding  $N - d$  rows, in which case  $u_1, \dots, u_N$  are the columns of the extended matrix. This extension is possible if and only if the rows of  $M$  are orthonormal vectors; in other words, it is equivalent to  $MM^\dagger = I_d$ .

To analyze  $M$ , we can write it as  $M = \sum_{i=1}^N v_i e_i^\dagger$ , where  $e_1, \dots, e_N$  is the standard orthonormal basis of  $K^N$ . Then

$$MM^\dagger = \sum_{i,j=1}^N v_i e_i^\dagger e_j v_j^\dagger = \sum_{i=1}^N v_i v_i^\dagger.$$

Thus, the extension is possible if and only if

$$\sum_{i=1}^N v_i v_i^\dagger = I_d,$$

which is the condition for a projective 1-design once we rescale to account for  $|v_i|^2 = d/N$ .  $\square$

Under the 1-design condition from Proposition 2.11, consider the projection  $w_i$  of the vectors  $u_i$  to the orthogonal complement  $V^\perp$  of  $V$  in  $U$ . This code  $\{w_1, \dots, w_N\}$  in  $K\mathbb{P}^{N-d-1}$  is called the *Gale dual* of the code  $\{v_1, \dots, v_N\}$  in  $K\mathbb{P}^{d-1}$ . However, there is one technicality: the  $N$  points in  $K\mathbb{P}^{N-d-1}$  needn't be distinct in general, so the Gale dual must be considered a multiset of points. Aside from the need to allow multisets, Gale duality is an involution on projective 1-designs, defined up to isometry.

Gale duality preserves tight simplices when  $N > d + 1$ , and the multiplicity issue does not arise:

**Corollary 2.12.** *Let  $K$  be  $\mathbb{R}$ ,  $\mathbb{C}$ , or  $\mathbb{H}$ . For  $N > d + 1$ , the Gale dual of an  $N$ -point tight simplex in  $K\mathbb{P}^{d-1}$  is an  $N$ -point tight simplex in  $K\mathbb{P}^{N-d-1}$ .*

*Proof.* Because the 1-design property is preserved, we need only check that the Gale dual is a simplex. In the notation used above, for  $i \neq j$  we have

$$0 = \langle u_i, u_j \rangle = \langle v_i, v_j \rangle + \langle w_i, w_j \rangle.$$

Thus,  $\langle w_i, w_j \rangle$  is constant for  $i \neq j$  because  $\langle v_i, v_j \rangle$  is. The inequality  $N > d + 1$  merely rules out the degenerate case  $K\mathbb{P}^0$ .  $\square$

The inequality

$$N \leq d + \frac{(d^2 - d) \dim_{\mathbb{R}} K}{2}$$

from Proposition 2.2 shows that tight simplices cannot be too large. Combining Gale duality with the same inequality shows that they cannot be too small, either (see Theorem 2.30 in [44] and Corollary 2.19 in [28]):

**Corollary 2.13.** *Let  $K$  be  $\mathbb{R}$ ,  $\mathbb{C}$ , or  $\mathbb{H}$ . If there exists an  $N$ -point tight simplex in  $K\mathbb{P}^{d-1}$  with  $N > d + 1$ , then*

$$N \geq d + \frac{1 + \sqrt{1 + \frac{8d}{\dim_{\mathbb{R}} K}}}{2}.$$

### 3 Simplicies in Quaternionic Projective Spaces

#### 3.1 Generic case

The definition gives one characterization of tight  $N$ -point simplices; we simply impose  $|x_i|^2 = 1$  for each  $i$  and  $|\langle x_i, x_j \rangle|^2 = (N-d)/(d(N-1))$  for  $i < j$ . In fact, tight simplices can be characterized even more succinctly: it can be shown that  $\sum_{i,j} |\langle x_i, x_j \rangle|^2 \geq N^2/d$ , with equality iff  $\{x_1, \dots, x_N\}$  is a tight simplex. Both of these descriptions, though, suffer from the problem that the imposed conditions are *singular*; loosely put, if a set of points satisfies the conditions, then it does so “just barely.” In other words, if we define  $f: \mathbb{H}^N \rightarrow \mathbb{R}^{N+1}$  by

$$f(x_1, \dots, x_N) = \left( |x_1|^2 - 1, \dots, |x_N|^2 - 1, \sum_{\substack{i,j \\ i \neq j}} |\langle x_i, x_j \rangle|^2 - N^2/d \right),$$

then the fact that the last coordinate is always nonnegative implies that the last row of  $Df$  is zero at a tight simplex. Therefore it is hopeless to try to prove existence by applying Theorem I.2.1. Setting all the inner products equal to  $(N-d)/(d(N-1))$  suffers from the same problem, because

$$\frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^N |\langle x_i, x_j \rangle|^2 \geq \frac{N-d}{d(N-1)}$$

for all  $x_1, \dots, x_N$  (see the proof of Proposition 2.4).

Fortunately, it is generally possible to recast the conditions describing tight simplices so that the Jacobian of the associated polynomial map becomes surjective.

**Proposition 3.1.** *Suppose  $x_1, \dots, x_N \in \mathbb{H}^d$  ( $d > 1$ ) and  $w_1, \dots, w_N \in \mathbb{R}$  satisfy the following conditions:*

- (a)  $|x_i|^2 = 1$  for  $i = 1, \dots, N$ ,
- (b)  $|\langle x_i, x_j \rangle|^2 = |\langle x_{i'}, x_{j'} \rangle|^2$  for  $1 \leq i < j \leq N$  and  $1 \leq i' < j' \leq N$ , and
- (c)  $\sum_{i=1}^N w_i x_i x_i^\dagger = I_d$ .

Then  $w_1 = \dots = w_N = \frac{d}{N}$  and  $\{x_1, \dots, x_N\}$  is a tight simplex in  $\mathbb{H}\mathbb{P}^{d-1}$ .

*Proof.* Define  $\Pi_i = x_i x_i^\dagger$ , and let  $\alpha$  denote the common inner product  $|\langle x_i, x_j \rangle|^2$  for  $i \neq j$ . By the first condition we have  $\langle \Pi_i, I_d \rangle = 1$  for each  $i$ . Thus

$$d = \langle I_d, I_d \rangle = \sum_{i=1}^N w_i \langle \Pi_i, I_d \rangle = \sum_{i=1}^N w_i.$$

Moreover, using equation (2.1) we have  $\langle \Pi_i, \Pi_i \rangle = 1$  and  $\langle \Pi_i, \Pi_j \rangle = \alpha$  for all  $i \neq j$ . Thus, for any  $j$ ,

$$1 = \langle \Pi_j, I_d \rangle = \sum_{i=1}^N w_i \langle \Pi_j, \Pi_i \rangle = (1 - \alpha)w_j + \alpha \cdot \sum_{i=1}^N w_i = (1 - \alpha)w_j + \alpha d.$$

It follows that  $w_j = (1 - \alpha d)/(1 - \alpha)$  for each  $j$ . Substituting back into the equation  $\sum_{i=1}^N w_i = d$  yields  $\alpha = (N - d)/(d(N - 1))$ , from which the result follows.  $\square$

Using Proposition 3.1, we can view tight simplices of  $N$  points in  $\mathbb{H}\mathbb{P}^{d-1}$  as the solutions of a system of

$$N + \left( \frac{N(N-1)}{2} - 1 \right) + (2d^2 - d) \text{ real constraints}$$

in

$$N(4d + 1) \text{ real variables.}$$

In situations where Theorem I.2.1 apply to this system, we get a solution space of dimension (number of variables)  $-$  (number of constraints) by Proposition I.2.3. This separately counts each unit-norm lift of the  $N$  elements of  $\mathbb{H}\mathbb{P}^{d-1}$ , so the actual space of simplices has codimension  $3N$ . Moreover, the space of simplices is invariant under the action of the symmetry group of  $\mathbb{H}\mathbb{P}^{d-1}$ . This symmetry group, the compact symplectic group  $\text{Sp}(d)$  (strictly speaking, modulo its center  $\{\pm 1\}$ ), has real dimension  $d(2d + 1)$ . Thus the actual dimension of the space of simplices, local to this particular solution and modulo symmetries, is at least

$$r(N, \mathbb{H}\mathbb{P}^{d-1}) := (4d - 3)N - \frac{N(N-1)}{2} - 4d^2 + 1 \quad (3.1)$$

when Theorem I.2.1 (and thus also Proposition I.2.3) and Proposition 3.1 apply. Equality holds if the resulting simplices have only a discrete symmetry group; otherwise the solutions modulo symmetry have dimension greater than  $r(N, \mathbb{H}\mathbb{P}^{d-1})$ .

While *a priori* it is possible to have tight simplices of up to  $N = 2d^2 - d$  points, we only have  $r(N, \mathbb{H}\mathbb{P}^{d-1}) \geq 0$  for  $N$  between roughly  $2(2 - \sqrt{2})d$  and  $2(2 + \sqrt{2})d$ . This does not rule out larger tight simplices, but it does mean that this approach using Proposition 3.1 and Theorem I.2.1 could not prove their existence. We believe that outside of this range, only sporadic examples will exist in general.

Note that Gale duality (replacing  $d$  with  $N - d$ ) preserves  $r(N, \mathbb{H}\mathbb{P}^{d-1})$ , as one would expect. Furthermore, because  $r(N, \mathbb{H}\mathbb{P}^{d-1})$  is quadratic in  $N$ , it is also symmetric about the midpoint of the range in which it is positive. Specifically,  $r(N, \mathbb{H}\mathbb{P}^{d-1}) = r(8d - 5 - N, \mathbb{H}\mathbb{P}^{d-1})$ .

*Remark 3.2.* We emphasize that  $r(N, \mathbb{H}\mathbb{P}^{d-1})$  is *defined* by (3.1). The assertion that the space of simplices locally has dimension  $r(N, \mathbb{H}\mathbb{P}^{d-1})$  is true only when (i) we find a numerical solution of the conditions of Proposition 3.1 to which Theorem I.2.1 applies, and (ii) the action of the symmetry group on our simplex has finite (0-dimensional) stabilizer. Regarding (ii), we have checked this numerically but not rigorously (see §6.3); of all the cases in part (a) of Tables 3.1–4.1, only 5-point simplices in  $\mathbb{O}\mathbb{P}^2$  fail to satisfy the condition. In that case there is a 3-dimensional stabilizer. We accounted for this in Table 4.1.

When we attempt to apply Proposition 3.1, there are three possible outcomes:

- (a) we find an approximate numerical solution with surjective Jacobian, in which case we can prove existence using Theorem I.2.1,
- (b) we find an approximate numerical solution, but the Jacobian at that point is not surjective, or
- (c) we cannot even find an approximate numerical solution to the system, in which case we conjecture that there exists no tight simplex.

In a few cases we encountered a fourth possibility:

- (d) we find what appears to be an approximate solution but we are unable to converge to greater precision.

Table 3.1: Cases in  $\mathbb{H}\mathbb{P}^2$ : (a) proven existence of tight simplices; (b) singular Jacobian; (c) conjectured nonexistence.

$N$	$r(N, \mathbb{H}\mathbb{P}^2)$	$N$	rank deficiency	$N$
5	0	12	2	14
6	4	13	2	(c)
7	7	15	14	
8	9	(b)		
9	10			
10	10			
11	9			
(a)				

Table 3.2: Cases in  $\mathbb{H}\mathbb{P}^3$ : (a) proven existence of tight simplices; (c) conjectured nonexistence.

$N$	$r(N, \mathbb{H}\mathbb{P}^3)$	$N$	$r(N, \mathbb{H}\mathbb{P}^3)$	$N$
6	0	14	28	22–28
7	7	15	27	(c)
8	13	16	25	
9	18	17	22	
10	22	18	18	
11	25	19	13	
12	27	20	7	
13	28	21	0	
(a)				

When this situation arose we tried both Newton’s method and gradient descent for energy minimization (see §6.2), but we were unable to improve the error in the constraints beyond  $10^{-5}$  (as compared to a numerical error of about  $10^{-15}$  for cases (a) and (b)). In these cases we make no conjecture as to existence or non-existence of solutions.

Tables 3.1, 3.2, 3.3, and 3.4 list our results for  $d = 3$ ,  $d = 4$ ,  $d = 5$ , and  $d = 6$ , respectively. Each table lists all values of  $N$  from  $d + 2$  to the upper bound  $2d^2 - d$  from Proposition 2.2. There is no intrinsic problem with extending to larger dimensions, although the calculations become increasingly time-consuming.

**Theorem 3.3.** *For the values of  $(N, d)$  listed in part (a) of Tables 3.1 through 3.4, there exists a tight  $N$ -point simplex in  $\mathbb{H}\mathbb{P}^{d-1}$ .*

In fact, near the points found by our computer program and exhibited in the auxiliary files, the space of simplices modulo symmetries has dimension at least  $r(N, \mathbb{H}\mathbb{P}^{d-1})$ . As described above, we conjecture that these dimensions are exact, but they are at least lower bounds. In the case of a singular Jacobian (part (b) of the tables) we report the *rank deficiency* (i.e.,  $n - \text{rank}(Df)$  in the terminology of Theorem I.2.1).

In Table 3.3, i.e., in  $\mathbb{H}\mathbb{P}^4$ , we first observe a gap between the tight simplices of sizes  $d$  and  $d + 1$  that always exist in  $\mathbb{H}\mathbb{P}^{d-1}$  and the range of simplices for which our method

Table 3.3: Cases in  $\mathbb{H}\mathbb{P}^4$ : (a) proven existence of tight simplices; (c) conjectured nonexistence (proven for  $N = 7$ ); (d) ambiguous numerical results.

$N$	$r(N, \mathbb{H}\mathbb{P}^4)$	$N$	$r(N, \mathbb{H}\mathbb{P}^4)$	$N$	$r(N, \mathbb{H}\mathbb{P}^4)$	$N$
8	9	15	51	22	44	7
9	18	16	53	23	39	29–45
10	26	17	54	24	33	(c)
11	33	18	54	25	26	(c)
12	39	19	53	26	18	(c)
13	44	20	51	27	9	(c)
14	48	21	48			(d)

(a)

(d)

Table 3.4: Cases in  $\mathbb{H}\mathbb{P}^5$ : (a) proven existence of tight simplices; (c) conjectured nonexistence (proven for  $N = 8$ ); (d) ambiguous numerical results.

$N$	$r(N, \mathbb{H}\mathbb{P}^5)$	$N$	$r(N, \mathbb{H}\mathbb{P}^5)$	$N$	$r(N, \mathbb{H}\mathbb{P}^5)$	$N$
9	10	18	82	27	73	8
10	22	19	85	28	67	36–66
11	33	20	87	29	60	(c)
12	43	21	88	30	52	(c)
13	52	22	88	31	43	(c)
14	60	23	87	32	33	(c)
15	67	24	85	33	22	(c)
16	73	25	82	34	10	(c)
17	78	26	78			(d)

(a)

(d)



proves existence. The gap is real: there exists no 7-point tight simplex in  $\mathbb{H}\mathbb{P}^4$ , because of Corollary 2.13. Similarly, there exists no 8-point tight simplex in  $\mathbb{H}\mathbb{P}^5$ .

### 3.2 12- and 13-point simplices

The cases of 12- and 13-point simplices are somewhat special: the system of constraints specified by Proposition 3.1 has a rank deficiency. To prove existence of solutions using Theorem I.2.1, a different approach is needed.

We take as our starting point the following observation: not only do tight 12-point simplices exist (numerically), but actually 12-point cyclic-symmetric simplices exist (again, numerically). By this we mean a simplex such that, if  $(x, y, z) \in \mathbb{H}^3$  is a point in it, then so are  $(y, z, x)$  and  $(z, x, y)$ , and these are three distinct points in  $\mathbb{H}\mathbb{P}^2$ .

We would like to adapt Proposition 3.1 to find simplices with cyclic symmetry. Imposing this symmetry reduces the number of degrees of freedom we have, but it also reduces the number of conditions we need to check. Fortunately, we end up with a set of constraints that has a surjective Jacobian at a tight simplex.

For convenience we will state the result only for  $d = 3$ , but it naturally generalizes to any dimension.

**Proposition 3.4.** *Let  $\sigma$  be the cyclic-shift automorphism  $\sigma(a, b, c) = (b, c, a)$ . Suppose  $x_1, \dots, x_{3m} \in \mathbb{H}^3$  and  $w_1, \dots, w_{3m} \in \mathbb{R}$  satisfy the following conditions:*

- (a)  $x_{m+i} = \sigma(x_i)$  for  $i = 1, \dots, 2m$ ,
- (b)  $w_{m+i} = w_i$  for  $i = 1, \dots, 2m$ ,
- (c)  $|x_i|^2 = 1$  for  $i = 1, \dots, m$ ,
- (d) the squared inner products  $|\langle x_i, x_j \rangle|^2$  for  $i = 1, \dots, m$  and the following values of  $j$  are all equal: (i)  $j = i + m$ , (ii)  $i < j \leq m$ , (iii)  $i + m < j \leq 2m$ , (iv)  $i + 2m < j \leq 3m$ , and
- (e) the matrix  $\sum_{i=1}^{3m} w_i x_i x_i^\dagger$  has 1,1 entry equal to 1 and vanishing 1,2 entry.

Then  $w_1 = \dots = w_{3m} = \frac{1}{m}$  and  $\{x_1, \dots, x_{3m}\}$  is a tight simplex in  $\mathbb{H}\mathbb{P}^2$ .

*Proof.* By repeatedly applying the identities  $\langle x_i, x_j \rangle = \langle \sigma(x_i), \sigma(x_j) \rangle$ , it easily follows from (d) that  $\{x_1, \dots, x_{3m}\}$  is a simplex.

Having shown that, now consider the matrix  $M = \sum_{i=1}^{3m} w_i x_i x_i^\dagger$ . Rewriting  $M$  as  $\sum_{i=1}^m w_i (x_i x_i^\dagger + \sigma(x_i) \sigma(x_i)^\dagger + \sigma^2(x_i) \sigma^2(x_i)^\dagger)$ , we see that  $M$  is cyclic-symmetric, i.e., it is invariant under conjugation by the permutation  $\sigma$ . Of course  $M$  is also Hermitian. Combining these two properties, it must be of the form

$$M = \begin{pmatrix} r & s & \bar{s} \\ \bar{s} & r & s \\ s & \bar{s} & r \end{pmatrix}$$

for some  $r \in \mathbb{R}$  and  $s \in \mathbb{H}$ . The last condition in the proposition statement forces  $r = 1$  and  $s = 0$ , so in fact  $M = I_3$ .

Therefore,  $\{x_1, \dots, x_{3m}\}$  is a simplex with  $\sum_{i=1}^{3m} w_i x_i x_i^\dagger = I_3$ , and we complete the proof by applying Proposition 3.1.  $\square$

Applying the constraints in the above proposition with  $m = 4$ , we get a surjective Jacobian in Theorem I.2.1, which proves the following result.

**Theorem 3.5.** *There is a tight simplex of 12 points in  $\mathbb{H}\mathbb{P}^2$ . In fact, there is such a tight simplex with cyclic symmetry.*

Experimentally it appears that tight simplices with cyclic symmetry exist in other cases (e.g., 6- and 9-point simplices in  $\mathbb{H}\mathbb{P}^2$ ). In those cases we do not need to use the symmetry to establish the existence of tight simplices, though.

For 13-point simplices, we wish to follow a similar approach to bypass the rank-deficiency issue, but we must allow fixed points of the cyclic shift. In fact, there are cyclic-symmetric 13-point tight simplices consisting of 12 points with cyclic symmetry as above (i.e., four equivalence classes under the cyclic-shift operator) plus one extra point which is invariant under the cyclic-shift operator.

**Proposition 3.6.** *Let  $\sigma$  be the cyclic-shift automorphism  $\sigma(a, b, c) = (b, c, a)$ . Suppose  $x_1, \dots, x_{3m} \in \mathbb{H}^3$  satisfy the following conditions:*

- (a)  $x_{m+i} = \sigma(x_i)$  for  $i = 1, \dots, 2m$ ,
- (b)  $|x_i|^2 = 1$  for  $i = 1, \dots, m$ ,
- (c) the squared inner products  $|\langle x_i, x_j \rangle|^2$  for  $i = 1, \dots, m$  and the following values of  $j$  are all equal: (i)  $j = i + m$ , (ii)  $i < j \leq m$ , (iii)  $i + m < j \leq 2m$ , (iv)  $i + 2m < j \leq 3m$ , and
- (d) the 1,2 entry of the matrix  $\sum_{i=1}^{3m} x_i x_i^\dagger$  has real part  $1/6$  and magnitude  $1/3$ .

Then there is a unique point  $x_{3m+1} \in \mathbb{H}\mathbb{P}^2$  such that  $\{x_1, \dots, x_{3m}, x_{3m+1}\}$  is a tight simplex, and that point satisfies  $\sigma(x_{3m+1}) = x_{3m+1}$ .

*Proof.* A tight  $(3m + 1)$ -point simplex  $\{x_1, \dots, x_{3m+1}\}$  must satisfy  $\sum_{i=1}^{3m+1} x_i x_i^\dagger = \frac{3m+1}{3} I_3$ . Thus the matrix  $x_{3m+1} x_{3m+1}^\dagger$  is determined by the other data; since a point in projective space is determined by its projection matrix, this proves uniqueness. It also proves that, if such a point  $x_{3m+1}$  exists, then it must satisfy  $\sigma(x_{3m+1}) = x_{3m+1}$ ; this is because otherwise

$$\sigma(\{x_1, \dots, x_{3m}, x_{3m+1}\}) = \{x_1, \dots, x_{3m}, \sigma(x_{3m+1})\}$$

would be a distinct tight simplex.

Define  $M = \sum_{i=1}^{3m} x_i x_i^\dagger$ . This matrix is Hermitian and cyclic-symmetric, so as in the proof of Proposition 3.4 it is of the form

$$M = \begin{pmatrix} r & s & \bar{s} \\ \bar{s} & r & s \\ s & \bar{s} & r \end{pmatrix}$$

for some  $r \in \mathbb{R}$  and  $s \in \mathbb{H}$ . Each projection  $x_i x_i^\dagger$  has trace 1, so  $\text{Tr } M = 3m$ . Thus  $r = m$ , so  $\Pi := \frac{3m+1}{3} I_3 - M$  equals

$$\begin{pmatrix} 1/3 & -s & -\bar{s} \\ -\bar{s} & 1/3 & -s \\ -s & -\bar{s} & 1/3 \end{pmatrix}.$$

Being Hermitian and of trace 1,  $\Pi$  is a projection matrix of rank 1 iff  $3s^2 = -\bar{s}$ , as one can see by solving  $\Pi^2 = \Pi$ . The last hypothesis in the proposition statement implies that  $-3s$  is a cube root of unity in  $\mathbb{H}$ , from which we see that this condition is satisfied.

Let  $x_{3m+1} \in \mathbb{H}\mathbb{P}^2$  be the point satisfying  $\Pi = x_{3m+1}x_{3m+1}^\dagger$ . We know that  $\{x_1, \dots, x_{3m}\}$  is a regular simplex, as in Proposition 3.4. For  $i = 1, \dots, 3m$  define  $\Pi_i = x_i x_i^\dagger$  and let  $\alpha$  be the common inner product  $\langle \Pi_i, \Pi_j \rangle$  (for  $i, j \leq 3m$  with  $i \neq j$ ). We know

$$\sum_{i=1}^{3m+1} x_i x_i^\dagger = \frac{3m+1}{3} I_3, \quad (3.2)$$

by the definition of  $x_{3m+1}$ . Taking the inner product with  $\Pi_i$  for any  $i \leq 3m$ , we get

$$\frac{3m+1}{3} = \langle \Pi_i, \Pi_i \rangle + \langle \Pi_i, \Pi \rangle + \sum_{\substack{j \leq 3m \\ j \neq i}} \langle \Pi_i, \Pi_j \rangle = 1 + (3m-1)\alpha + \langle \Pi_i, \Pi \rangle.$$

Thus  $\langle \Pi_i, \Pi \rangle = (3m-2)/3 - (3m-1)\alpha$  for each  $i = 1, \dots, 3m$ . Similarly, taking the inner product of (3.2) with  $\Pi$  gives

$$\frac{3m+1}{3} = \langle \Pi, \Pi \rangle + \sum_{i \leq 3m} \langle \Pi, \Pi_i \rangle = 1 + 3m \cdot \left( \frac{3m-2}{3} - (3m-1)\alpha \right).$$

Solving this equation yields  $\alpha = \frac{3m-2}{9m} = \frac{N-3}{3(N-1)}$  (with  $N = 3m+1$ ). It follows that  $\langle \Pi, \Pi_i \rangle = \alpha$ , and that  $x_1, \dots, x_{3m+1}$  form a tight simplex.  $\square$

We get a surjective Jacobian when we apply the conditions of the above proposition in Theorem I.2.1 with  $m = 4$ , proving the following result.

**Theorem 3.7.** *There is a tight simplex of 13 points in  $\mathbb{H}\mathbb{P}^2$ . In fact, there is such a tight simplex with cyclic symmetry.*

Theorems 3.5 and 3.7 establish the existence of tight simplices, and their proof could also provide the dimension of the space of tight simplices with cyclic symmetry. They cannot, though, tell us the dimension of the full space of tight simplices.

If Proposition 3.1 had applied then we would have concluded that, in some neighborhood, the space of tight simplices of 12 (resp., 13) points in  $\mathbb{H}\mathbb{P}^2$  has dimension 7 (resp., 4). The observed rank deficiency of two has several possible explanations, including the following: it might mean that two of the constraints are redundant, so that the space of tight simplices is two dimensions larger than predicted; it might mean that the constraints become degenerate at the solutions, but the space of tight simplices is still a manifold of the predicted dimension; or it might mean that the space of tight simplices is not even locally a manifold. Based on numerical evidence (see §6.5), we conjecture that the first possibility holds.

**Conjecture 3.8.** *There exists a 12-point (resp., 13-point) tight simplex in  $\mathbb{H}\mathbb{P}^2$  such that, in a neighborhood thereof, the space of tight simplices has dimension 9 (resp., 6).*

### 3.3 15-point simplices

The case of 15 points in  $\mathbb{H}\mathbb{P}^2$  is special for a few reasons. First, it may be the only case in quaternionic projective spaces where the cardinality upper bound in Proposition 2.2

is achieved (beyond  $\mathbb{HP}^1$ , which is  $S^4$  and clearly contains a 6-point simplex). Also, in comparison with the other cases in Tables 3.1, this case has especially large rank deficiency. This suggests that the space of simplices is of a larger dimension than  $r(15, \mathbb{HP}^2)$ . That turns out to be correct, as we now show.

**Proposition 3.9.** *Suppose  $x_1, \dots, x_{15} \in \mathbb{H}^3$  satisfy*

$$\langle \Gamma_i, \Gamma_j \rangle = -\frac{1}{21} \quad \text{for } i \neq j,$$

where

$$\Gamma_i := x_i x_i^\dagger - \frac{1}{3} |x_i|^2 I_3.$$

Suppose additionally that  $|x_i|^4 \in [1 - 10^{-6}, 1 + 10^{-6}]$  for each  $i$ . Then  $|x_i| = 1$  and  $\{x_1, \dots, x_{15}\}$  is a tight simplex in  $\mathbb{HP}^2$ .

We do not think the assumption  $|x_i|^4 \in [1 - 10^{-6}, 1 + 10^{-6}]$  is necessary for the proposition to hold, but it is easy to verify in our applications and lets us prove the result with local calculations. More specifically, it lets us prove the result in a straightforward manner using interval arithmetic (see §I.3.2).

*Proof.* For each  $i$  write  $|x_i|^4 = 1 + \delta_i$ . It suffices to show  $\delta_i = 0$  for all  $i$ , because  $\{x_1, \dots, x_{15}\}$  is then a tight simplex. Specifically, define  $\eta_i = (1 + \delta_i)^{-1/2}$  and let  $\Pi_i = \frac{1}{|x_i|^2} x_i x_i^\dagger = \eta_i x_i x_i^\dagger$  denote the projection matrix associated to  $x_i$ . Then

$$\langle \Pi_i, \Pi_j \rangle = \eta_i \eta_j \langle \Gamma_i, \Gamma_j \rangle + \frac{1}{3} = \begin{cases} 1 & \text{if } i = j, \text{ and} \\ -\eta_i \eta_j / 21 + 1/3 & \text{if } i \neq j. \end{cases}$$

If  $\eta_i = 1$  for all  $i$ , then these inner products agree with the desired value  $2/7$  in a tight simplex of 15 points.

Our strategy is to show that nonnegativity of the second zonal harmonic sum forces  $\delta_i = 0$  for all  $i$ , given a rank condition coming from the fact that 15 equals the dimension of the space of Hermitian matrices.

Recall that the zonal harmonics on  $\mathbb{HP}^{d-1}$  are given by Jacobi polynomials  $P_k^{(2d-3,1)}(2t-1)$ . Specifically, the functions

$$K_k(x, y) = P_k^{(2d-3,1)}(2|\langle x, y \rangle|^2 - 1)$$

are positive definite kernels on  $\mathbb{HP}^{d-1}$ . Let  $\Sigma_k$  be the sum of the kernel  $K_k(x, y)$  over the projective code determined by  $\{x_1, \dots, x_{15}\}$ . Then positive definiteness implies  $\Sigma_k \geq 0$ .

We will require only  $\Sigma_2$ . As  $P_2(3,1)(2t-1) = 28t^2 - 21t + 3$ , we can write  $\Sigma_2$  in terms of the moments  $\sum_{i,j=1}^{15} \langle \Pi_i, \Pi_j \rangle^k$  with  $k \leq 2$ . These moments, in turn, we can write as functions of  $\delta_i$ . If  $\delta = 0$ , then  $\Sigma_2 = 0$ . Moreover, expanding to second order in  $\delta_1, \dots, \delta_{15}$ , a direct calculation shows that

$$\Sigma_2 = -\frac{10}{3} m_1 + \frac{23}{252} m_1^2 + \frac{719}{252} m_2 + O(\delta^3), \quad (3.3)$$

where  $m_1 := \sum_i \delta_i$  and  $m_2 := \sum_i \delta_i^2$ .

If  $\Sigma_2$  were locally a negative-definite function of  $\delta_1, \dots, \delta_{15}$ , then  $\Sigma_2 \geq 0$  would imply  $\delta_i = 0$ . However, the approximation in (3.3) is not negative-definite. To make it so, we must

add correction terms based on additional constraints satisfied by the perturbations  $\delta_i$ .

These additional constraints come from a singular Gram matrix. We have  $\langle \Gamma_i, \Gamma_i \rangle = \frac{2}{3}(1 + \delta_i)$ , and the Gram matrix of the elements  $\sqrt{\frac{2}{3}}\Gamma_i$  is

$$G = \begin{pmatrix} 1 + \delta_1 & & -\frac{1}{14} \\ & \ddots & \\ -\frac{1}{14} & & 1 + \delta_N \end{pmatrix}.$$

Each of  $\Gamma_1, \dots, \Gamma_{15}$  is a traceless Hermitian matrix, so they must be linearly dependent, because the space of such matrices has dimension 14. Thus, the Gram matrix  $G$  must be singular. Let  $D := 14^{14} \det(G)/15^{12}$  be its determinant (normalized as written); then  $D = 0$ . But consider writing  $D$  as a function of the  $\delta_i$ . A short computation shows that

$$\det \left( \text{diag}(a_1, \dots, a_n) + b \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix} \right) = b \sum_{i=1}^n \left( \prod_{j \neq i} a_j \right) + \left( \prod_{j=1}^n a_j \right).$$

Applying this with  $a_i = 1 + \delta + 1/14$  and  $b = -1/14$  gives us a reasonably compact expression for  $D$ . Using this, we can check that

$$D = 15m_1 + 14(m_1^2 - m_2) + O(\delta^3).$$

Because  $D$  (and so  $D^2$ ) must vanish and  $\Sigma_2$  must be nonnegative,

$$\Sigma'_2 := 4200D - 269D^2 + 18900\Sigma_2$$

must be nonnegative as well. However, from the above inequalities, we have

$$\Sigma'_2 = -4875m_2 + O(\delta^3),$$

and this is negative-definite. It remains negative-definite in a suitably small neighborhood of the point  $(\delta_1, \dots, \delta_{15}) = (0, \dots, 0)$ , and so if we know that the  $\delta_i$  lie in such a neighborhood, then we can conclude (as desired) that  $\delta_i = 0$  for all  $i$ .

It remains only to compute an explicit radius for such a neighborhood. This is routine using interval arithmetic. Specifically, one can compute intervals containing the entries of the Hessian for all  $\delta = (\delta_1, \dots, \delta_{15})$  with  $|\delta_i| \leq 10^{-6}$  for all  $i$ . (This is easy to implement in any computer algebra system with support for interval arithmetic; example MATHEMATICA code is available from the author by request.) This allows one to prove that the (squared) Frobenius norm of the difference between the Hessian at any such  $\delta$  and the Hessian at  $(0, \dots, 0)$  (which is  $-9750$  times the identity matrix) is bounded by  $8 \cdot 10^6$ . It follows that all of the eigenvalues of the Hessian at  $\delta$  differ from  $-9750$  by at most  $\sqrt{8 \cdot 10^6}$ . In particular, the eigenvalues are all negative. That completes the proof.  $\square$

Using this system of constraints, we do get a nonsingular Jacobian matrix and hence we can apply Theorem I.2.1. Proposition I.2.3 then yields a 75-dimensional solution space; after subtracting overcounting and symmetries, we arrive at the following.

**Theorem 3.10.** *There is a tight simplex of 15 points in  $\mathbb{HP}^2$ . In fact, locally there is a 9-dimensional space of such simplices, as opposed to the  $-5$  predicted by  $r(15, \mathbb{HP}^2)$ .*

Theorem 3.10 establishes the existence of a tight 2-design in  $\mathbb{H}\mathbb{P}^2$ . The common inner product in this simplex is  $2/7$ , contrary to a theorem of Bannai and Hoggar asserting that the inner products in tight designs are always reciprocals of integers [7, Corollary 1.7(b)]. The case of 2-designs is not addressed in their proof, and Bannai has informed us that this was an oversight. See also [33] for another correction (the icosahedron is a tight 5-design in  $\mathbb{C}\mathbb{P}^1$  with irrational inner products).

## 4 Simplices in $\mathbb{O}\mathbb{P}^2$

The study of simplices in  $\mathbb{O}\mathbb{P}^2$  bears a strong resemblance to that in  $\mathbb{H}\mathbb{P}^2$ ; we get essentially the same results as long as we take care to work in an affine chart. In particular, we can handle the generic case, 24- and 25-point simplices, and 27-point simplices using adaptations of Propositions 3.1, 3.4 and 3.6, and 3.9, respectively.

### 4.1 Generic case

**Proposition 4.1.** *For  $i = 1, \dots, N$ , suppose  $x_i = (a_i, b_i, c_i) \in \mathbb{R}_+ \times \mathbb{O}^2$  and  $w_i \in \mathbb{R}$  satisfy*

- (a)  $|a_i|^2 + |b_i|^2 + |c_i|^2 = 1$  for  $i = 1, \dots, N$ ,
- (b)  $\rho(x_i, x_j)^2 = \rho(x_{i'}, x_{j'})^2$  for  $1 \leq i < j \leq N$  and  $1 \leq i' < j' \leq N$ , and
- (c)  $\sum_{i=1}^N w_i \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} \begin{pmatrix} \bar{a}_i & \bar{b}_i & \bar{c}_i \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ .

Then  $w_1 = \dots = w_N = \frac{3}{N}$  and  $\{x_1, \dots, x_N\}$  is a tight simplex.

We omit the proof of Proposition 4.1 as it is nearly identical to that of Proposition 3.1.

We can attempt to apply Proposition 4.1 with Theorem I.2.1 and Proposition I.2.3, just as we did for simplices in quaternionic projective spaces. There are

$$N + \left( \frac{N(N-1)}{2} - 1 \right) + 27 \text{ real constraints} \quad \text{in} \quad 18N \text{ real variables,}$$

so, when the Jacobian is nonsingular, we get a solution space of dimension  $(N-1)(34-N)/2 - 9$ . As before, we should deduct the dimension of the symmetry group. The symmetry group of  $\mathbb{O}\mathbb{P}^2$  is the exceptional Lie group  $F_4$ , which has dimension 52. Thus, our final expression for the dimension of the space of simplices modulo symmetries is

$$r(N, \mathbb{O}\mathbb{P}^2) := \frac{(N-1)(34-N)}{2} - 61.$$

Again, as with  $r(N, \mathbb{H}\mathbb{P}^{d-1})$ , this formula only applies when, at our numerical solution, Proposition I.2.3 applies to the conditions of Proposition 4.1 and the simplex has 0-dimensional stabilizer.

**Theorem 4.2.** *For the values of  $N$  listed in part (a) of Table 4.1, there exists a tight  $N$ -point simplex in  $\mathbb{O}\mathbb{P}^2$ .*

Table 4.1: Cases in  $\mathbb{O}\mathbb{P}^2$ : (a) proven existence of tight simplices; (b) singular Jacobian; (c) conjectured nonexistence.

$N$	$r(N, \mathbb{O}\mathbb{P}^2)$	$N$	$r(N, \mathbb{O}\mathbb{P}^2)$	$N$	$r(N, \mathbb{O}\mathbb{P}^2)$
5	0 <sup>†</sup>	12	60	19	74
6	9	13	65	20	72
7	20	14	69	21	69
8	30	15	72	22	65
9	39	16	74	23	60
10	47	17	75		
11	54	18	75		

(a)

$N$	rank deficiency
24	2
25	2
27	26

(b)

$N$
26

(c)

<sup>†</sup> Actually  $r(5, \mathbb{O}\mathbb{P}^2)$  is not 0; rather, it equals  $-3$ . This is the only case in which the simplex we found has a positive-dimensional stabilizer. The stabilizer is 3-dimensional, so the actual dimension of the space of solutions modulo symmetries, which  $r(N, \mathbb{O}\mathbb{P}^2)$  is really intended to capture, is 0.

## 4.2 24- and 25-point simplices

The following proposition is proven similarly to Proposition 3.4.

**Proposition 4.3.** *Let  $\sigma$  be the cyclic-shift automorphism  $\sigma(a, b, c) = (b, c, a)$ . Suppose  $x_1, \dots, x_{3m} \in \mathbb{O}^3$  and  $w_1, \dots, w_{3m} \in \mathbb{R}$  satisfy the following conditions:*

- (a)  $x_{m+i} = \sigma(x_i)$  for  $i = 1, \dots, 2m$ ,
- (b)  $w_{m+i} = w_i$  for  $i = 1, \dots, 2m$ ,
- (c)  $x_i \in \mathbb{R}_+ \times \mathbb{O}^2$  and  $|x_i|^2 = 1$  for  $i = 1, \dots, m$ ,
- (d) the squared distances  $\rho(x_i, x_j)^2$  for  $i = 1, \dots, m$  and the following values of  $j$  are all equal: (i)  $j = i + m$ , (ii)  $i < j \leq m$ , (iii)  $i + m < j \leq 2m$ , (iv)  $i + 2m < j \leq 3m$ , and
- (e) the matrix  $\sum_{i=1}^{3m} w_i x_i x_i^\dagger$  has 1,1 entry equal to 1 and vanishing 1,2 entry.

Then  $w_1 = \dots = w_{3m} = \frac{1}{m}$  and  $\{x_1, \dots, x_{3m}\}$  is a tight simplex.

Using the conditions of Proposition 4.3 with  $m = 8$  in Theorem I.2.1 yields a surjective Jacobian, allowing us to prove the following theorem.

**Theorem 4.4.** *There is a tight simplex of 24 points in  $\mathbb{O}\mathbb{P}^2$ . In fact, there is such a tight simplex with cyclic symmetry.*

Similarly, to prove the existence of tight simplices with 25 points, we use the following adaptation of Proposition 3.6.

**Proposition 4.5.** *Let  $\sigma$  be the cyclic-shift automorphism  $\sigma(a, b, c) = (b, c, a)$ . Suppose  $x_1, \dots, x_{3m} \in \mathbb{O}^3$  satisfy the following conditions:*

- (a)  $x_{m+i} = \sigma(x_i)$  for  $i = 1, \dots, 2m$ ,
- (b)  $x_i \in \mathbb{R}_+ \times \mathbb{O}^2$  and  $|x_i|^2 = 1$  for  $i = 1, \dots, m$ ,
- (c) the squared distances  $\rho(x_i, x_j)^2$  for  $i = 1, \dots, m$  and the following values of  $j$  are all equal: (i)  $j = i + m$ , (ii)  $i < j \leq m$ , (iii)  $i + m < j \leq 2m$ , (iv)  $i + 2m < j \leq 3m$ , and
- (d) the 1,2 entry of the matrix  $\sum_{i=1}^{3m} x_i x_i^\dagger$  has real part  $1/6$  and magnitude  $1/3$ .

Then there is a unique point  $x_{3m+1} \in \mathbb{O}\mathbb{P}^2$  such that  $\{x_1, \dots, x_{3m}, x_{3m+1}\}$  is a tight simplex, and that point satisfies  $\sigma(x_{3m+1}) = x_{3m+1}$ .

Using the conditions above with  $m = 8$  in Theorem I.2.1 yields a surjective Jacobian.

**Theorem 4.6.** *There is a tight simplex of 25 points in  $\mathbb{O}\mathbb{P}^2$ . In fact, there is such a tight simplex with cyclic symmetry.*

Continuing the correspondence with 12- and 13-point simplices in  $\mathbb{H}\mathbb{P}^2$ , based on numerical evidence we conjecture the following.

**Conjecture 4.7.** *There exists a 24-point (resp., 25-point) tight simplex in  $\mathbb{O}\mathbb{P}^2$  such that, in a neighborhood thereof, the space of tight simplices has dimension 56 (resp., 49).*



### 4.3 27-point simplices

**Proposition 4.8.** *Suppose  $x_i = (a_i, b_i, c_i) \in \mathbb{R}_+ \times \mathbb{O}^2$  satisfy*

$$\langle \Gamma_i, \Gamma_j \rangle = -\frac{1}{39} \quad \text{for } i \neq j,$$

where

$$\Gamma_i := \begin{pmatrix} a_i \\ b_i \\ c_i \end{pmatrix} (\bar{a}_i \quad \bar{b}_i \quad \bar{c}_i) - \frac{1}{3}(a_i^2 + |b_i|^2 + |c_i|^2)I_3.$$

*Suppose additionally that  $|x_i|^4 \in [1 - 10^{-7}, 1 + 10^{-7}]$  for each  $i$ . Then  $|x_i| = 1$  and  $\{x_1, \dots, x_{27}\}$  determines a tight simplex in  $\mathbb{O}\mathbb{P}^2$ .*

*Proof.* We use the same proof technique as Proposition 3.9, with the only difference being the constants appearing in the proof. As before, we write  $|x_i|^4 = 1 + \delta_i$ , and let  $\delta = \max_i |\delta_i|$ . Let  $G$  be the Gram matrix of  $\sqrt{2/3}\Gamma_1, \dots, \sqrt{2/3}\Gamma_{27}$ . Then  $\det(G) = 0$ , and the normalized determinant  $D := 26^{26} \det(G)/27^{24}$  satisfies

$$D = 27m_1 + 26(m_1^2 - m_2) + O(\delta^3).$$

The second zonal harmonic on  $\mathbb{O}\mathbb{P}^2$  is given by the Jacobi polynomial

$$P_2^{(7,3)}(2t - 1) = 91t^2 - 65t + 10.$$

Let  $\Sigma_2$  be the sum of the kernel  $K_2(x, y) := P_2^{(7,3)}(2|\langle x, y \rangle|^2 - 1)$  over the projective code determined by  $\{x_1, \dots, x_{27}\}$ , so  $\Sigma_2 \geq 0$ . We compute

$$\Sigma_2 = -6m_1 + \frac{41}{468}m_1^2 + \frac{2429}{468}m_2 + O(\delta^3).$$

Because  $D = 0$ ,

$$\Sigma_2' := 75816D - 2745D^2 + 341172\Sigma_2$$

must be nonnegative. We have  $\Sigma_2' = -200475m_2 + O(\delta^3)$ , so in a neighborhood of  $(0, \dots, 0)$  the function  $\Sigma_2'$  is negative-definite. We just need to identify such a neighborhood. Using interval arithmetic as in Proposition 3.9, we find that the squared Frobenius norm of the difference between the Hessian at  $(0, \dots, 0)$  and the Hessian at any point  $\delta = (\delta_1, \dots, \delta_{27})$  with  $|\delta_i| \leq 10^{-7}$  for all  $i$  is bounded by  $4 \cdot 10^9$ . Thus the eigenvalues of the Hessian at any such  $\delta$  differ from  $-400950$  by at most  $\sqrt{4 \cdot 10^9}$ . In particular, the eigenvalues are all negative, from whence the proposition follows.  $\square$

Applying Theorem I.2.1 with the conditions of the above proposition, we find a suitable point for which the Jacobian is surjective.

**Theorem 4.9.** *There is a tight simplex of 27 points in  $\mathbb{O}\mathbb{P}^2$ . In fact, there is a 56-dimensional space of such simplices.*

Theorem 4.9 establishes the existence of a tight 2-design in  $\mathbb{O}\mathbb{P}^2$ . Such designs were previously conjectured not to exist [23, p. 251]. It is known [25] that tight  $t$ -designs in  $\mathbb{O}\mathbb{P}^2$  can only exist for  $t = 2$  and  $t = 5$ , and there is an explicit construction of a 819-point tight 5-design [12], so Theorem 4.9 completes the list of  $t$  for which tight  $t$ -designs exist in  $\mathbb{O}\mathbb{P}^2$ .

## 5 Simplices in Grassmannians $G(m, n, \mathbb{R})$

Our techniques also apply to show the existence of many simplices on Grassmannian spaces. The Grassmannian  $G(m, n) = G(m, n, \mathbb{R})$  is the space of all  $m$ -dimensional subspaces of  $\mathbb{R}^n$ . It is a homogeneous space for the orthogonal group  $O(n)$ , isomorphic to  $O(n)/(O(m) \times O(n-m))$ , and it has dimension  $m(n-m)$ . These spaces generalize (real) projective space  $\mathbb{R}P^{n-1}$ , which is the space of lines in  $\mathbb{R}^n$ . The spaces  $G(m, n)$  and  $G(n-m, n)$  can be identified by associating to each subspace its orthogonal complement, so in what follows we always assume  $m \leq n/2$ .

Though Grassmannians are generally not 2-point homogeneous spaces, it is still possible to find linear programming bounds [2]. Here we will just consider the special case of the simplex bound.

When  $m \leq n/2$ , a pair of points in  $G(m, n)$  is described by  $m$  parameters, namely the *principal angles* between the  $m$ -dimensional subspaces. Given two  $m$ -dimensional subspaces  $U$  and  $U'$ , define sequences of unit vectors  $u_1, \dots, u_m \in U$  and  $u'_1, \dots, u'_m \in U'$  inductively so that  $\langle u_i, u'_i \rangle$  is maximized subject to  $\langle u_i, u_j \rangle = \langle u'_i, u'_j \rangle = 0$  for  $j < i$ . Then the principal angles are  $\theta_i := \arccos \langle u_i, u'_i \rangle$ .

The *chordal distance* on  $G(m, n)$  is given by

$$d_c(U, U') = \sqrt{\sin^2 \theta_1 + \dots + \sin^2 \theta_m}.$$

Unlike in projective space, the chordal metric on Grassmannians is generally not equivalent to the geodesic metric  $\sqrt{\theta_1^2 + \dots + \theta_m^2}$ . See [17] for discussion of why the chordal metric is preferable.

A *generator matrix* for an element of  $G(m, n)$  is a  $m \times n$  matrix whose rows form an orthonormal basis of the subspace. Given a generator matrix  $X$ , the orthogonal projection onto the subspace is  $X^t X$ . Suppose  $X_1$  and  $X_2$  are generator matrices for the subspaces  $U_1$  and  $U_2$ , and let  $\Pi_i = X_i^t X_i$  (for  $i = 1, 2$ ) be the orthogonal projection matrices. Then the singular values of the matrix  $X_1 X_2^t$  are  $\cos \theta_i$  for  $1 \leq i \leq m$ . It follows that

$$d_c(U_1, U_2)^2 = \frac{1}{2} \|\Pi_1 - \Pi_2\|^2 = m - \langle \Pi_1, \Pi_2 \rangle. \quad (5.1)$$

Let  $\Pi^0 = \Pi - (m/n)I_n$  be the traceless part of the projection matrix. It is easily checked that  $\|\Pi^0\|^2 = m(n-m)/n$ . Thus  $\Pi^0$  can be thought of as a point in  $\mathbb{R}^D$ , where  $D = m(m+1)/2 - 1$ , if we view  $\mathbb{R}^D$  as the space of trace zero symmetric matrices. This gives an isometric embedding  $U \mapsto \Pi^0$  of  $G(m, n)$  into the sphere of radius  $\sqrt{m(n-m)/n}$  in  $\mathbb{R}^D$ , with  $d_c(U_1, U_2) = \|\Pi_1^0 - \Pi_2^0\|/\sqrt{2}$ . The simplex bound for spherical codes gives us the following result.

**Proposition 5.1** (Conway, Hardin, and Sloan [17]). *Every  $N$ -point simplex in  $G(m, n)$  satisfies*

$$N \leq \binom{m+1}{2},$$

*and every code of  $N$  points has squared chordal distance at most*

$$\frac{m(n-m)}{n} \cdot \frac{N}{N-1}.$$

This squared chordal distance is equivalent to inner product  $\frac{m(Nm-n)}{n(N-1)}$  between orthogonal

projection matrices.

*Remark 5.2.* The  $m = 1$  case of Proposition 5.1 is the same as the  $K = \mathbb{R}$  case of Proposition 2.2 (together with Proposition 2.4). Indeed, the proofs of these two results are essentially the same; they are just phrased in different language.

We say that a simplex in  $G(m, n)$  is tight if its minimal chordal distance meets the upper bound above. Analogously to simplices in projective space, a Grassmannian simplex is tight iff it is a 1-design (i.e., a 2-design in the language of Bachoc, Coulangéon, and Nebe [3]), which holds iff the linear programming bound is sharp [2]. If the projection matrices of the simplex are  $\Pi_1, \dots, \Pi_N$ , then another equivalent condition for tightness is  $\sum_{i=1}^N \Pi_i = \frac{Nm}{n} I_n$ .

Conway, Hardin, and Sloane [17] reported a number of putative tight simplices based on numerical evidence, but except for some explicit constructions they did not present any techniques for rigorous existence proofs. The cases with explicit constructions are listed in Table 5.1. By applying our methods, we can certify the existence of simplices for many of the cases previously identified but not settled.

**Proposition 5.3.** *Suppose  $\{x_{i,j} \in \mathbb{R}^n\}_{\substack{i=1,\dots,N \\ j=1,\dots,m}}$  and  $w_1, \dots, w_N$  satisfy the following conditions:*

- (a)  $|x_{i,j}| = 1$  for all  $i, j$ ,
- (b) for all  $i$  and all  $j < j'$ ,  $\langle x_{i,j}, x_{i,j'} \rangle = 0$ ,
- (c) the inner products  $\langle \sum_{j=1}^m x_{i,j} x_{i,j}^t, \sum_{j=1}^m x_{i',j} x_{i',j}^t \rangle$  are equal for all distinct pairs  $i, i'$ , and
- (d)  $\sum_{i=1}^N w_i \left( \sum_{j=1}^m x_{i,j} x_{i,j}^t \right) = I_n$ .

Then  $w_1 = \dots = w_N = \frac{n}{Nm}$  and the subspaces  $\text{span}\{x_{i,1}, \dots, x_{i,m}\}$  form a tight simplex in  $G(m, n)$ .

*Proof.* For each  $i$ , define  $\Pi_i = \sum_{j=1}^m x_{i,j} x_{i,j}^t$ . Because  $\{x_{i,1}, \dots, x_{i,m}\}$  is an orthonormal system, this is the projection matrix associated to the plane  $\text{span}\{x_{i,1}, \dots, x_{i,m}\}$ . Using (5.1), the third condition guarantees that we have a simplex. Arguing as in the proof of Proposition 3.1, we deduce from the last condition that  $w_1 = \dots = w_N = \frac{n}{Nm}$ . Thus  $\sum_{i=1}^N \Pi_i = \frac{Nm}{n} I_n$ ; as noted above, this is equivalent to tightness.  $\square$

In many cases the system specified by Proposition 5.3 is nonsingular, allowing us to apply Theorem I.2.1. This yields the following.

**Theorem 5.4.** *For the values of  $(N, m, n)$  listed in the “proven” column of Table 5.2, there exists a tight  $N$ -point simplex in  $G(m, n, \mathbb{R})$ .*

In the context of Proposition 5.3, we have  $Nmn + N$  real variables and

$$N \cdot \binom{m+1}{2} + \left( \frac{N(N-1)}{2} - 1 \right) + \binom{n+1}{2}$$

real constraints. Thus, when Proposition I.2.3 applies, we locally get a solution space whose dimension is the difference of these counts. Because  $O(m)$  acts on the different representations of each plane, we are overcounting the dimension by  $N \cdot \binom{m}{2}$ . Moreover, when the symmetry group  $O(n)$  of  $G(m, n)$  acts with finite stabilizer on the simplex, we should

Table 5.1: Previously known tight simplices with explicit constructions in  $G(m, n)$  for  $n \leq 8$

$(m, n)$	$N$	Reference
(2, 4)	2–6	[17, pp. 145–146]
(2, 4)	10	[17, p. 147]
(2, 6)	9	[17, p. 154]
(3, 7)	28	[17, p. 152]
(2, 8)	8	[17, p. 154]
(2, 8)	20	[11, p. 135]
(2, 8)	28	[17, p. 154]

Table 5.2: Tight Grassmannian simplices in  $G(m, n)$ .

$(m, n)$	Proven	Singular Jacobian	Ambiguous
(2, 4)	4–6	2, 3, 7, 8, 10	
(2, 5)	5–10	4, 11	
(2, 6)	5–14	3, 4	
(3, 6)	5–16	2–4	17
(2, 7)	6–17		18
(3, 7)	5–22	4, 28	23
(2, 8)	6–21	4, 5, 28	
(3, 8)	5–28	4	
(4, 8)	5–30	2–4	

deduct  $\binom{n}{2}$  from our final dimension count. Putting this all together, when these conditions are satisfied (c.f. Remark 3.2), we get a neighborhood in which the space of simplices has dimension

$$r(N, G(m, n)) := Nm n - \frac{N(N-3)}{2} - Nm^2 - n^2 + 1. \quad (5.2)$$

We tested all cases up to dimension  $n = 8$ , using our own software to search for numerical solutions and also comparing with the numerical results of Conway, Hardin, and Sloane [17]. As with simplices in projective spaces, sometimes the system specified by Proposition 5.3 was singular, and sometimes the numerical evidence was unclear (cf. Tables 3.1 and 3.3). These cases are in the third and fourth columns, respectively, of Table 5.2.

In addition to our existence proofs and the previously known explicit constructions, several Grassmannian tight simplices can be proven to exist using the following observation: if there is a tight  $N$ -point simplex in  $G(m, n)$  for some  $m, n$ , then there is a tight  $N$ -point simplex in  $G(km, kn)$  for all  $k \geq 1$ . This is immediate from block repetition [19, Proposition 12]. It proves existence for 11 of the singular cases in Table 5.2. This leaves us with only 7 hitherto unresolved cases in which there is strong numerical evidence for a tight simplex: 4-point simplices in  $G(2, 5)$ ,  $G(3, 6)$ ,  $G(3, 7)$ , and  $G(3, 8)$ ; 7- and 8-point simplices in  $G(2, 4)$ ; and 11-point simplices in  $G(2, 5)$ . For completeness, we will settle all of these in the following subsection.

There should be no difficulty in applying our techniques to complex or quaternionic Grassmannians, but we have not done so.

## 5.1 Miscellaneous special cases in Grassmannians

We begin with the case of 11-point tight simplices in  $G(2, 5)$ . This can be handled in the same way as 13-point tight simplices in  $\mathbb{H}\mathbb{P}^2$  and 25-point tight simplices in  $\mathbb{O}\mathbb{P}^2$ ; i.e., we can prove existence of simplices with cyclic symmetry. We will state the analogous result in greater generality than we attempted in Proposition 3.6 (which was written in the special case of  $\mathbb{H}\mathbb{P}^2$  rather than a general projective space  $\mathbb{H}\mathbb{P}^{d-1}$ ), at the cost of some additional complexity.

**Proposition 5.5.** *Fix dimensions  $n \geq m$  and let  $\sigma$  be the cyclic-shift automorphism  $\sigma(x_1, x_2, \dots, x_n) = (x_2, \dots, x_n, x_1)$  on  $\mathbb{R}^n$ . Set  $N = nk + 1$  and suppose we have vectors  $\{x_{i,j} \in \mathbb{R}^n\}_{i=1, \dots, nk, j=1, \dots, m}$ . For each  $i$ , define  $\Pi_i = \sum_{j=1}^m x_{i,j} x_{i,j}^t$ . Define  $\Pi_N = \frac{Nm}{n} I_n - \sum_{i < N} \Pi_i$ . Suppose that, for some  $\eta \in (\frac{m}{m+1}, \frac{m}{m-1})$ , the following conditions are satisfied:*

- (a)  $x_{k+i,j} = \sigma(x_{i,j})$  for all  $i \leq (n-1)k$  and all  $j$ ,
- (b)  $|x_{i,j}| = 1$  for all  $i \leq k$  and all  $j$ ,
- (c) for all  $i \leq k$  and all  $j < j'$ ,  $\langle x_{i,j}, x_{i,j'} \rangle = 0$ ,
- (d) the inner products  $\langle \Pi_i, \Pi_{i'} \rangle$  are all equal for (i)  $i \leq k$ ,  $i' = i + qk$ , and  $q = 1, \dots, \lfloor \frac{n}{2} \rfloor$  and (ii)  $i \leq k-1$ ,  $i' = i'_0 + qk$ ,  $i < i'_0 \leq k$ , and  $q = 0, \dots, n-1$ , and
- (e) the first  $\lfloor \frac{n}{2} \rfloor + 1$  entries in the first row of  $\Pi_N^2 - \eta \Pi_N$  are all zero.

Then  $\eta = 1$ ,  $\Pi_N$  is a projection matrix of rank  $m$ , and the projection matrices  $\{\Pi_i\}_{i < N}$  determine a tight  $N$ -point simplex in  $G(m, n)$ .

*Proof.* The automorphism  $\sigma$  on  $\mathbb{R}^n$  determines an automorphism on  $G(m, n)$  by acting simultaneously on basis vectors, and this latter automorphism is an isometry. The first condition states that the planes spanned by  $\{x_{i,1}, \dots, x_{i,m}\}$  and  $\{x_{k+i,1}, \dots, x_{k+i,m}\}$  are related by this isometry; thus, taking all  $i < N$ , we have  $k$  orbits under the cyclic-shift action, each of size  $n$ . The next two conditions ensure that the matrices  $\Pi_i$  for  $i < N$  are orthogonal projections of rank  $m$ . Thus the inner products amongst them determine distances in  $G(m, n)$ . Now, by applying the cyclic-shift isometry we see that the fourth condition is sufficient to force  $\{\Pi_i\}_{i < N}$  to determine a regular simplex. Let  $\alpha = \langle \Pi_i, \Pi_{i'} \rangle$  be the common inner product.

Consider now the matrix  $\Pi_N$ . It is symmetric, being a linear combination of symmetric matrices. Moreover, it is cyclic-symmetric, since  $\sum_{i < N} \Pi_i$  is a sum over orbits of the cyclic shift. It follows that  $\Pi_N^2 - \eta \Pi_N$  also shares these properties. Now a matrix with cyclic-symmetry is determined by its first row, as the other rows are just shifts thereof. A matrix which is also symmetric is determined by the first  $\lfloor \frac{n}{2} \rfloor + 1$  entries in the first row. Therefore, by the last condition,  $\Pi_N^2 - \eta \Pi_N = 0$ .

It follows that the eigenvalues of  $\Pi_N$  are all either 0 or  $\eta$ . Let  $r$  be the rank of  $\Pi_N$ , so that  $\text{Tr } \Pi_N = r\eta$ . But, since  $\text{Tr } \Pi_i = m$  for all  $i < N$ , we have  $\text{Tr } \Pi_N = m$ . Hence  $\eta = m/r$  is  $m$  times the reciprocal of an integer. The assumption  $\eta \in (\frac{m}{m+1}, \frac{m}{m-1})$  then forces  $\eta = 1$ , from which we conclude that  $\Pi_N$  is an orthogonal projection matrix of rank  $m$ .

$$\begin{aligned} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} & \frac{1}{5} \begin{pmatrix} -\sqrt{3} & 2 & \sqrt{6} & 2\sqrt{3} & 0 \\ 0 & \sqrt{3} & -\sqrt{2} & 0 & \sqrt{20} \end{pmatrix} \\ \\ & \frac{1}{5} \begin{pmatrix} 3 & 0 & 0 & 4 & 0 \\ 0 & 1 & -2\sqrt{6} & 0 & 0 \end{pmatrix} & \frac{1}{5} \begin{pmatrix} \sqrt{3} & 2 & \sqrt{6} & -2\sqrt{3} & 0 \\ 0 & -\sqrt{3} & \sqrt{2} & 0 & \sqrt{20} \end{pmatrix} \end{aligned}$$

Figure 5.1: Generator matrices for a tight 4-point simplex in  $G(2, 5)$ .

It remains only to check that  $\langle \Pi_i, \Pi_N \rangle = \alpha$  for all  $i < N$  (so that  $\{\Pi_i\}_{i \leq N}$  is an  $N$ -point simplex) and that  $\alpha = \frac{m(Nm-n)}{n(N-1)}$ , so that the simplex is tight. This follows from the same argument as in Proposition 3.6. Namely, using the definition of  $\Pi_N$  we compute

$$\langle \Pi_i, \Pi_N \rangle = \frac{Nm}{n} \langle \Pi_i, I_n \rangle - \langle \Pi_i, \Pi_i \rangle - \sum_{i' \neq i} \langle \Pi_i, \Pi_{i'} \rangle = \frac{Nm^2}{n} - m - (N-2)\alpha.$$

Expanding  $\frac{Nm^2}{n} = \langle \Pi_N, \frac{Nm}{n} I_n \rangle = \sum_{i \leq N} \langle \Pi_N, \Pi_i \rangle$  then gives

$$\frac{Nm^2}{n} = m + (N-1) \left( \frac{Nm^2}{n} - m - (N-2)\alpha \right),$$

and solving this equation gives the desired conclusion.  $\square$

Note that the plane with projection matrix  $\Pi_N$  is the unique plane completing  $\{\Pi_i\}_{i < N}$  into a tight simplex. This plane is a fixed point for the cyclic-shift action.

In our case of interest we found a point in which the conditions described in Proposition 5.5 are nonsingular. This yields the following.

**Theorem 5.6.** *There exists a tight 11-point simplex in  $G(2, 5)$ . In fact, there is such a tight simplex with cyclic symmetry.*

We remark in passing that every approximate 11-point tight simplex in  $G(2, 5)$  we found numerically exhibited a symmetry group conjugate to the cyclic-symmetry discussed here. With this evidence as well as the fact that  $r(11, G(2, 5)) = -2 < 0$ , we conjecture that every tight 11-point simplex in  $G(2, 5)$  has a nontrivial symmetry group.

We will settle the remaining cases with algebraic constructions. The four cases of 4-point simplices afford constructions using only rationals and quadratic irrationals, so we give them explicitly here. Given the provided matrices, the proof of the following theorem consists only of a straightforward calculation.

**Theorem 5.7.** *The four  $2 \times 5$  matrices in Figure 5.1 are generator matrices whose corresponding planes form a tight simplex in  $G(2, 5)$ , i.e., they have orthonormal rows and the spans of those rows constitute a tight simplex. Similarly, the matrices in Figures 5.2, 5.3, and 5.4 determine tight simplices in  $G(3, 6)$ ,  $G(3, 7)$ , and  $G(3, 8)$ , respectively.*

We are now left with the cases of 7- and 8-point tight simplices in  $G(2, 4)$ . These cases are more interesting; the simplest explicit coordinates we have been able to find for them require algebraic numbers of degree 4 and 6, respectively. Because of this, instead of presenting the algebraic numbers here we rely on a computer algebra system to (rigorously) verify the calculation. The computational method is discussed in §6.4. Here we simply record the result.

$$\begin{array}{cc}
\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} & \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \sqrt{2} \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix} \\
\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & \sqrt{2} & 0 & 0 \end{pmatrix} & \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & \sqrt{2} & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}
\end{array}$$

Figure 5.2: Generator matrices for a tight 4-point simplex in  $G(3, 6)$ .

$$\begin{array}{cc}
\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} & \frac{1}{7} \begin{pmatrix} 0 & -2\sqrt{2} & 3 & 2\sqrt{3} & 2\sqrt{5} & 0 & 0 \\ -\sqrt{5} & 0 & -2\sqrt{2} & \sqrt{6} & 0 & \sqrt{30} & 0 \\ 0 & -\sqrt{5} & 0 & 0 & -\sqrt{2} & 0 & \sqrt{42} \end{pmatrix} \\
\frac{1}{7} \begin{pmatrix} 5 & 0 & 0 & 0 & 0 & 2\sqrt{6} & 0 \\ 0 & 3 & 0 & 0 & 2\sqrt{10} & 0 & 0 \\ 0 & 0 & -1 & 4\sqrt{3} & 0 & 0 & 0 \end{pmatrix} & \frac{1}{7} \begin{pmatrix} 0 & 2\sqrt{2} & 3 & 2\sqrt{3} & -2\sqrt{5} & 0 & 0 \\ \sqrt{5} & 0 & -2\sqrt{2} & \sqrt{6} & 0 & -\sqrt{30} & 0 \\ 0 & \sqrt{5} & 0 & 0 & \sqrt{2} & 0 & \sqrt{42} \end{pmatrix}
\end{array}$$

Figure 5.3: Generator matrices for a tight 4-point simplex in  $G(3, 7)$ .

$$\begin{array}{cc}
\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} & \frac{1}{2} \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & -\sqrt{3} \\ 0 & 1 & 0 & 0 & 0 & -\sqrt{3} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \sqrt{3} & 0 \end{pmatrix} \\
\frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 & \sqrt{3} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \sqrt{3} & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \end{pmatrix} & \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 & -\sqrt{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \sqrt{3} \\ 0 & 0 & 1 & 0 & 0 & 0 & -\sqrt{3} & 0 \end{pmatrix}
\end{array}$$

Figure 5.4: Generator matrices for a tight 4-point simplex in  $G(3, 8)$ .

**Theorem 5.8.** *There exist 7- and 8-point tight simplices in  $G(2, 4)$ .*

We remark in passing that  $G(2, 4)$  contains tight simplices of  $N$  points for all  $N \leq 10$  (the theoretical maximum) except for  $N = 9$ . Compared with the other spaces studied in this paper, only the quaternionic and octonionic projective planes have such a wealth of simplices. Note also that there does not seem to exist a tight simplex of size one less than the upper bound in any of these spaces.

## 6 Algorithms and Computational Methods

We used computer assistance in several different aspects of this work. Our main results involve two different computational steps: finding approximate solutions and then rigorously proving existence of a nearby solution. We also require a method for computing with real algebraic numbers for Theorem 5.8, and we must discuss how to compute stabilizers of simplices and estimate the dimensions of solution spaces. This section describes the algorithms and programs used for each of these tasks.

### 6.1 Rigorous proof

Our existence proofs rely on Theorem I.2.1. To apply the theorem, we need to choose  $\varepsilon > 0$ , the starting point  $x_0$ , and a matrix  $T$  and then compute the operator norms of  $T$  and  $Df(x_0) \circ T - \text{id}_{\mathbb{R}^n}$ . We use the  $\ell^\infty$  norm on the domain and codomain. In addition to being amenable to the computation of operator norms (see §I.3.1), this norm allows balls to be naturally represented using interval arithmetic (see §I.3.2). In all of the proofs given in this chapter, we choose  $\varepsilon = 10^{-9}$ . Thus the conclusion of Theorem I.2.1 is that there is an exact solution, each of whose coordinates differs from our starting point  $x_0$  by less than  $10^{-9}$ . In other words, the error is less than one nanounit.

To check the hypotheses rigorously, we used a new computer package called QNEWTON, written by the author. It can be obtained from him upon request. QNEWTON is a C++ library with a PYTHON front end for convenience; it was designed to find and prove the existence of solutions to polynomial equations over real algebras. QNEWTON has native support for multiplication in  $\mathbb{R}$ ,  $\mathbb{C}$ ,  $\mathbb{H}$ , and  $\mathbb{O}$ . Also, it uses automatic (reverse) differentiation to compute the Jacobian of the constraint function. These two features substantially increase its performance, so that it is practical even for large problems (on the order of  $10^3$  variables/constraints).

The finding part of this mission statement is discussed in §6.2. As to the proving part, the basic algorithm is clear. We first rigorously compute  $f(x_0)$ , which is straightforward using interval arithmetic. Then we compute the Jacobian matrix, again using interval arithmetic; moreover, before we do so, we expand the intervals containing our starting point  $x_0$  by  $\varepsilon$ . Thus, for each entry we get an interval that contains the corresponding entry of  $Df(x)$  for every  $x \in B(x_0, \varepsilon)$ . We then compute an interval guaranteed to contain  $\|Df(x) \circ T - \text{id}_{\mathbb{R}^n}\|$  for all such  $x$ , and an interval guaranteed to contain  $1 - \|T\| \cdot |f(x_0)|/\varepsilon$ . If the upper bound of the first interval is less than the lower bound of the second interval, then we are assured that Theorem I.2.1 applies.

We should also remark upon the computation of the matrix  $T$ . It is supposed to be approximately a right inverse of  $Df(x_0)$ , but otherwise we are free in choosing it. Internally in QNEWTON we compute  $T$  in the following manner. First we compute the matrix  $Df(x_0)$  approximately, using floating-point arithmetic. Then we find its pseudo-inverse (i.e., the



least-squares right inverse) again using inexact floating-point arithmetic. Finally, we take the result, which is a floating-point number, and replace it with an interval of width 0. This approach is fast and, since  $T$  need not be exact, still gives rigorous results. It is possible to compute  $Df(x_0)$  in interval arithmetic and then compute the pseudo-inverse in the same way; this is a bad idea, though, because inverting a matrix in interval arithmetic is both slow and can result in very large intervals.

Finally, another issue that arises in the course of rigorous proof is that, in three cases (Propositions 3.9, 4.8, and 5.5), we require certain quantities to be close to 1. For example, in Proposition 3.9 we need  $||x_i|^4 - 1|$  to be at most  $10^{-6}$  for each  $i$ . This could easily be checked by direct computation using the  $10^{-9}$  bound for distance from the starting point, but it is simpler to use the following approach. For each  $i$ , we add a new variable  $v_i$ , add a new constraint  $v_i = |x_i|^4$ , and initialize  $v_i$  to be 1 at the starting point. Then we can conclude that  $||x_i|^4 - 1| < 10^{-9}$  in the exact solution with no additional computation.

## 6.2 Finding approximate solutions

The QNEWTON package also has some support for finding approximation solutions. The computations involved in rigorous proof of existence — in particular, computing the Jacobian matrix — are also the computations needed for Newton’s method. (In fact, exploiting the templating feature of C++, QNEWTON shares the same computation code between rigorous, interval arithmetic computations and approximate, floating-point computations.) After we specify the polynomials and constraints for the problem and an initial point, QNEWTON attempts to find a solution using a damped Newton’s method algorithm. Newton’s method converges rapidly (quadratically) in a neighborhood of a solution, but it is ill-behaved away from solutions; thus we damp the steps so that no coordinate changes in a single step by more than a specified upper bound. The damping is just a bound on the maximum step size; if that bound is violated then we scale the step proportionally.

In our computations, we used random Gaussian variables for the initial points and a maximum step size of 0.1. Because our variables represent unit vectors, the step size is approximately one order of magnitude less than the natural scale. By using this approach we were able to find a solution in all cases in which we think there should exist one, using just a few different random starting positions. In most cases we found a solution on the first try. These approximate calculations use double-precision floating-point arithmetic, so we can only expect convergence up to an error of approximately  $10^{-15}$ . In all cases this was more than sufficient for our goals of rigorous proof.

Suppose that, as in Theorem I.2.1, we are solving for a zero of a function  $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$ . Newton’s method calls for taking repeatedly taking steps  $\Delta x$  satisfying  $Df(x) \cdot \Delta x = -f(x)$ . In particular, we must repeatedly solve linear systems. When  $m > n$  the system is underdetermined. Also,  $Df(x)$  may fail to be surjective. Hence we need a linear solver tolerant of such problems. QNEWTON uses a *least-squares* solver that treats small singular values of  $Df(x)$  as zero; specifically, it uses the DGELSD function in LAPACK [1]. By using such a solver we can handle cases with redundant constraints. This was particularly useful when we were first determining a minimal set of constraints for our problems.

Because the codes we seek are energy minimizers, another approach to find them would have been gradient descent. In practice, we have found that gradient descent is much slower than Newton’s method.

For a quick example of the usage of QNEWTON, Figure 6.1 shows a set of PYTHON commands to find a set of standard generators for  $\mathbb{H}$ , i.e., anticommuting elements  $a$  and  $b$  of

```

q = HNewton.Newton()
q.parse_block("""
CONSTRAIN Rx(a)
CONSTRAIN Rx(b)
y(len_a_minus_1) := x(a)~ x(a) - 1
CONSTRAIN Ry(len_a_minus_1)
y(len_b_minus_1) := x(b)~ x(b) - 1
CONSTRAIN Ry(len_b_minus_1)
y(ab) := x(a)~ x(b)
CONSTRAIN Ry(ab)
""")
q.initialize_variable('a', [0.1,0.2,0.3,0.4])
q.initialize_variable('b', [0.5,0.6,0.7,0.8])
q.Newton_steps(15)

```

Figure 6.1: A PYTHON session finding a set of generators for  $\mathbb{H}$ .

norm 1 with zero real part. (Because  $\operatorname{Re} a = \operatorname{Re} b = 0$ , the condition  $ab = -ba$  is equivalent to  $\operatorname{Re} \bar{a}b = 0$ .) Note that the starting point we use is far from any solution.

### 6.3 Finding stabilizers

In all but one case, namely 5-point simplices in  $\mathbb{O}\mathbb{P}^2$ , our reported dimension for the space of tight simplices has the dimension of the full symmetry group deducted. That assumes that the simplex has finite (i.e., zero-dimensional) stabilizer. We checked this by (i) finding a basis for the Lie algebra of the symmetry group, (ii) applying each element of that basis to the points of the simplex, and (iii) checking that the resulting vectors are linearly independent. As described below, we used floating point arithmetic, so these calculations are not rigorous.

The relevant symmetry groups are  $\operatorname{Sp}(d)/\{\pm 1\}$  for  $\mathbb{H}\mathbb{P}^{d-1}$  and  $F_4$  for  $\mathbb{O}\mathbb{P}^2$ . The Lie algebra of  $\operatorname{Sp}(d)$  has dimension  $2d^2 + d$ . Viewed as a space of operators on the Jordan algebra of Hermitian matrices, it consists of commutation with skew-Hermitian matrices. The Lie algebra of  $F_4$  has dimension 52; it consists of commutation with traceless skew-Hermitian matrices and of derivations of the algebra  $\mathbb{O}$  (see [5]).

For each element of our basis for the Lie algebra, we applied that element to the projection matrices corresponding to the  $N$  points in our approximation to the simplex. We then concatenated the resulting  $N$  Hermitian matrices to form one vector. Finally, we used a singular value decomposition (cf. §6.5) to compute the dimension of the span of those vectors. In every case except for 5 points in  $\mathbb{O}\mathbb{P}^2$  the resulting vectors had full-dimensional span; in that case the dimensionality of the span was 3 smaller.

As yet this procedure is not fully rigorous, but it would not be difficult to make it so. This is because, being an open condition, proving a lower bound on the dimension of the span of a set of vectors can be done with approximate computations. In all but the one exceptional case we are asserting linear independence; this could be verified by using interval arithmetic to compute a bound, valid for all codes within  $10^{-10}$  of the given code, on the determinant of a full square submatrix. Since the true simplex lies in this set, that bound would apply to it (and to a neighborhood thereof).

The same methods could show that, for the 5-point simplex in  $\mathbb{O}\mathbb{P}^2$ , the stabilizer has

dimension at most 3. This is good enough because, translated into a dimension for the space of simplices, that bound says that the dimension is at most 0; hence the dimension must equal 0.

*Remark 6.1.* Based on similar numerical evidence, we conjecture that (modulo symmetries) the space of SIC-POVMs, simplices of  $n^2$  points in  $\mathbb{C}\mathbb{P}^{n-1}$ , has dimension 1 when  $n = 3$  and 0 when  $n \geq 4$ . In particular, we conjecture that, except in  $\mathbb{C}\mathbb{P}^2$ , SIC-POVMs are isolated. This is in accordance with the numerical results in [38], although they searched only for SIC-POVMs that are invariant under the Weyl-Heisenberg group.

## 6.4 Real algebraic numbers

To verify equations involving algebraic numbers of moderately high degree (in particular, higher degree than we care to manipulate by hand), we require a computational method for rigorously doing basic arithmetic with such numbers. We will use the standard approach of “isolating intervals,” which is implemented in many modern computer algebra systems. We used the PARI/GP package because it is freely available and has support for arbitrary-precision rational numbers [35]. There is no explicit support for the isolating interval method in PARI/GP, so we provide a short implementation in addition to the pertinent data files for our applications.

The technique is as follows. A real algebraic number  $\alpha$  is represented by a triple  $(p(x), \ell, u)$ , where  $p(x)$  is a polynomial with integer coefficients such that  $p(\alpha) = 0$ ,  $\ell$  and  $u$  are rational numbers such that  $\alpha \in [\ell, u]$ , and  $p(x)$  has a unique root in the interval  $[\ell, u]$  (namely,  $\alpha$ ). We always take  $p(x)$  to be (a scalar multiple of) the minimal polynomial of  $\alpha$ , and we use Sturm sequences to rigorously count the number of real roots in a given interval. Given representations  $(p_\alpha, \ell_\alpha, u_\alpha)$  and  $(p_\beta, \ell_\beta, u_\beta)$  for two real algebraic numbers  $\alpha, \beta$ , we compute a representation for  $\alpha + \beta$  by first taking the resultant, in the variable  $t$ , of the polynomials  $p_\alpha(t)$  and  $p_\beta(x - t)$ . This gives a polynomial in  $x$  for which  $\alpha + \beta$  is a root. We then factor the resulting polynomial and count the number of roots for each irreducible factor in the interval  $[\ell_\alpha + \ell_\beta, u_\alpha + u_\beta]$ . If there is more than one factor that has a root in that interval or some factor has multiple roots, then we bisect the starting intervals  $[\ell_\alpha, u_\alpha]$  and  $[\ell_\beta, u_\beta]$ , using Sturm sequences for  $p_\alpha$  and  $p_\beta$  to choose the halves containing  $\alpha$  and  $\beta$ , respectively. After a finite number of steps we are left with a valid representation for  $\alpha + \beta$ . Computing a representation for  $\alpha \cdot \beta$  proceeds similarly, beginning with the resultant of  $p_\alpha(t)$  and  $t^{\deg p_\beta} p_\beta(x/t)$ .

Using this system, we can now elucidate the proof of existence for 7- and 8-point tight simplices in  $G(2, 4)$ .

*Proof of Theorem 5.8.* Our data files provide isolating interval representations for the entries of the  $4 \times 4$  projection matrices  $\{\Pi_i\}_{i \leq N}$  for the  $N = 7$  or 8 points in each simplex. To verify the construction we need only perform a few calculations. First we need to check that each provided matrix  $\Pi$  satisfies  $\Pi = \Pi^t$ ,  $\Pi^2 = \Pi$ , and  $\text{Tr } \Pi = 2$ , as together these conditions imply that  $\Pi$  is an orthogonal projection onto a plane. Then we just need to verify that  $\text{Tr } \Pi_i \Pi_j = \frac{N-2}{N-1}$  for  $i < j \leq N$ . These calculations are straightforward given our implementation of the isolating interval method.  $\square$

## 6.5 Estimating dimensions

In Conjectures 3.8, 4.7, and 7.4, we conjecture the dimension of certain solution spaces; here we describe the basis for those conjectures.

Suppose, as is the case in our examples, that we are studying the zero set  $Z$  of some function  $f$ . Suppose moreover that we have a procedure for converging to zeros of  $f$ , using, for example, Newton’s method with least-squares solving to handle degeneracy. Thus we have the ability to generate points on  $Z$ , and we wish to use that ability to calculate its dimension. This is a simple instance of *manifold learning*, the problem of describing a manifold given sample points embedded in some higher-dimensional space.

For our purposes we use following heuristic. Fix  $\varepsilon > 0$ . Starting with a solution  $x_0$ , we compute  $N$  nearby solutions  $x_1, \dots, x_N$  by, for each  $i$ , setting  $x'_i = x_0 + \varepsilon g_i$ , where  $g_i$  is a vector of standard normal random variables, and using our iterative solver to find a zero  $x_i$  of  $f$  near  $x'_i$ . Then, to first order in  $\varepsilon$ , the vectors  $\frac{x_i - x_0}{|x_i - x_0|}$  are random (normalized) samples from the tangent space of  $Z$  at  $x_0$ . We then form the matrix whose rows are those  $N$  vectors and compute its singular values. There should be  $d$  singular values of order approximately 1, where  $d$  is the dimension of  $Z$ . The remaining singular values should be smaller by a factor of  $\varepsilon$ .

This procedure is certainly not rigorous, but in suitably nice cases, and with proper choice of parameters, one can have a fair amount of confidence in the result. In particular,  $N$  should be at least as large as the dimension  $d$  and  $\varepsilon$  should be chosen small enough that, in a ball of radius  $\varepsilon$ ,  $Z$  is well-approximated by its tangent space. One pitfall to avoid is that, while  $\varepsilon$  needs to be small for the tangent space approximation, it should also be large enough that the precision of the solver is better than (approximately)  $\varepsilon^2$ . If this is violated then we may erroneously identify extra null vectors of  $Df(x_0)$  as elements of the tangent space.

In our applications we used  $N = 1000$  and  $\varepsilon = 10^{-3}$  and we required that Newton’s method converge to within  $10^{-12}$ . It was usually easy to identify the jump in singular values after the  $d$  corresponding to the tangent space. For instance, Conjecture 3.8 says that, before accounting for overcounting and symmetries, we conjecture a 66-dimensional space of 12-point tight simplices in  $\mathbb{H}\mathbb{P}^2$ . This is based on the following observation: when we ran the procedure just discussed, the first 66 singular values were all in the interval  $[2, 6]$ , but the 67<sup>th</sup> was 0.04139564.

## 7 Explicit Constructions

With the exception of Theorems 5.7 and 5.8, all of the new results we have presented so far involve computer-assisted proofs using Theorem I.2.1. This allowed us to sidestep explicit constructions, and it also gave local dimensions as a collateral benefit. When an explicit construction is available, though, it can sometimes give insight not proffered by a general existence theorem. We conclude the paper with a few examples of this.

### 7.1 Two universal optima in $\text{SO}(4)$

Most results in the literature concerning universal optima in continuous spaces are set in two-point homogeneous spaces, i.e., spheres and projective spaces. We have already seen another family of spaces (namely, real Grassmannians) but there are many others.

Consider the special orthogonal group  $\text{SO}(n)$ , endowed with the chordal distance  $d_c(U_1, U_2) = \|U_1 - U_2\|$  coming from the embedding  $\text{SO}(n) \hookrightarrow \mathbb{R}^{n^2}$  as  $n \times n$  matrices (i.e., the Frobenius norm). This is not the Killing metric, but rather it has the advantage of being smooth in the matrix entries. Note that every element of  $\text{SO}(n)$  has norm  $n$ , so up to this scaling factor we have an embedding into  $S^{n^2-1}$ .

By a *universally optimal* code in  $\text{SO}(n)$ , we mean a code that minimizes energy for every completely monotonic function of squared chordal distance (see [14]). In this section we present two particularly attractive universal optima in  $\text{SO}(4)$ .

**Theorem 7.1.** *There is a 17-point code in  $\text{SO}(4)$  with the following properties: it is a regular simplex, it is universally optimal, and it has a transitive symmetry group. Moreover, there is no larger regular simplex in  $\text{SO}(4)$ .*

*Proof.* Given  $a, b \in \mathbb{Z}/17\mathbb{Z}$ , define the rotation matrix

$$R_{a,b} = \begin{pmatrix} \cos(a\theta) & -\sin(a\theta) & 0 & 0 \\ \sin(a\theta) & \cos(a\theta) & 0 & 0 \\ 0 & 0 & \cos(b\theta) & -\sin(b\theta) \\ 0 & 0 & \sin(b\theta) & \cos(b\theta) \end{pmatrix},$$

where  $\theta = 2\pi/17$ . For any  $a, b, c, d$ , not all zero, the map  $\sigma_{a,b,c,d}: \text{SO}(4) \rightarrow \text{SO}(4)$  defined by  $X \mapsto R_{a,b}XR_{c,d}$  is an isometry of  $\text{SO}(4)$  of order 17. Set

$$X_0 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \in \text{SO}(4)$$

and let  $\{X_i = R_{1,3}^i X_0 R_{4,5}^i\} \in \text{SO}(4)$  be the orbit of  $X_0$  under  $\sigma_{1,3,4,5}$ . This is a 17-point code which, by construction, has a transitive symmetry group. Moreover, direct calculation shows that it forms a regular simplex.

By virtue of the Euclidean embedding  $\text{SO}(4) \hookrightarrow S^{15}$ , there can be no regular simplices of more than 17 points, and a 17-point regular simplex must be universally optimal (indeed, it is even universally optimal as a code on the sphere). That proves the remaining claims of the theorem.  $\square$

**Theorem 7.2.** *There is a 32-point code in  $\text{SO}(4)$  with the following properties: it is a subgroup, it is universally optimal, and it forms the vertices of a cross-polytope in  $S^{15}$ .*

*Proof.* The code consists of all matrices of the form

$$\begin{pmatrix} a & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \end{pmatrix}, \begin{pmatrix} 0 & a & 0 & 0 \\ b & 0 & 0 & 0 \\ 0 & 0 & 0 & c \\ 0 & 0 & d & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & a & 0 \\ 0 & 0 & 0 & b \\ c & 0 & 0 & 0 \\ 0 & d & 0 & 0 \end{pmatrix}, \text{ or } \begin{pmatrix} 0 & 0 & 0 & a \\ 0 & 0 & b & 0 \\ 0 & c & 0 & 0 \\ d & 0 & 0 & 0 \end{pmatrix},$$

where  $a, b, c, d = \pm 1$  with an even number of  $-1$ 's. In other words, we use signed permutation matrices in which the underlying permutation is either trivial or a product of disjoint 2-cycles and the number of minus signs is even. It is not difficult to check that this defines a subgroup of  $\text{SO}(4)$ .

The supports of these four types of matrices are disjoint, so the corresponding points in  $\mathbb{R}^{16}$  are orthogonal. The inner product between two matrices of the same type is simply the inner product of the vectors  $(a, b, c, d)$ , which is 0 or  $\pm 4$  because of the even number of  $-1$ 's. Thus, the code forms a cross polytope in  $S^{15}$ .

As in Theorem 7.1, universal optimality of  $\mathcal{C}$  in  $\text{SO}(4)$  follows from universal optimality as a subset of  $S^{15}$  (see [14]).  $\square$

## 7.2 39 points in $\mathbb{O}\mathbb{P}^2$

**Theorem 7.3.** *There exists a tight code  $\mathcal{C}$  of 39 points in  $\mathbb{O}\mathbb{P}^2$ . It consists of 13 orthogonal triples such that, for any two points  $x_i, x_j$  in distinct triples,  $\rho(x_i, x_j) = \sqrt{2/3}$ . In other words, if  $\Pi, \Pi'$  are the projection matrices corresponding to two distinct points in  $\mathcal{C}$ , then  $\langle \Pi, \Pi' \rangle$  equals 0 if the two points are in the same triple and otherwise equals  $1/3$ .*

*Proof.* First we recall from [18, p. 127] the standard construction of a 12-point universal optimum in  $\mathbb{C}\mathbb{P}^2$ : in terms of unit-length representatives, it consists of the standard basis

$$(1, 0, 0), (0, 1, 0), (0, 0, 1)$$

together with the 9 points

$$\frac{1}{\sqrt{3}}(1, \omega^a, \omega^b), \tag{7.1}$$

where  $\omega = e^{2\pi i/3}$  and  $a, b = 0, 1, 2$ .

To construct the desired code, we will use the standard basis together with four rotated copies of (7.1). More precisely, let  $\{1, i, j, k\}$  be the standard basis of  $\mathbb{H}$  and let  $\ell$  be any one of the remaining four standard basis elements of  $\mathbb{O}$ . We identify  $\omega \in \mathbb{C}$  as an element of  $\text{span}\{1, i\} \subset \mathbb{O}$ . Set  $n = j\ell$ . Then we define  $\mathcal{C} \subset \mathbb{O}\mathbb{P}^2$  to be the code obtained from the standard basis and the points

$$\begin{aligned} (1, \omega^a, \omega^b)/\sqrt{3}, & \quad (1, \omega^a j, \omega^b \ell)/\sqrt{3}, \\ (1, \omega^a \ell, \omega^b n)/\sqrt{3}, & \quad (1, \omega^a n, \omega^b j)/\sqrt{3} \end{aligned} \tag{7.2}$$

for  $a, b = 0, 1, 2$ . Direct computation shows that this code has the desired distances. Specifically, we must split the code into 13 distinguished triples of points. The standard basis yields one such triple, and each of the four types of points in (7.2) yields three triples according to the value of  $a + b$  modulo 3.

The sum over  $\mathcal{C}$  of the first and second harmonics

$$\begin{aligned} P_1^{(7,3)}(2t-1) &= 12t - 4, \\ P_2^{(7,3)}(2t-1) &= 91t^2 - 65t + 10 \end{aligned}$$

of  $\mathbb{O}\mathbb{P}^2$  both vanish; thus  $\mathcal{C}$  is a 2-design. As it has only two inner products between distinct points, and one of those is 0, it is tight [32] and in fact universally optimal [14].  $\square$

The code  $\mathcal{C}$  in Theorem 7.3 is a system of mutually unbiased bases. It follows easily from linear programming bounds that it is the largest such system possible.

This code is not unique: we can deform it to a four-dimensional family of tight codes by replacing  $\ell, n, n$ , and  $j$  in the second line of (7.2) with  $\xi_1 \ell, \xi_2 n, \xi_3 n$ , and  $\xi_4 j$ , where  $\xi_1, \dots, \xi_4$  are complex numbers of absolute value 1. The group of isometries of  $\mathbb{O}\mathbb{P}^2$  fixing the remaining 21 unchanged points is zero-dimensional, so we have a four-dimensional family even modulo the action of the isometry group  $F_4$  of  $\mathbb{O}\mathbb{P}^2$ . We think the actual space of tight codes is much larger, though. On the basis of numerical evidence (see §6.5), we conjecture the following.

**Conjecture 7.4.** *In a neighborhood of the code constructed in (7.2), the space of tight 39-point codes, modulo the action of  $F_4$ , is a manifold of dimension 24.*

At present this remains just a conjecture, though, as we have been unable to identify a nonsingular system of equations to which we can apply Theorem I.2.1.

The existence of a code of this form was conjectured by Hoggar [24, Table 2] after classifying the permissible parameters for strongly regular graphs. Excepting a hypothetical 26-point tight simplex, which we conjecture does not exist, there are no remaining cases in which the existence of a tight code in  $\mathbb{O}\mathbb{P}^2$  is conjectured but not resolved. In fact, based on computations of optimal quasicodes (two-point correlation functions subject to linear programming bounds [16]), we are confident there are no other tight codes in  $\mathbb{O}\mathbb{P}^2$  with at most  $10^4$  points. We believe there are no more of any size.





# Bibliography

- [1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK users' guide*, third ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.
- [2] Christine Bachoc, *Linear programming bounds for codes in Grassmannian spaces*, IEEE Trans. Inform. Theory **52** (2006), no. 5, 2111–2125. MR 2234468 (2007h:94095)
- [3] Christine Bachoc, Renaud Coulangeon, and Gabriele Nebe, *Designs in Grassmannian spaces and lattices*, J. Algebraic Combin. **16** (2002), no. 1, 5–19. MR 1941981 (2003m:05045)
- [4] Christine Bachoc and Frank Vallentin, *New upper bounds for kissing numbers from semidefinite programming*, J. Amer. Math. Soc. **21** (2008), no. 3, 909–924. MR 2393433 (2009c:52029)
- [5] John C. Baez, *The octonions*, Bull. Amer. Math. Soc. (N.S.) **39** (2002), no. 2, 145–205. MR 1886087 (2003f:17003)
- [6] Eiichi Bannai and Stuart G. Hoggar, *On tight  $t$ -designs in compact symmetric spaces of rank one*, Proc. Japan Acad. Ser. A Math. Sci. **61** (1985), no. 3, 78–82. MR 796472 (87b:05040)
- [7] ———, *Tight  $t$ -designs and squarefree integers*, European J. Combin. **10** (1989), no. 2, 113–135. MR 988506 (90d:05055)
- [8] Marcel Berger, *A panoramic view of Riemannian geometry*, Springer-Verlag, Berlin, 2003. MR 2002701 (2004h:53001)
- [9] A. V. Bondarenko, *On a spherical code in the space of spherical harmonics*, Ukrainian Math. J. **62** (2010), no. 6, 993–996. MR 2888653
- [10] Ulrich Brehm and Wolfgang Kühnel, *15-vertex triangulations of an 8-manifold*, Math. Ann. **294** (1992), no. 1, 167–193. MR 1180457 (94e:57033)
- [11] Arthur R. Calderbank, Ronald H. Hardin, Eric M. Rains, Peter W. Shor, and Neil J. A. Sloane, *A group-theoretic framework for the construction of packings in Grassmannian spaces*, J. Algebraic Combin. **9** (1999), no. 2, 129–140. MR 1679247 (2000e:51015)
- [12] Arjeh M. Cohen, *Exceptional presentations of three generalized hexagons of order 2*, J. Combin. Theory Ser. A **35** (1983), no. 1, 79–88. MR 704257 (84f:51038)
- [13] Henry Cohn, *Order and disorder in energy minimization*, Proceedings of the International Congress of Mathematicians. Volume IV (New Delhi), Hindustan Book Agency, 2010, pp. 2416–2443. MR 2827978 (2012k:05082)
- [14] Henry Cohn and Abhinav Kumar, *Universally optimal distribution of points on spheres*, J. Amer. Math. Soc. **20** (2007), no. 1, 99–148. MR 2257398 (2007h:52009)
- [15] Henry Cohn, Abhinav Kumar, and Gregory Minton, *Optimal simplices and codes in projective spaces*, arXiv:1308.3188, 2013.

- [16] Henry Cohn and Yufei Zhao, *Energy-minimizing error-correcting codes*, arXiv:1212.1913, 2012.
- [17] John H. Conway, Ronald H. Hardin, and Neil J. A. Sloane, *Packing lines, planes, etc.: packings in Grassmannian spaces*, Experiment. Math. **5** (1996), no. 2, 139–159. MR 1418961 (98a:52029)
- [18] H. S. M. Coxeter, *Regular complex polytopes*, first ed., Cambridge University Press, London, 1974. MR 0370328 (51 #6555)
- [19] Jean Creignou, *Constructions of Grassmannian simplices*, arXiv:cs/0703036, 2007.
- [20] P. Delsarte, J. M. Goethals, and J. J. Seidel, *Spherical codes and designs*, Geometriae Dedicata **6** (1977), no. 3, 363–388. MR 0485471 (58 #5302)
- [21] Noam D. Elkies and Benedict H. Gross, *The exceptional cone and the Leech lattice*, Internat. Math. Res. Notices (1996), no. 14, 665–698. MR 1411589 (97g:11070)
- [22] H. Hadwiger, *Über ausgezeichnete Vektorsterne und reguläre Polytope*, Comment. Math. Helv. **13** (1940), 90–107 (German). MR 0003718 (2,260d)
- [23] Stuart G. Hoggar, *t-designs in projective spaces*, European J. Combin. **3** (1982), no. 3, 233–254. MR 679208 (85b:05052)
- [24] ———, *Parameters of t-designs in  $\mathbf{F}P^{d-1}$* , European J. Combin. **5** (1984), no. 1, 29–36. MR 746042 (85m:05027)
- [25] ———, *Tight t-designs and octonions*, Mitt. Math. Sem. Giessen (1984), no. 165, 1–16. MR 745865 (85i:05062)
- [26] G. A. Kabatiansky and V. I. Levenshtein, *Bounds for packings on a sphere and in space*, Prob. Inf. Transm. **14** (1978), no. 1, 1–17, English translation from Russian. MR 0514023 (58 #24018)
- [27] William M. Kantor, *Quaternionic line-sets and quaternionic Kerdock codes*, Linear Algebra Appl. **226/228** (1995), 749–779. MR 1344596 (97b:94033)
- [28] Mahdad Khatirinejad, *Regular structures of lines in complex spaces*, Ph.D. thesis, Simon Fraser University, 2008, <http://summit.sfu.ca/item/9188>.
- [29] Hermann König, *Cubature formulas on spheres*, Advances in multivariate approximation (Witten-Bommerholz, 1998), Math. Res., vol. 107, Wiley-VCH, Berlin, 1999, pp. 201–211. MR 1797231 (2002f:65036)
- [30] P. W. H. Lemmens and J. J. Seidel, *Equiangular lines*, J. Algebra **24** (1973), 494–512. MR 0307969 (46 #7084)
- [31] V. I. Levenshtein, *Bounds on the maximal cardinality of a code with bounded modulus of the inner product*, Soviet Math. Dokl. **25** (1982), no. 2, 526–531, English translation from Russian.
- [32] ———, *Designs as maximum codes in polynomial metric spaces*, Acta Appl. Math. **29** (1992), no. 1-2, 1–82, Interactions between algebra and combinatorics. MR 1192833 (93j:05012)
- [33] Yu. I. Lyubich, *On tight projective designs*, Des. Codes Cryptogr. **51** (2009), no. 1, 21–31. MR 2480685 (2010b:05041)
- [34] Arnold Neumaier, *Combinatorial configurations in terms of distances*, Memorandum 81-09 (Wiskunde), TH Eindhoven, <http://solon.cma.univie.ac.at/scan/combcon.pdf>, 1981.
- [35] *PARI/GP, version 2.6.0*, Bordeaux, 2013, <http://pari.math.u-bordeaux.fr/>.
- [36] Joseph M. Renes, *Equiangular tight frames from Paley tournaments*, Linear Algebra Appl. **426** (2007), no. 2-3, 497–501. MR 2350673 (2008j:42008)

- [37] Joseph M. Renes, Robin Blume-Kohout, A. J. Scott, and Carlton M. Caves, *Symmetric informationally complete quantum measurements*, J. Math. Phys. **45** (2004), no. 6, 2171–2180. MR 2059685 (2004m:81043)
- [38] A. J. Scott and M. Grassl, *Symmetric informationally complete positive-operator-valued measures: a new computer study*, J. Math. Phys. **51** (2010), no. 4, 042203, 16. MR 2662471 (2011f:81022)
- [39] Mátyás A. Sustik, Joel A. Tropp, Inderjit S. Dhillon, and Robert W. Heath, Jr., *On the existence of equiangular tight frames*, Linear Algebra Appl. **426** (2007), no. 2-3, 619–635. MR 2350682 (2008f:15066)
- [40] J. H. van Lint and J. J. Seidel, *Equilateral point sets in elliptic geometry*, Nederl. Akad. Wetensch. Proc. Ser. A 69=Indag. Math. **28** (1966), 335–348. MR 0200799 (34 #685)
- [41] Lloyd R. Welch, *Lower bounds on the maximum cross correlation of signals*, IEEE Trans. Information Theory **20** (1974), no. 3, 397–399.
- [42] William K. Wootters and Brian D. Fields, *Optimal state-determination by mutually unbiased measurements*, Ann. Physics **191** (1989), no. 2, 363–381. MR 1003014 (90e:81019)
- [43] Pengfei Xia, Shengli Zhou, and Georgios B. Giannakis, *Achieving the Welch bound with difference sets*, IEEE Trans. Inform. Theory **51** (2005), no. 5, 1900–1907. MR 2235693 (2007b:94148a)
- [44] Gerhard Zauner, *Quantum designs: foundations of a noncommutative design theory*, Int. J. Quantum Inf. **9** (2011), no. 1, 445–507. MR 2931102



# Chapter III

## Gravitational Choreographies

Gravitational  $n$ -body choreographies are particularly beautiful configurations of  $n$  equal-mass point particles, interacting through Newtonian gravity. In this chapter we present an application of computer-assisted proof techniques to the problem of existence of choreographies. The plan is as follows. In §1 we introduce the problem and survey previous results from the literature. We then set up the problem in formal terms in §2, and in §3 we summarize our results. In §4, we present the technical details of our proof technique. We elaborate in §5 on some issues arising in the efficient software implementation of this approach. In §6 we compare our techniques with previous work. Finally, we close in §7 with some questions for further study.

### 1 Background

The study of gravity has an illustrious history. Using Tycho Brahe’s data, Johannes Kepler formulated his three laws describing in detail the elliptical orbits of the planets [60, 59]. Isaac Newton, inspired by Kepler’s work, formulated his “law of universal gravitation,” stating that massive bodies attract each other with a force proportional to the inverse-square of their distance [75]. Using his development of differential calculus, Newton deduced the elliptical nature of orbits in the Solar System as a mathematical consequence of his law. In fact, he solved the 2-body system entirely, showing that the only solutions for two point masses, attracting each other through the inverse-square force law, are conic sections [43].

With the 2-body problem solved, research progressed to studying the  $n$ -body problem with  $n \geq 3$ . This proved significantly harder than the 2-body problem, and in fact remains today “one of the oldest unsolved problems in the exact sciences” [45, p. xiv]. One seminal work on the subject is Henri Poincaré’s submission for King Oscar II’s prize, which eventually expanded into a three-volume tome [77]. Rather than providing a solution of the 3-body problem, which is what the prize solicited, Poincaré’s work showed that a complete solution in the sense of the solution of the 2-body problem was impossible [30].

*Aside.* This does not rule out solutions in other, weaker senses, however. For instance, Sundman’s infinite series can be viewed as a solution of the 3-body problem, albeit an impractical one [11, pp. 190–191]. One estimate for the number of terms in the series needed for astronomical accuracy is  $10^{8000000}$  [92, p. 39]. This is captured in quotes of George Birkhoff; he wrote of Sundman’s series that it is “one of the most remarkable contributions to the problem of three bodies which has ever been made” — but speaking to its practical usefulness, he said “[u]nfortunately these series are valueless” [11, p. 191].

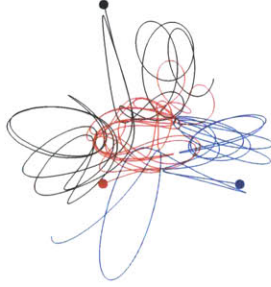


Figure 1.1: An approximate solution of the “Burrau problem” (also known as the “Pythagorean problem”), a particular set of simple initial conditions for the 3-body problem.

Some sense of the difficulty of the problem is captured by Figure 1.1, a depiction of the trajectories for a particular 3-body problem with a simple set of initial conditions. Despite the simplicity of the starting configuration, the resulting orbit is quite complicated.

### 1.1 Periodic orbits

One approach that Poincaré suggested was to focus on special solutions that we can better understand, instead of attacking the general problem. In particular, he championed the study of periodic solutions:

D’ailleurs, ce qui nous rend ces solutions périodiques si précieuses, c’est qu’elles sont, pour ainsi dire, la seule brèche par où nous puissions essayer de pénétrer dans une place jusqu’ici réputée inabordable [77, §36].

What makes periodic orbits so valuable is that they are the only breach, so to speak, through which we can try to enter a place up to now deemed unapproachable [32, p. 18].

The simplest example of a periodic orbit in the  $n$ -body problem is the “circular orbit” in which  $n$  equal-mass bodies form a regular  $n$ -gon rotating with constant angular velocity. It is a simple exercise to write down such orbits explicitly. In the 3-body case, this generalizes to a periodic solution for any set of masses [92, §5.7]. More generally, there are periodic *homographic* motions with *central configuration* in the plane; these are solutions in which the configuration of the bodies remains constant, up to scaling and rotation [2]. For instance, there is a homographic motion of the equal-mass 4-body problem in which three of the bodies form an isosceles triangle and the fourth is inside the triangle, on the axis of symmetry [1].

### 1.2 Choreographies

Rather than studying these, however, we shall study special cases of periodic orbits which have a remarkable property: all  $n$  masses actually follow the *same curve*. More precisely, we use the following definition.

**Definition 1.1.** An  $n$ -body *choreography* is a periodic solution of the  $n$ -body problem in which the  $n$  bodies have equal mass and they all follow the same curve, spaced equally in time. That is, if  $x(t)$  denotes the position function for one of the bodies and  $T$  is its period, then at all times  $t$  the  $n$  bodies are at positions  $\{x(t), x(t + \frac{1}{n}T), \dots, x(t + \frac{n-1}{n}T)\}$ .

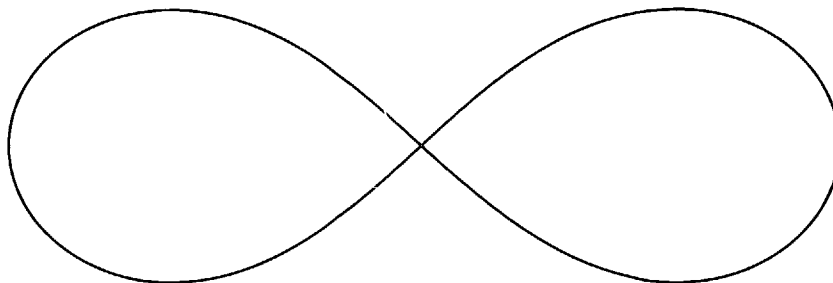


Figure 1.2: The figure-eight 3-body choreography.

We wish to emphasize that, throughout this document, the  $n$ -body problem refers to the problem of  $n$  point masses under Newtonian gravity. We do not consider relativistic effects (except for a brief remark in §7). When we refer to a *solution* of the  $n$ -body problem, we implicitly mean a nonsingular solution. That means in particular that there are no collisions, which is the relevant statement for our work.

*Aside.* In addition to collisions, there is one other type of singularity to rule out: the size of the system could diverge *in finite time* [95]. Surprisingly, this exotic type of singularity *can* occur in Newtonian gravity [40, 98].

*Remark.* Choreographies are defined as *spatial*, i.e., three-dimensional orbits. However, here we will study choreographies that are *planar*, i.e., two-dimensional. There is no mathematical reason to impose planarity *a priori*, but aside from a few scattered remarks we will put off discussion of non-planar orbits until a later paper.

The term “choreography” for these orbits was coined by Carles Simó [84]. He referred to orbits satisfying our definition as “simple choreographies.” They are sometimes referred to as “absolute choreographies,” as opposed to “relative choreographies” where the periodic motion takes place in a rotating frame. Note also that some sources do not explicitly impose the equal-mass or the equal-time-spacing properties in the definition (see Problem 7.3).

In the remainder of this section we give an overview of the literature on choreographies. For more information, the reader is referred to Susanna Terracini’s comprehensive survey article [91].

### 1.3 Figure-eight

The circular orbit is clearly a choreography, but it is not a terribly interesting one. The existence of nontrivial choreographies was implicit in the 1983 work of Davies et al. [29], but it was first explicitly highlighted in a 1993 paper by Cris Moore [71]. Moore numerically discovered the exemplar of choreographies: the figure-eight orbit depicted in Figure 1.2.

Prior to Moore’s work, the existence of such simple but nontrivial choreographies had not been anticipated. In addition to being surprising and novel, the figure-eight orbit has several fine qualities with which to recommend itself. It is planar and possesses 12-fold symmetry (see §1.4). It has zero angular momentum, in stark contrast to circular orbits (which require nonzero angular momentum in order to avoid collisions). Moreover, and perhaps most interestingly, it has been proven to be (linearly) *stable* [85, 56, 57]. While many more choreographies have been found since the figure-eight, its stability remains exceptional. No other (planar)  $n$ -body choreography ( $n \geq 3$ ), including the circular orbit,

has been proven to be stable. In fact, to our knowledge, there is only one other planar choreography which has even been conjectured to be stable (see §3.6).

*Aside.* The stability of the figure-eight raises the possibility of such a configuration actually existing naturally in the universe. Douglas Heggie carried out simulations to estimate the likelihood of binary–binary scattering (encounters between two pairs of bodies) yielding a figure-eight orbit [44]. Based on this, he reportedly estimated that the density of figure-eights in nature should be between one per galaxy and one per universe [87, p. 147]. Applications to physics, in particular in the form of gravitational waves, have been considered by Chiba et al. [25].

## 1.4 Variational proofs

In a celebrated 2000 paper, Alain Chenciner and Richard Montgomery rediscovered the figure-eight orbit and proved its existence [23]. Their proof relied on the variational formulation of the physical equations of motion, which we now recall.

Given a physical system, let  $T$  be its total kinetic energy and let  $V$  be its total potential energy. These are functions of time. Conservation of energy dictates that the total energy  $H := T + V$  is constant over time. Instead of focusing on  $H$ , though, consider the *Lagrangian*  $L := T - V$ , and its time integral, the *action*  $S := \int L dt$ . The action is a functional, i.e., a real-valued function of trajectory functions. The importance of the action functional is expressed in the following principle, often attributed to Hamilton.

**Principle of Stationary Action.** *The physical trajectories are the stationary points of the action.*

*Remark.* We have been intentionally vague in the statement of this principle. In addition to the types of physical systems to which this applies, we are especially leaving unspecified the limits of integration in the definition of the action and the ambient function space over which the action is supposed to be stationary. These points will be elaborated upon in §2.1 and §4.1, respectively.

The principle of stationary action was inspired by Fermat’s principle of least time, which describes the path that light takes through media of varying refraction index. It, in turn, inspired Feynman’s path integral formulation of quantum mechanics. Viewing classical physical systems through the lens of the action principle falls under the heading of *Lagrangian mechanics*. It is also referred to as the *variational approach*.

For choreographic purposes, the primary importance of the variational formulation is that it views entire paths at once. In particular, one can look for periodic paths by searching over spaces of periodic paths. It is not as natural to represent periodic solutions using the usual dynamical perspective, wherein one fixes initial conditions and then solves forward in time.

As the potential energy from gravitational attraction is nonpositive, the action of any path is nonnegative; in particular, it is bounded below. This motivates one to search for periodic orbits by finding minimizers of the action in some restricted space of paths. There are three possible obstacles to this course of action [17, pp. 74–75]:

- (a) *non-coercivity*, the problem that the infimum of the action amongst all periodic trajectories comes from having each body follow its own infinitely short, infinitely slow loop, and putting these loops infinitely far apart;



- (b) *triviality*, the possibility that an uninteresting (e.g., circular) orbit may minimize the action; and
- (c) *collisions*.

The problem of coercivity is usually addressed by imposing either topological constraints or symmetries. The topological constraints one finds in the literature consist of restricting to a particular homology class [41, 78, 94] or homotopy class [71, 70]. The other approach, symmetry constraints, is of more relevance to us. Certain symmetries are coercive, in the sense that if we restrict to orbits possessing such symmetries, then orbits of infinite size no longer minimize the action. The first example of such a symmetry [28] was the so-called “Italian symmetry” [19]: at every time, the configuration after advancing half of a period is the antipode of the current configuration. Imposing this symmetry forces coercivity, because it means that in order for bodies to become infinitely separated, they also have to travel infinitely far each period. More than the Italian symmetry, of specific interest to us is the choreographic constraint, the condition that moving forward in time by  $1/n$  of a period just effects a cyclic shift of the positions of the  $n$  bodies. This symmetry also forces coercivity. Thus, since we are only focused on choreographies, coercivity will not be a problem.

Imposing the choreographic constraint does not rule out trivial solutions, as the circular orbit is a choreography. However, triviality is typically not a major obstacle. For instance, imposing additional symmetries can rule out the circular orbit, as we shall see shortly with the proof of the figure-eight.

By far the trickiest of the three problems is that of collisions. By a result of Sundman, for times  $t$  near a binary collision at  $t_0$ , the distance  $r(t)$  between the colliding bodies satisfies  $r(t) \sim |t - t_0|^{2/3}$  [10, §2.2.3]. Thus the potential and kinetic energies, and so also the Lagrangian, scale as  $|t - t_0|^{-2/3}$ . This has an integrable singularity, so the action is *finite* through the collision. In particular, a minimizer of the action is not *a priori* collision-free. This is an unfortunate property of the Newtonian inverse potential  $1/r$ ; “strong force laws” such as potentials  $r^a$  with  $a \leq -2$  have the property that collisions have infinite action [78], and consequently one can prove theorems about the abundance of choreographies under such potentials [22, 71].

To prove existence of the figure-eight, Chenciner and Montgomery imposed a particular symmetry group (containing the choreographic symmetries) and then used a delicate calculation to rule out collisions [23]. They essentially showed that, while finite, the action through a collision would be larger than that achieved by a non-collision test path. This yields the following theorem (with presentation following Chenciner’s later article [17]).

Fix a period  $T > 0$ . Let

$$D_{12} = \langle r, s \mid r^6 = 1, s^2 = 1, rs = sr^{-1} \rangle$$

be the dihedral group of order 12. Let  $\beta$  be the group action of  $D_{12}$  on  $\mathbb{R}/T\mathbb{Z}$  defined by

$$\beta(r) \cdot t = t + \frac{T}{6} \quad \text{and} \quad \beta(s) \cdot t = -t,$$

and let  $\alpha$  be the group action of  $D_{12}$  on  $(\mathbb{R}^2)^3$  given by

$$\begin{aligned} \alpha(r) \cdot (x^{(1)}, x^{(2)}, x^{(3)}) &= (R(x^{(3)}), R(x^{(1)}), R(x^{(2)})) \\ \alpha(s) \cdot (x^{(1)}, x^{(2)}, x^{(3)}) &= (-x^{(1)}, -x^{(3)}, -x^{(2)}), \end{aligned}$$

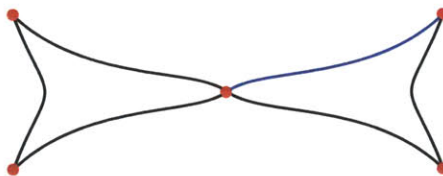


Figure 1.3: A loop with the sum-zero and  $D_{12}$ -equivariance properties imposed upon the figure-eight. The entire loop is determined by its values in the first  $(1/6)^{\text{th}}$  period (depicted in blue). The positions at each  $(1/6)^{\text{th}}$  period are marked in red.

where  $x^{(1)}, x^{(2)}, x^{(3)} \in \mathbb{R}^2$  and  $R: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  is the reflection map  $R(a, b) = (-a, b)$ .

**Theorem** ([23]). *There exists a periodic solution  $x^{(1)}(t), x^{(2)}(t), x^{(3)}(t): \mathbb{R}/T\mathbb{Z} \rightarrow \mathbb{R}^2$  of the planar, equal-mass 3-body problem satisfying the following properties:*

[sum-zero:] for all times  $t$ ,  $x^{(1)}(t) + x^{(2)}(t) + x^{(3)}(t) = 0$ , and

[ $D_{12}$ -equivariance:] for all  $g \in D_{12}$  and all times  $t$ ,

$$\alpha(g) \cdot (x^{(1)}(t), x^{(2)}(t), x^{(3)}(t)) = (x^{(1)}(\beta(g) \cdot t), x^{(2)}(\beta(g) \cdot t), x^{(3)}(\beta(g) \cdot t)).$$

*It is an “eight-shaped” choreography with zero angular momentum.*

The above discussion should lend credence to this theorem. We do not wish to repeat an actual proof, as the further details are fairly technical and tangential to our aims. Instead, we just take a moment to emphasize the philosophy of the proof: existence follows from taking a minimizer of the action subject to the chosen symmetry, with the one caveat that we must check if the minimizer is collision-free. The desired structure of the orbit is forced by judicious choice of symmetry group. In the particular case of the symmetries prescribed in the Chenciner-Montgomery theorem, one can check that an equivalent formulation is (i)  $x_2(t) = x_1(t + T/3)$  and  $x_3(t) = x_1(t + 2T/3)$  (i.e., the orbit is a choreography), (ii)  $x_1$  is an odd function, and (iii)  $x_1$  satisfies

$$x_1\left(t + \frac{T}{3}\right) = -x_1(t) - R\left(x_1\left(t + \frac{T}{6}\right)\right).$$

Consequently the entire orbit is determined by the function  $x_1$  on  $[0, T/6]$ , with the only additional constraint being that  $x_1(0) = 0$ . As demonstrated by example in Figure 1.3, this forces the orbit to have essentially the correct structure.

The variational proof technique has more applications, of course. It was applied to the  $n$ -body problem (although not to choreographies) as early as 1977, by William Gordon [41]. Another example (closer to our topic of study) is that shortly before the publication of the figure-eight orbit with Montgomery, Chenciner together with Andrea Venturelli demonstrated a proof of existence for a spatial periodic orbit with the Italian symmetry [24]. Known as the “hip-hop orbit,” it is a member of a family discovered previously by Davies et al. [29]. (The use of Italian symmetry is conducive to finding spatial orbits, because for  $n \geq 4$  bodies an action minimizer subject to said symmetry is necessarily nonplanar [19].)

Stronger machinery for proving existence using this variational method then developed quickly. Using a novel averaging argument due to Christian Marchal [65] to rule out isolated collisions, and then building on work of Montgomery, Terracini, and Venturelli, in 2002 Chenciner presented a proof of a general theorem showing that action-minimizing paths between two fixed endpoint configurations are collision-free [18]. As a consequence, one finds that periodic minimizers subject to certain symmetries, including the Italian symmetry, are collision-free.

Marchal’s technique applies to both planar and spatial orbits, but with different proofs. This is because his core averaging idea arose from the fact that the Newtonian potential is a harmonic function on  $\mathbb{R}^3$ . (Note that neither case implies the other, as a planar orbit is a spatial orbit but a minimizer in the space of planar orbits need not be a minimizer in the space of spatial orbits.)

In 2004, Davide Ferrario and Terracini generalized these ideas greatly by proving that *local* minimizers subject to a certain wide class of symmetry groups are collision-free [35, Theorem 10.10]. (As we shall see demonstrated in §1.5, the set of local minimizers is significantly richer than the set of global minimizers.) The “wide class” handled by Ferrario and Terracini’s result includes the choreographical symmetry group, the dihedral symmetry group of the figure-eight, and the symmetry group of the hip-hop orbit [35, Examples 1, 2, 6], so it generalizes all of the variational results mentioned heretofore. Their result also applies to many more symmetry groups, including a cyclic symmetry group yielding nontrivial spatial choreographies [35, Example 8].

Following this work, Vivina Barutello et al. analyzed all of the symmetry groups possible for the 3-body planar periodic problem, showing that local minima of the action with respect to each are collision-free [12]. Ferrario later proved the same for the 3-body spatial periodic problem [33].

At this point the collision-free proof technology for the variational method is well established, so that in most if not all desired cases, one knows that there exists a periodic orbit with specified symmetry. The approach continues to be applied to yield new existence proofs [80].

## 1.5 More choreographies

The variational proof technique discussed above is quite powerful, but it gives existence proofs for abstract orbits. In most cases (namely, every choreographic case known except the circle [13, Theorem 1]), we do not get a concrete description, or even approximation, of the global minimizer of the action. For example, Figure 1.2 actually plots a *local* minimum of the action subject to the choreographic symmetry. We do not know with certainty that it plots the Chenciner-Montgomery figure-eight, i.e., the global minimum of the action subject to  $D_{12}$ -symmetry. (We do suspect it does, because as Chenciner pointed out [17, p. 83], “... computations indicate that ‘the’ eight orbit is probably unique.” However, this remains unproven.)

As this discussion highlights, to supplement the proofs discussed above, the study of choreographies can be illuminated by explicit computational results searching for local minima. One takes a starting path and then numerically minimizes the action until it converges, using a finite-dimensional approximation to the function space to make the problem algorithmically tractable. Modulo numerical errors and the imprecision of the finite-dimensional approximation, the limit is a stationary point of the action and so is a choreography. This was the original perspective taken by Moore [71] when he numerically

discovered the figure-eight orbit.

*Remark.* Moore used discrete time samples to approximate the positions in his orbits and used finite differences to approximate velocities [72]. However, in all of the recent work and in our own work, orbits are instead approximated as trigonometric polynomials (i.e., one represents functions by truncated Fourier series). This will be made explicit in §2.4.

While the approach of numerically solving variational problems continued to be widely used outside of celestial mechanics (see, e.g., Gray et al.’s summary [42]), its application to choreographies lay dormant until Chenciner and Montgomery’s announcement of the figure-eight. Immediately thereafter, Simó began numerical computations in earnest [84]. In addition to finding the (or rather, as just discussed, “the”) figure-eight orbit numerically, he found many more choreographies as local minima of the action. (Indeed, since the action is a highly nonconvex function, one may expect it to have many critical points.) Simó presented 34 new  $n$ -body choreographies with  $n \geq 4$ . Position data for these and 12 more  $n$ -body choreographies are available at his website [81]. Using his data, we have plotted Simó’s orbits in Figures 3.5, 7.1, and 8.1–8.3. Notice that these orbits cannot be fully characterized by their symmetry groups, as (for instance) there are ten different 5-body choreographies in Figure 8.2 exhibiting  $\mathbb{Z}/2$  bilateral symmetry. Even more compellingly, in Figure 3.5 there are four choreographies exhibiting no symmetry in the trace of the orbit.

This work is numerical and non-rigorous in nature. However, in a series of three papers, Tomasz Kapela and coauthors developed computer-assisted proofs for a few of these numerically-found choreographies [58, 55, 56]. Figure 8.4 shows plots of the orbits handled by this work. We will explore their proof technique in more detail in §6.1.

In addition to Simó’s work, several others considered the problem of finding choreographies by numerically minimizing the action. Chenciner et al. (where this “et al.” includes Simó) presented several choreographies in the style of Simó’s orbits [22], some of which are and some of which are not contained in Simó’s previous tables [84]. (Specifically, there are five new choreographies shown in their paper [22, Fig. 3d,3e,4b,4c,4d].) Michael Nauenberg studied spatial variants of the figure-eight [73] and, with Moore, other periodic orbits with symmetries [72]. Another example is Robert Vanderbei, who considered periodic orbits as an application of his general-purpose nonconvex optimization software [93]. We suspect that there are many others who have experimented with action minimization as a means of finding choreographies.

The variants of the figure-eight mentioned in the last paragraph have been particularly well studied. They were discovered independently by Nauenberg (numerically) [73] and Marchal (with proof, using the variational method) [64]. They were later studied in more depth by Chenciner et al. [21] and by Nauenberg [74]. They belong to a family of *relative* choreographies. (Recall that a relative choreography is an orbit which is choreographic in a rotating frame) The angular velocity of the rotating frame varies continuously in the family. It connects the figure-eight orbit, which is a planar absolute choreography with zero angular momentum, to the circular orbit, which is a planar absolute choreography with nonzero angular momentum. The intermediary relative choreographies are spatial, and some (those for which the angular velocity of the rotating frame satisfies certain integrality conditions) are actually absolute choreographies, albeit with a longer period than their period in the rotating frame:

**Observation 1.2** ([85, p. 223]). *Suppose we have an  $n$ -body relative choreography which has period  $T$  and whose rotating frame (with respect to which it is a choreography) has period  $\tau$ .*

Suppose  $\tau/T$  is rational, say  $\tau/T = p/q$  in lowest terms. If  $(p, n) = 1$ , then the orbit is an absolute choreography with period  $pT$ .

A similar family was found by Arioli et al. [7]. Their work builds on the “mountain pass theorem;” recall that this is a theorem in analysis which (under suitable conditions) guarantees the existence of a saddle point between two distinct minima. Using a numerical procedure for locating such saddle points [14], they found a saddle point between the circular orbit and the doubled circular orbit. This approach was mentioned by Simó [84, p. 13], but Arioli et al.’s work appears to have been the first implementation thereof. They then took their mountain pass orbit and showed that it continues into a family of relative choreographies, and finally they gave a computer-assisted proof of the existence of this family. We will review their proof technique in §6.2 and remark upon their “mountain pass orbit” in §3.3.

The existence of one-parameter families of relative periodic orbits was first described by Michel Hénon in 1974 [47]. Using his ideas from that paper, Simó reported that Hénon found a continuous family of *planar* relative choreographies containing the figure-eight [85, p. 223]; the members of the family are called “satellites” of the figure-eight by Simó [85] and Chenciner et al. [22]. This family was also found by Nauenberg [73], and it is one of the three studied by Chenciner et al. [21].

The main difference amongst the three families just mentioned is which axis is chosen for the frame to rotate around.

In passing we observe that the existence of such families (in particular, the third family), together with Observation 1.2, immediately proves the following result. It is certainly well-known, although we have not seen it explicitly stated in the literature. For purposes of this statement, we view choreographies as “equivalent” if they are related by an overall scaling, spatial isometry, and/or affine time transformation.

**Corollary 1.3.** *There exist infinitely many inequivalent  $n$ -body (absolute) choreographies. In fact, there already exist infinitely many inequivalent 3-body planar choreographies.*

In addition to searching for local minima of the action subject to the choreographic symmetry, one can find minima of the action subject to a larger symmetry group. We have already seen the power of this idea in the context of Chenciner and Montgomery’s proof of existence of the figure-eight. Numerical results were pursued in some cases by, e.g., Moore and Nauenberg [72], but systematic exploration of such minima really began with the work of Barutello et al. [12] and Ferrario [33, 34] in the periodic 3-body problem. Recently James Montaldi and Katrina Steckles have given a comprehensive survey of the symmetry groups possible for planar  $n$ -body choreographies [69]. In addition to classifying the possible symmetry groups theoretically, they numerically found some choreographies with those symmetries; Montaldi’s website contains animations thereof [68]. Some of these are extremely beautiful and novel. We will revisit their classification in §3.5.

## 1.6 Non-variational approaches

In the previous subsection we discussed approaches to finding choreographies based on numerically minimizing the action. This has proven effective, but there are other ways to identify solutions. For instance, one can imagine searching the parameter space of initial conditions directly for choreographic solutions. The dimension of the space is not particularly large; for instance, in the planar 3-body problem (after removing symmetries) the phase

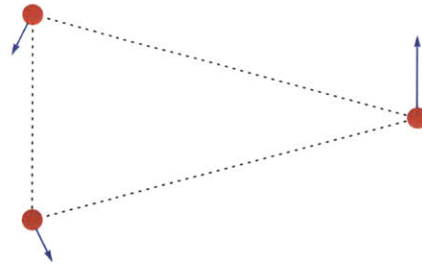


Figure 1.4: The initial conditions considered in Simó’s search for 3-body choreographies.

space is 6-dimensional. However, the time evolution map is ill-behaved, so that small changes in initial conditions lead to large (and perhaps very large) changes in the time-evolved conditions [83, p. 279].

Nonetheless, one might try to search this parameter space directly for choreographies. With suitably sophisticated differential equation solvers, Newton’s method operating on the phase space can be used to converge to a choreography given a reasonably good starting approximation [84, §7.2]. If one does not start with an approximation, though, then it is difficult to find a choreography in the full parameter space. Removing degrees of freedom from the problem can greatly improve the odds of success.

The circular orbit is an extreme example of this. It reduces the  $n$ -body problem to a two-parameter problem: the radius of the circle and the angular velocity. Finding solutions of the reduced problem is easy. More generally, given a central configuration, it is easy to (numerically) identify homographic solutions: one just needs to consider the scaling and rotation of the configuration over time, and conservation of energy and angular momentum impose stringent conditions on those functions.

*Aside.* Restricting the problem is a pervasive idea in physics. In the specific context of periodic orbits, E. Myles Standish, Jr. restricted to configurations starting from rest in order to identify what may be the first example of a nontrivial periodic solution of the comparable-mass 3-body problem [86]. (Previously there were known periodic solutions with collisions, but Standish’s solution is singularity-free.) There is also a large body of research concerning the *restricted three-body problem* in which one of the three bodies has negligible mass as compared with the others [10, §2.5].

In search of more 3-body choreographies, Simó considered configurations with the following restriction: the three bodies begin in an isosceles triangle and the orbit is invariant (up to relabeling of two of the bodies) under the simultaneous application of time reversal and reflection across the axis of symmetry of the triangle (see Figure 1.4) [85, §7]. If we (without loss of generality) fix the energy of the system, then this restriction leaves just three degrees of freedom.

Simó then systematically searched the reduced parameter space. For each initial condition he numerically integrated until either (1) the bodies returned to an isosceles configuration with the same sort of reflection symmetry as the starting configuration, but with a different body at the apex of the triangle; (2) a close approach occurred; or (3) a prechosen time threshold passed. In the event of (1), the solution was necessarily a relative choreography; if instead (2) or (3) occurred, then he dismissed the test case. Starting from each relative choreography, Simó used a “continuation process” to find a family of nearby relative choreographies with varying angular velocity for the rotating frame. He then selected from the resulting families

those relative choreographies in which the angular velocity satisfies the integrality condition so that the relative choreography is actually an absolute choreography (with longer period).

The results of these extensive computations were 345 choreographies, of which all but one (the figure-eight orbit) were new [85, §7]. Data files describing these choreographies are available at Simó's website [82].

Another study of periodic solutions of the 3-body problem, comparable in spirit but less extensive in scope, has recently received a lot of attention [90]. This work searched for periodic orbits starting from a certain two-dimensional parameter space.

## 2 Formal Problem Statement

### 2.1 Physics

Let  $d$  be the ambient dimension; in the present document we will consider  $d = 2$ .

The (nonrelativistic)  $n$ -body problem refers to a physical system in which there are  $n$  point masses interacting solely through Newtonian gravity. Recall that, in the theory of Newtonian gravity, a pair of point particles with masses  $m_1$  and  $m_2$  and separation  $r$  generate a force between them which is attractive and which has magnitude

$$F = \frac{Gm_1m_2}{r^2}.$$

Here  $G$  is the gravitational constant, a physical constant with the units of distance<sup>3</sup> · mass<sup>-1</sup> · time<sup>-2</sup>. Thus, if  $x^{(1)}(t), \dots, x^{(n)}(t): \mathbb{R} \rightarrow \mathbb{R}^d$  are the positions of the  $n$  bodies, then the dynamics of the  $n$ -body problem are determined by the equations

$$\ddot{x}^{(i)} = \sum_{j \neq i} \frac{Gm_j}{|x^{(j)} - x^{(i)}|^3} (x^{(j)} - x^{(i)}).$$

(In this document we follow the physicist's convention of using dot accents to denote time derivatives.)

An equivalent description of Newton's law of universal gravitation is that each pair interaction contributes  $-Gm_1m_2/r$  to the potential energy of the system. Thus the total kinetic and potential energies, respectively, are given by

$$T = \sum_{i=1}^n \frac{1}{2} m_i |\dot{x}^{(i)}|^2 \quad \text{and} \quad V = - \sum_{i < j} \frac{Gm_i m_j}{|x^{(j)} - x^{(i)}|}.$$

We will always consider *periodic* paths  $x^{(i)}(t)$ . Periodicity provides a clear choice for the limits of integration in the definition of the action: we simply integrate over a full period. Thus, letting  $T$  denote the period, the action of the periodic  $n$ -body system is

$$S = \int_0^T \left( \sum_{i=1}^n \frac{1}{2} m_i |\dot{x}^{(i)}|^2 + \sum_{i < j} \frac{Gm_i m_j}{|x^{(j)} - x^{(i)}|} \right) dt. \quad (2.1)$$

## 2.2 Conserved quantities

There are 10 classical conserved quantities of the spatial  $n$ -body problem. The simplest are the 6 coming from translation invariance, which may be viewed as the starting position and velocity of the center of mass:

$$\frac{1}{\sum_{i=1}^n m_i} \sum_{i=1}^n m_i x^{(i)}.$$

Rotational invariance gives conservation of angular momentum, a three-dimensional vector:

$$L = \sum_{i=1}^n m_i (x^{(i)} \times \dot{x}^{(i)}).$$

(Note that, in the planar case, the angular momentum is always perpendicular to the plane and so may be viewed as a scalar quantity. In general the angular momentum is properly defined in the exterior square  $\wedge^2(\mathbb{R}^d)$ .)

The final conserved quantity, the energy, arises from time invariance of the system:

$$H = T + V = \sum_{i=1}^n \frac{1}{2} m_i |\dot{x}^{(i)}|^2 - \sum_{i < j} \frac{G m_i m_j}{|x^{(j)} - x^{(i)}|}.$$

It is a nontrivial theorem of Bruns that there are no other first integrals [16, 54].

## 2.3 Normalization

Because the time and space scales are arbitrary, we will fix them in such a way as to simplify the problem. Consider replacing the position functions  $x^{(i)}(t)$  with  $\tilde{x}^{(i)}(t) = \alpha x^{(i)}(t/\beta)$  for some  $\alpha, \beta > 0$ ; in other words, consider rescaling time and space (by  $\beta$  and  $\alpha$ , respectively). The period of the new functions is  $\beta T$ , and the new action is

$$\beta \cdot \int_0^T \left( \sum_{i=1}^n \frac{\alpha^2}{\beta^2} \cdot \frac{1}{2} m_i |\dot{x}^{(i)}|^2 + \frac{1}{\alpha} \cdot \sum_{i < j} \frac{G m_i m_j}{|x^{(j)} - x^{(i)}|} \right) dt.$$

The physical meaning of the action is unchanged under multiplication by a constant. Thus there are essentially three degrees of freedom, *viz.*:  $\alpha$ ,  $\beta$ , and overall scale. We can use them to set, without loss of generality,  $T$ ,  $G$ , and the overall scaling of masses arbitrarily. This is easiest to see in two separate steps.

- (a) If we set  $\alpha = \gamma\mu^2$  and  $\beta = \gamma\mu^{3/2}$  for  $\gamma, \mu > 0$ , then the net effect of the transformation is to replace  $G$  with  $G/\gamma$  and replace the masses  $m_i$  with  $m_i/\mu$ .
- (b) If we then perform a transformation with  $\alpha = \beta^{2/3}$ , then the net effect is to multiply the period by  $\beta$  (and leave  $G$  and the masses unchanged).

Using this freedom, we will make the following choice throughout.

**Convention 2.1.** In the  $n$ -body problem, we set  $T = 1$ ,  $G = (2\pi)^4$ , and  $m_1 = (2\pi)^{-2}$ .

In particular, since all of the problems we will consider are equal-mass, all of the masses are set to  $1/(4\pi^2)$ .



We may also fix some of the conserved quantities. Firstly, we need to fix an inertial frame, i.e., the origin of our coordinate system. The following is, we claim, the only sensible choice.

**Convention 2.2.** Our  $n$ -body systems have their center of mass fixed at the origin. In particular, the center of mass is stationary, i.e., there is no net linear momentum.

Of course, asking for a periodic solution already forced the center of mass to be stationary.

There is remaining freedom in how to choose the orientation of the reference frame and the initial time. However, there is not a clear, canonical choice for these and so we shall not impose any condition.

*Aside.* The choice of normalization varies widely in the literature. As it happens, in his  $n$ -body choreographies discussed in §1.5, Simó used essentially the same normalization conventions as we have chosen (his choice was  $T = 2\pi$ ,  $m = 1$ ,  $G = 1$ , which differs from ours only superficially) [84]. However, in his search for 3-body choreographies discussed in §1.6, he used a significantly different normalization. In general the data one finds in the literature needs to be renormalized before it can be compared with ours. This problem of discrepancy in normalization conventions for numerical results was already bemoaned in 1974 [47, Appendix], when Hénon presented a set of conventions for the 3-body problem and “propose[d] that after discussion this set or a similar one be adopted as a standard” [47, p. 386]. We did not follow his conventions.

Finally, for later convenience, we shall agree to the following.

**Convention 2.3.** All energies are henceforth understood to be divided by  $2n$ .

For instance, the kinetic energy is  $\frac{1}{16n\pi^2} \sum_{i=1}^n |\dot{x}^{(i)}|^2$ . The reason for this rescaling will become clear in §4.2.

## 2.4 Fourier series representation

Because we only consider periodic orbits, it is natural to represent the coordinate functions as Fourier series. Recalling that we have normalized so that the period is  $T = 1$ , we write

$$x_c(t) := x_c^{(1)}(t) = \sum_{k \in \mathbb{Z}} A_k^{(c)} e^{-2\pi i k t},$$

where  $x_c^{(1)}$  is the  $c^{\text{th}}$  component ( $c = 1, \dots, d$ ) of the position function for the first body. Here  $A_k^{(c)} = \int_0^1 x_c(t) e^{2\pi i k t} dt$  is the  $k^{\text{th}}$  *Fourier coefficient* of  $x_c(t)$ .

*Aside.* We find the application of Fourier series to celestial mechanics quite pleasing, as the original development of the fast Fourier transform was by Gauss to aid his calculation of ephemerides [46].

Note that for choreographies, wherein all  $n$  bodies follow the same curve, we do not need to write down series for the positions of the other bodies. Moreover, because we fixed the center of mass at the origin, under the choreographic constraint it follows that the mean of  $x(t)$  must be zero. That is,  $A_0^{(c)} = 0$  for all  $c$ .

*Remark.* The condition  $A_0^{(c)} = 0$  states that, on average, the center of mass is at the origin. This requires only the part of the choreographic constraint that the bodies follow the same curve, not that they do so with equal time spacing. But the center of mass is

always at the origin, and (under our definition) the bodies are equally spaced in time. Thus  $\sum_{i=1}^n x_c^{(i)}(t) = n \sum_{k:n|k} A_k^{(c)} e^{-2\pi i k t}$  is identically zero, so  $A_k^{(c)} = 0$  for all  $k$  such that  $n \mid k$ . We do not exploit this anywhere, but it can provide a useful spot-check of numerical results.

Now, because the functions  $x_c(t)$  are real, the Fourier coefficients  $A_k^{(c)}$  satisfy  $A_{-k}^{(c)} = \overline{A_k^{(c)}}$ . Thus, writing  $A_k^{(c)} = a_k^{(c)} + i b_k^{(c)}$  for  $k \geq 1$ , we have

$$x_c(t) = 2 \sum_{k=1}^{\infty} \left( a_k^{(c)} \cos 2\pi k t + b_k^{(c)} \sin 2\pi k t \right).$$

Given coordinate functions  $x(t) = (x_1(t), \dots, x_d(t))$  with their corresponding Fourier coefficients, we define the  $\ell^1$  norm

$$\|x\| = \sum_{k=1}^{\infty} \sum_{c=1}^d k (|a_k^{(c)}| + |b_k^{(c)}|).$$

This choice of norm is really at the heart of our proof technique; it will be used crucially in §4.

*Aside.* This norm is an  $\ell^1$  norm in terms of varying Fourier modes and also an  $\ell^1$  norm in terms of the real coefficients for each mode. We could instead consider the mixed  $\ell^1, \ell^2$  norm

$$\|x\|_{1,2} := \sum_{k=1}^{\infty} k \left( \sum_{c=1}^d |A_k^{(c)}|^2 \right)^{1/2}.$$

This norm is actually mathematically better suited to our goals (as we can see already in Lemma 2.5). However, we found the computer implementation to be simpler using the purely  $\ell^1$  norm, and in any event the norms are equivalent:  $\|x\|_{1,2} \leq \|x\| \leq \sqrt{2d} \|x\|_{1,2}$ . Therefore we shall content ourselves with the analysis as-is.

**Definition 2.4.** By  $\mathcal{X}$  we understand the space of mean-zero  $\mathbb{R}^d$ -valued periodic functions  $x(t)$  such that  $\|x\| < \infty$ .

Note that  $\mathcal{X}$  is a Banach space isomorphic to  $\ell^1$ . When convenient we identify it with  $\ell^1$  by the mapping  $x \mapsto (ka_k^{(c)}, kb_k^{(c)} : c = 1, \dots, d, k \geq 1)$ . It contains all sufficiently differentiable (for instance,  $C^3$  suffices) functions. In particular, because a collision-free solution of the  $n$ -body problem is necessarily smooth, it contains all of the orbits we wish to study.

The following bounds are immediate from the definition.

**Lemma 2.5.** *Given any  $x = (x_1, \dots, x_d) \in \mathcal{X}$ , the functions  $x_c$  are  $C^1$ . Moreover, for any  $t \in \mathbb{R}$ ,*

$$\sum_{c=1}^d |x_c(t)| \leq 2\|x\| \quad \text{and} \quad \sum_{c=1}^d |\dot{x}_c(t)| \leq 4\pi\|x\|.$$

In other words, in defiance of Heisenberg, the norm on  $\mathcal{X}$  simultaneously controls position and velocity.

## 2.5 The action and its gradient

Applying the conventions from §2.3 to (2.1) and distributing the integral, we have the following expression for the action:

$$\frac{1}{16n\pi^2} \sum_{i=1}^n \int_0^1 |\dot{x}^{(i)}|^2 dt + \frac{1}{4n} \sum_{i \neq j} \int_0^1 \frac{1}{|x^{(j)} - x^{(i)}|} dt.$$

Consider first the kinetic terms  $\int_0^1 |\dot{x}^{(i)}|^2 dt$ . By the choreographic condition, these are all equal to  $\int_0^1 |\dot{x}|^2 dt = \sum_{c=1}^d \int_0^1 |\dot{x}_c|^2 dt$ . By Parseval's identity, the  $L^2$  norms can be computed in terms of the Fourier coefficients of  $\dot{x}$ :

$$\int_0^1 |\dot{x}_c|^2 dt = \sum_{k \in \mathbb{Z}} \left| (-2\pi i k) A_k^{(c)} \right|^2 = 8\pi^2 \sum_{k=1}^{\infty} k^2 |A_k^{(c)}|^2.$$

The potential terms are not as simple. We can make one simplification, though: using the choreographic condition, we can replace the double sum over  $i, j$  with  $n$  copies of a single sum over  $j$ .

Putting all of this together, our final expression for the action of a choreographic orbit is

$$S(x) = \frac{1}{2} \sum_{k=1}^{\infty} k^2 \sum_{c=1}^d |A_k^{(c)}|^2 + \frac{1}{4} \sum_{j=1}^{n-1} \int_0^1 \frac{1}{|x(t+j/n) - x(t)|} dt. \quad (2.2)$$

Let  $U_{\text{cf}} \subset \mathcal{X}$  be the “collision-free” subset of  $\mathcal{X}$ , i.e., the set of  $x(t)$  such that  $x(t+j/n) \neq x(t)$  for all  $t \in \mathbb{R}$  and  $j = 1, \dots, n-1$ . Because the values of  $x$  vary continuously with respect to the norm on  $\mathcal{X}$  (by Lemma 2.5),  $U_{\text{cf}}$  is an open set. The action defines a functional  $S: U_{\text{cf}} \rightarrow \mathbb{R}$ .

This function is twice-differentiable, but some care needs to be taken in how we define the derivatives. We will examine this in detail in §4.1 and especially §4.2, and content ourselves here with simply defining the norm on the first derivative. By abuse of notion we often refer to the first derivative as the “gradient,” denote it by  $\nabla S$  (see Definition 4.7), and think of it as a vector of partial derivatives with respect to the coordinates  $ka_k^{(c)}, kb_k^{(c)}$ . Using this viewpoint, we choose the  $\ell^1$  norm:

$$\|\nabla S(x)\| = \sum_{k=1}^{\infty} \sum_{c=1}^d \left( \left| \frac{\partial S}{\partial (ka_k^{(c)})}(x) \right| + \left| \frac{\partial S}{\partial (kb_k^{(c)})}(x) \right| \right). \quad (2.3)$$

Note that this is *not* the dual norm that one typically uses with the Fréchet derivative; that would be an  $\ell^\infty$  norm. This norm is much stronger.

## 3 Our Results

Our primary mathematical result is the proof technique developed in §4. As a demonstration of its power, we have applied it to prove the existence of a slew of choreographies — manifold, although not yet myriad. To date, we have proven the existence of 192 choreographies. This number can be compared with the 16 choreographies which had previously had their existence certified by Kapela et al. (see Figure 8.4) — all of which we have re-proven with our

software — and the one family treated by Arioli et al. [7]. In addition to being substantially larger, our storehouse of choreographies includes behavior far more complicated than had been handled before.

To feed the prover, we also developed techniques for finding choreographies. While these techniques are not novel and not especially sophisticated, they did prove to be powerful tools; we describe our methods in §3.2 and explore their main limitation in §3.4. Finally, we use the remainder of this section to discuss some observations and speculations regarding the phenomena of symmetry and stability.

We display a selection of 64 proven orbits in this document, in the present section and then continuing in §8, *viz.*: Figures 3.1, 3.4, 3.6, 3.7, 3.8, 7.2, 8.5, 8.6, 8.7, 8.8, 8.9, 8.10, 8.11, and 8.12. For each of these we show a plot of the trajectory, record the number of bodies, name the original discoverer (to the best of our knowledge) of the choreography, and give the name of the choreography (if one was given). An updated, complete laundry list of proven orbits, with notes and animations, can be found at the author’s website

<http://gminton.org/choreo.html>

(which also contains the CHOREO.JS program discussed below) [67].

### 3.1 Proving existence of choreographies

The mathematical presentation of our proof technique is delayed until §4, but to understand the statements of our existence results we need to preview a few aspects of it. There are several parameters to be chosen per proof. First, recall that, as input to Theorem I.2.1, we have to choose  $\varepsilon$ , the radius of the ball in which we are going to (attempt to) prove existence. The constraining factors are that a larger  $\varepsilon$  means that we do not need to have as accurate of a starting point, but it also means that we have to control the behavior of the problem on a larger ball. (In particular, the bounds we give in §4.7 depend critically on  $\varepsilon$  being small.) Our choice of  $\varepsilon$  varied somewhat in these proofs. In most of our proofs we used double-precision approximate calculations to find a solutions whose gradient had norm less than  $10^{-13}$ , and then we proved existence inside a ball of radius  $\varepsilon = 10^{-10}$ . There were a few exceptions, though, where we had to use a smaller  $\varepsilon$  and consequently a better starting approximation (see Figure 3.1). In those cases we improved our approximation by running our solver with double-double-precision floating-point arithmetic (as provided by the QD library [48]). In the most extreme case, we found a solution whose gradient had norm less than  $10^{-18}$  and then proved existence within a ball of radius  $\varepsilon = 10^{-13}$ .

In all cases  $\varepsilon$  was orders of magnitude less than the diameter of the choreography, and so a successful computer-assisted proof tells us that there is a choreography whose plot is essentially indistinguishable from that given. In particular, the largest  $\varepsilon$  we ever used was  $10^{-8}$  and the diameter of every orbit was at least 0.5. Recalling Lemma 2.5, that lets us state the following master theorem.

**Theorem 3.1.** *In each case of proven existence, there is a true choreography whose graph would be indistinguishable from that of our approximate solution even if the resolution of the plot was one part in a million.*

To understate, the plots given here may be thought of as reasonably representative.

Besides  $\varepsilon$ , another parameter is the number of Fourier modes that we need to control, both in our approximate solution (see §3.2) and in bounding the infinite matrix in our proof (see “ $\kappa$ ” in §4.6). This number, times  $2d$ , is the dimensionality of the space we use to

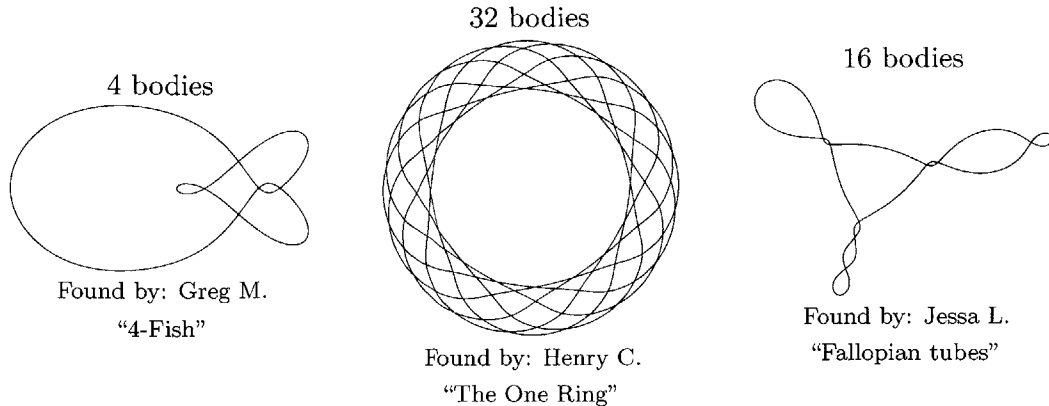


Figure 3.1: Three choreographies for which we used QD to find a higher-precision approximation and then proved existence using smaller  $\varepsilon$ ; the first and second used  $\varepsilon = 10^{-11}$ , and the third used  $\varepsilon = 10^{-13}$ .

approximate  $\mathcal{X}$ . Simó stated that the dimension he used was typically in the range  $[10^3, 10^4]$ . This generally lines up with our results. We found that 800 modes was often sufficient to give a satisfactory approximation for a moderately complicated planar choreography. That corresponds to a 3200-dimensional approximation space.

In addition to choreographies found using our software, which is described below in §3.2, we proved the existence of some choreographies found by others. In particular, we used data from Simó’s  $n$ -body and 3-body data sets, mentioned in §1.5 and §1.6, respectively. Simó made his  $n$ -body data set available in the form of time series of positions, using essentially the same normalization as we chose (see §2.3); it was easy to import these. We proved existence of 33 of the 47 choreographies in this data set. We anticipate no major obstacle to proving existence of the others, but the default parameters were not sufficient and we have not yet experimented with other settings.

By contrast, Simó’s 3-body data set is available in the form of positions which are *not* uniformly sampled in time. We dealt with this by taking an approximate time rescaling based on making the angular momentum constant, and then using our software to converge to a solution. In some cases this was successful and in some it was not; in some of the successful cases we have proven existence, and in some we have not. For the converged-but-not-yet-proven cases, like the  $n$ -body choreographies discussed in the last paragraph, we anticipate no fundamental problem in proving them. We just have not yet done so. Of the 345 orbits in the data set, we have proven existence of 19 and have found approximate solutions for another 50.

Figures 8.5 and 8.6 show a handful of Simó’s choreographies for which we have proven existence.

### 3.2 Finding choreographies numerically

While our primary results concern computer-assisted proof, we also developed software for numerically finding choreographies. Our software differs from previous work in two ways: firstly, we eschew action minimization and instead use second-order methods to look for critical points, and secondly, we start our search from a user specified starting loop.

Concerning the first difference, recall that, as we discussed in §1.5, the primary method discussed in the literature for finding choreographies is action minimization. Indeed, the principle of stationary action is often sloppily called the principle of least action, which strongly suggests that one look for minima. More specifically, the common approach is to minimize the action by using gradient descent in the space of trigonometric polynomials of a given trigonometric degree (i.e., Fourier series truncated at a certain mode). Gradient descent has definite advantages, of which the foremost is that (having arranged for coercivity by imposing the choreographic constraint) it will converge to some local minimum. That minimum may be uninteresting (e.g., the circular orbit) or it may have collisions, but at least we know that the algorithm will converge to something.

However, gradient descent also has some disadvantages. Barring extreme numerical coincidences, it can only find local minima, so we lose the ability to identify saddle points. It is also slower to converge than higher-order methods, which is relevant because we require high-precision inputs for the prover. For these reasons, instead of using gradient descent, in our software we use Newton’s method to search for a zero of the gradient. This has the advantages of being able to find saddle points (which we will return to in §3.3) and of converging rapidly, to wit, converging quadratically in a neighborhood of a solution. Simó wrote that the number of iterations needed to “achieve a good approximation” is quite large, on the order of  $10^3$  to  $10^4$  [84, p. 12]. Orders of magnitude fewer steps suffice with our Newton techniques.

The most notable downside of using Newton’s method is that it is ill-behaved when the starting point is far from a solution. To combat this, we actually use a *damped* Newton’s method algorithm. More specifically, we use the same damping algorithm as in §II.6.2, i.e., we just compute the norm of each step and scale it back if it exceeds some threshold. We have found this to be satisfactory in practice, although there is certainly room for improvement. A more principled damping method like the Levenberg-Marquardt algorithm [63, 66] could be useful.

Even after instituting damping, we still lose the guarantee of convergence that gradient descent has. This problem is mitigated in practice by applying a random perturbation when the algorithm is unable to make progress.

We suspect that this approach is not used in the literature due to a combination of the lack of convergence guarantees and the increased programming complexity. The code needed to compute the Hessian of the action is significantly more complicated than that needed to compute the gradient. However, we had to develop code for Hessian computations anyway for the proof procedure, so reusing that work for Newton’s method was natural. The computer-assisted proof technique used by Arioli et al. also required Hessian computations, and they do some calculations with Newton’s method, but only to improve the precision of an already-close approximation [7].

At the start of this subsection we mentioned another difference, that of the choice of starting curve. In 2000, Simó observed that the starting point for action minimization could be a very rough approximation, even a “hand drawn curve” [84, p. 12]. We do not know if he ever tried this, but we have implemented exactly such an input method. Our software accepts a user-drawn curve — sketched with a computer mouse or with a finger on touchscreen devices. This curve can be quite crude; for instance, we have found that haphazard scribbles in the vague shape of a figure-eight converge rapidly in the  $n = 3$  case to “the” figure-eight.

We developed three programs for finding choreographies: COMPUTE, CHOREOGRAPHER, and CHOREO.JS. The first, COMPUTE, just takes a set of Fourier coefficients and performs

some number of steps of Newton’s method. It does not accept interactive input, but accepts various options like a variable number of Fourier modes. We use this to find higher-precision approximations to hand to the prover. The other two, CHOREOGRAPHER and CHOREO.JS, are the interactive programs that start with user sketches. CHOREOGRAPHER is a C++ program which actually performs the damped Newton’s method algorithm just described. At the time of writing, CHOREOGRAPHER has not been widely released and so it is responsible for finding relatively few choreographies, and almost all of them were found by the author.

By contrast, the last program, CHOREO.JS, is a JavaScript program which can be accessed from the author’s website [67]. Users can access it, run the program inside their web browser, and then upload any choreographies they may find. Uploaded choreographies are stored for later processing. In this way we have crowdsourced the search for interesting choreographies.

Unfortunately, inverting large matrices is impractical inside of a web browser. Therefore we needed a different approach in order to make CHOREO.JS practical. So, instead of using Newton’s method, we used a *quasi-Newton method* [79, §10.9]. Such methods are analogs of the one-dimensional technique of approximating Newton’s method by using secant lines to approximate tangent lines. In the multidimensional setting there is not a unique “secant” approximation, so there are multiple quasi-Newton methods. We used the symmetric rank-one (SR1) method. This has the property that, unlike other popular quasi-Newton methods (e.g., BFGS), SR1 allows for indefinite matrices. This is sometimes considered a disadvantage of the method, but it is an advantage for us because we are explicitly interested in saddle points.

Using CHOREOGRAPHER, we found 8 new choreographies for which we then proved existence. Two particularly interesting members of this set are shown in Figure 3.6, and the remaining 6 can be found in Figure 8.7.

At the time of writing, 658 trajectories have been uploaded using CHOREO.JS. Many of these are false solutions; they are not collision-free critical points of the action, but rather they are artifacts of finite-dimensional approximation (see §3.4). There are also many duplicates. Even so, we have proven the existence of 128 new choreographies from these submissions. The 44 proven choreographies depicted in this document and not in Figures 3.6, 8.5, 8.6, or 8.7 were all found using CHOREO.JS.

### 3.3 Saddle points of the action

Recall that the “mountain pass choreography” of Arioli et al. [7] is a spatial choreography which is a saddle point of the action and lies between the local minima of the circular orbit and the doubled circular orbit. A significant amount of work was involved in constructing it, as it is nontrivial to implement the mountain pass theorem in a numerical algorithm [14]. Following such a procedure is *guaranteed* to construct a saddle point, but one might wonder if saddle points are actually that hard to find.

*Aside.* Our early efforts to reconcile the results of Arioli et al. with our own work suggested that the mountain pass choreography they found might not be new, but rather a familiar choreography viewed in a rotating frame. We have not yet investigated this fully, though.

One of our stated reasons for adopting Newton’s method is that it could, at least in principle, find saddle points of the action. But more than finding saddle points, we need to be able to recognize one when we see it. To check if a given choreography is a saddle point, we just need to check if the Hessian of the action at that choreography is indefinite. Since we have to compute the Hessian anyway to perform steps of Newton’s method, it is

easy enough to check its signature. COMPUTE has this capability; more specifically, it can compute the eigenvalues of the Hessian.

We should emphasize that these computations are *not rigorous*, both because they take place in floating-point arithmetic and because they work with a finite-dimensional approximation of the “true” Hessian. It is possible to make them rigorous, for instance by using the analysis given in §4.6 to determine conditions under which the finite-dimensional approximation of the Hessian would necessarily have the same number of negative eigenvalues as the true Hessian. We content ourselves, though, with just discussing saddle points and minima on the basis of nonrigorous, but convincing, numerical evidence.

It turns out to indeed be the case that Newton’s method with no additional modification can find saddle points; several of the choreographies we have found are saddle points. In particular, of the 8 choreographies found using CHOREOGRAPHER and depicted in Figures 3.6 and Figure 8.7, only one is a local minimum of the action (namely, the 12-point hexagon). For the other 7 choreographies, the Hessian has at least one negative eigenvalue (and, e.g., “Finger painting” in Figure 3.6 has 3 negative eigenvalues).

To date the basic version of CHOREO.JS has not found any new choreographies which are saddle points. This is somewhat disappointing, but not extremely surprising, because quasi-Newton methods (specifically, SR1 as applied to our problem) start(s) by performing gradient descent. Thus the algorithm can exit the vicinity of a saddle point before it detects the indefiniteness of the Hessian.

We said the “basic” version of CHOREO.JS because, as we will discuss in §3.5, CHOREO.JS has the ability of search for choreographies with certain symmetries imposed. Imposing symmetries amounts to looking for a stationary point of the action on a subspace; if we find a stationary point — even a minimum — on the subspace, there is no reason to expect that is a minimum on the full space. And indeed we do find that choreographies found with imposed symmetries are often saddle points of the action; for instance, “Whirlybird” in Figure 8.9 has 3 negative eigenvalues.

This observation is certainly not original to us. In fact, Simó’s  $n$ -body data set contains a saddle point, namely, the choreography “11 bodies on a circle with a flower” shown in Figure 8.5. This is the only saddle point in the data set. Simó does not comment on this, but we suspect he found this orbit by minimizing the action subject to a bilateral reflection symmetry.

Finally, we note that non-variational approaches for finding choreographies are very successful at finding saddle points. This is to be expected, because there is not reason *a priori* why a non-variational method would prefer minima. Recall that Simó’s 3-body data set came from a non-variational method (see §1.6). Of all the choreographies in that set for which we have good enough numerical approximations to believe the Hessian calculations (69, at current count), only the figure-eight is a minimum of action.

### 3.4 Identifying real orbits

In all of our computations, we truncate the Fourier series at a finite number of modes. This affects the quality of the solution, and it also masks whether or not a given solution is real. We have encountered many situations in which the algorithm converges to a critical point of the action given some mode cutoff, but as the mode cutoff increases it does not converge to an actual critical point.

A representative example of how this manifests is provided by an orbit submitted using CHOREO.JS by Yichen H., which he entitled “Tokyo love story.” Using COMPUTE, we took a



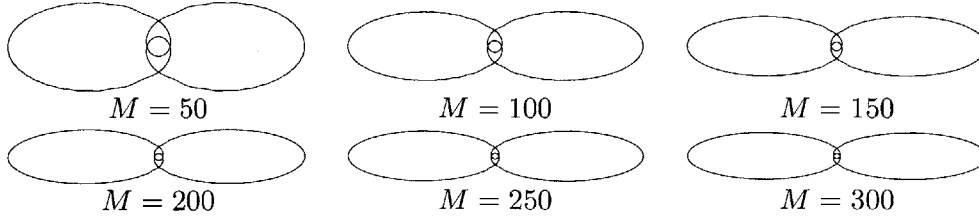


Figure 3.2: A family of numerical trajectories for the 2-body problem. These are critical points of the action as a function of the first  $M$  Fourier modes only.

starting approximation to Tokyo love story using Fourier coefficients up to mode 50 and then took enough steps of Newton’s method to converge to a critical point of the action (as a function of those variables only). By computing some additional derivatives of the action, COMPUTE can estimate the norm of the “full” gradient, i.e., the derivative in terms of all of the Fourier coefficients. We noted this norm, then iteratively increased the mode cutoff and repeated the process. If we were converging to a true choreography, then the norm of the full gradient would converge to 0. Instead, the following table shows the behavior we observed for Tokyo love story.

Mode cutoff	Norm of the gradient	Mode cutoff	Norm of the gradient
50	0.773	200	1.40
100	1.06	250	1.52
150	1.24	300	1.63

The qualitative behavior here is clear: the norm is increasing instead of decreasing to zero. We have plotted the corresponding orbits at each stage in Figure 3.2. These plots reveal the underlying problem: as the mode cutoff increases, the corresponding critical point of the action approaches a collision.

As the caption in Figure 3.2 says, Tokyo love story was supposed to be a 2-body choreography. But the 2-body problem is solved, and in particular the only choreographies are circular orbits. Thus the Tokyo love story was, in fact, never meant to be.

The driving-to-collision behavior is most common failure mode that we have encountered. Another example is demonstrated in Figure 3.3, this time for a 3-body (purported) choreography. In this case we know of no *a priori* reason why there could not have been a choreography of this type.

We dealt with this problem in a fairly unintelligent manner, by simply throwing out hypothetical choreographies such that the norm of the full gradient increased when we increased the mode cutoff. Another idea would be to reject any orbit in which two bodies come closer than a fixed distance cutoff; this is what Simó did in his 3-body searches [85, §7].

### 3.5 Symmetry

Examining, e.g., Figures 8.1–8.3 leads one to the conclusion that choreographies tend to possess nontrivial symmetry. To the best of our knowledge, in his  $n$ -body data set Simó did not explicitly search for choreographies with symmetries (except possibly for the 11-body choreography mentioned in §3.3), but nonetheless those are the solutions he found. This observation has been made repeatedly; Ferrario calls it the “recurring phenomenon of ‘more symmetries than expected’ in  $n$ -body problems” [33, p. 391].

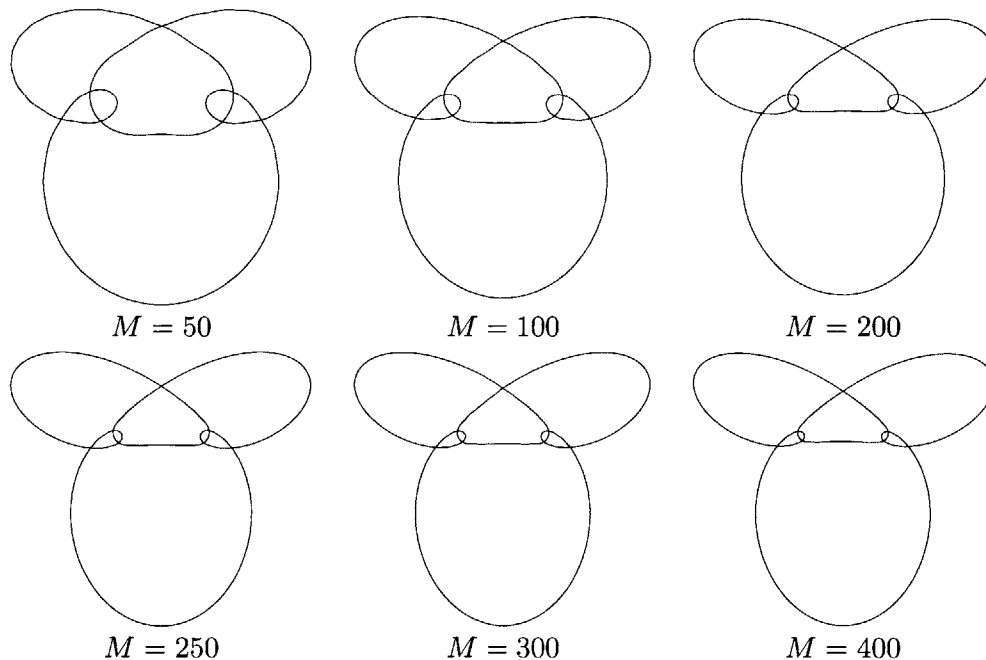


Figure 3.3: A family of numerical trajectories for the 3-body problem. These are critical points of the action as a function of the first  $M$  Fourier modes only.

This phenomenon recurred in the choreographies found by CHOREO.JS as well. Figure 3.4 shows three choreographies which were found without any symmetry being forced upon the solution.

In his 45-choreography corpus, Simó did find four asymmetric choreographies. They are plotted in Figure 3.5. The smallest has 6 bodies; to the best of our knowledge, before our work there were no asymmetric choreographies known with fewer bodies.

We have had more success finding asymmetric orbits. In particular, Figure 3.6 shows two 4-body asymmetric choreographies which we found using CHOREOGRAPHER. We already remarked in §3.3 that these choreographies are saddle points; it seems possible that there is a connection between saddle points and asymmetric choreographies, and such a connection may explain why we have had less difficulty finding asymmetric choreographies. We have still not found any asymmetric 3-body choreographies, though. To our knowledge, the existence of such is an open question.

The above discussion focuses on searching for asymmetric choreographies, but intentionally searching for symmetric choreographies can also be a fruitful endeavor. Recall from above that Montaldi and Steckles have classified the possible symmetry groups for planar choreographies. We refer the reader to their paper for the details of this classification [69], and here just give a quick summary. If  $n$  is even then the set of possible symmetry groups of an  $n$ -body choreography consists of two infinite families,  $\{C(n, k/\ell)\}$  and  $\{D(n, k/\ell)\}$ , where  $k \geq 1$  and  $\ell$  is prime to  $k$ . Possessing  $C(n, k/\ell)$  as a symmetry group means that advancing time by  $1/k$  results in a spatial rotation of  $2\pi\ell/k$ . The group  $D(n, k/\ell)$  is the group generated by  $C(n, k/\ell)$  and a time-reversal bilateral reflection symmetry. If  $n$  is odd then, in addition to these two infinite families, there are three additional exceptional symmetry groups.

The most important aspect of this classification for our present purposes is that the

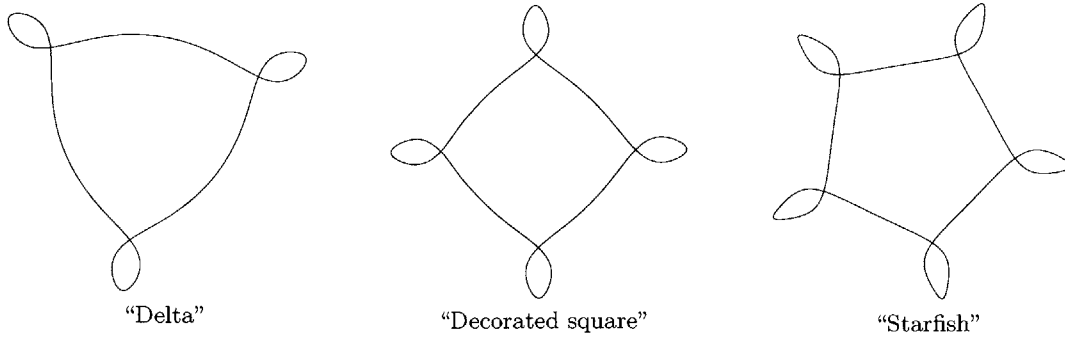


Figure 3.4: Three 8-body choreographies which were found by Abhinav K. using CHOREO.JS and which we proved to exist. These choreographies exhibit symmetry despite not being constrained to do so.

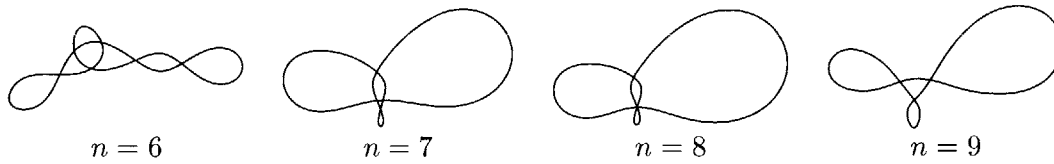


Figure 3.5: Asymmetric  $n$ -body choreographies found by Simó [81].

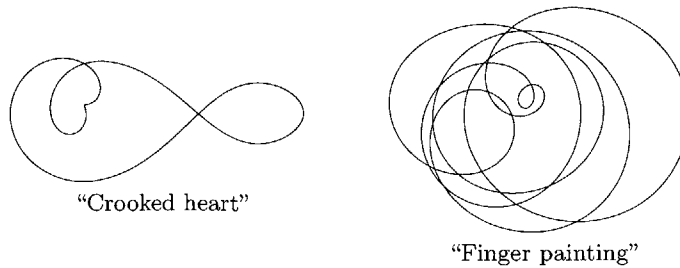
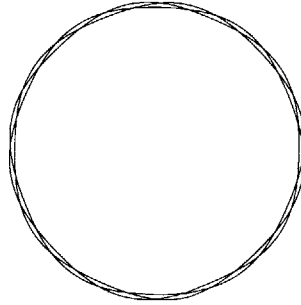


Figure 3.6: Two 4-body choreographies which we found using CHOREOGRAPHER and proved to exist. They exhibit no symmetries and they are both saddle points of the action.



“Drunken sailor”

Figure 3.7: A 32-body choreography which was found by Henry C. using CHOREO.JS, with an imposed symmetry group, and which we proved to exist.

property of a choreography possessing a given symmetry group is naturally expressed in terms of the Fourier coefficients. This made it easy to adapt CHOREO.JS to search for choreographies subject to possessing a given symmetry group; we just project the Fourier coefficients to the subspace obeying the symmetry (at every iteration of the quasi-Newton’s method procedure). This feature enabled the discovery of many more choreographies. One example of such a choreography is shown in Figure 3.7.

This choreography is surprisingly close to, but definitely not the same as, the circular orbit. It is a saddle point, and it would have been very difficult to find without imposing a nontrivial symmetry group.

Throughout this subsection we have referred to a choreography possessing or not possessing symmetry without rigorously justifying those assertions. The lack of a certain symmetry is an open condition, so it would be routine to verify that, e.g., the choreographies in Figure 3.6 actually are asymmetric. The opposite direction, showing that a choreography does possess symmetry, is more delicate. One could work in the subspace of loops possessing the symmetry in question and apply the action principle in that subspace (see the comment after Corollary 4.5). Another approach, which we find more appealing, is to argue that if a given choreography nearly possesses a given symmetry group, then it exactly possesses a conjugate of that symmetry group. Statements of that form can be derived from the uniqueness part of our existence theorem (Proposition I.2.4). Although we have not developed this theory yet, we are interested in doing so in the future; see Problem 7.5.

It is also worth noting that Montaldi and Steckles did not prove the existence of choreographies with (exactly) each specified symmetry group; they explicitly state that existence proofs were not part of their paper [69]. Thus if we did make rigorous the notion of possessing a given symmetry group, our work could provide some concrete existence theorems supporting their work.

### 3.6 Stability

The stability of the figure-eight is remarkable, but the idea seems fantastic that it is the *only* stable choreography. Recent work has proved the stability of some spatial choreographies in a family containing the figure-eight [57]; however, in this document we have limited our attention to planar choreographies, and in any event it is still of interest to find stable choreographies which are not just perturbations of the figure-eight.

Kapela and Simó remark that there is one other planar choreography which is a known

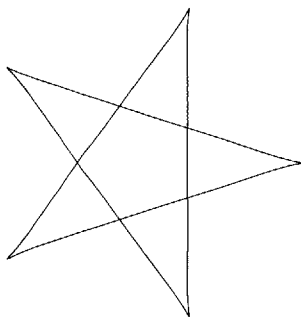


Figure 3.8: A 4-body symmetric choreography which was found by Kyle M. using CHOREO.JS and which we proved to exist.

candidate for stability: it is a 4-body choreography “looking like a pentagonal star” [56, p. 1252]. They attribute this observation to a preprint by a J. McLaughlin, but as of writing we have not obtained this preprint.

Independently of this remark, a user of CHOREO.JS found the choreography in Figure 3.8.

It seems likely that this choreography is the “pentagonal star” alluded to by Kapela and Simó. In addition to it having the right shape, it is plausibly stable: we simulated the choreography through two periods using a simple leap-frog integrator and, despite the numerical inaccuracy of this approach, the trajectory was very nearly periodic. We applied the same heuristic to the other choreographies in our repository, and (other than the figure-eight) no other choreography passed. This is not particularly convincing evidence because of the lack of sophistication in our choice of differential equation solver (but at least our choice is better than manual simulation using light bulbs [49]). We propose as further work to study stability in more detail (see §7).

## 4 Our Proof Technique

In this section we present and analyze our technique for computer-assisted, rigorous proof of existence for  $n$ -body choreographies. Our mindset here is theoretical; we aim to firmly establish the mathematical underpinnings of the approach and present the algorithm in enough detail to specify a reference implementation. Programming-specific issues and optimizations, which are necessary to make the approach work in practice but are tangential to the theory, are deferred to §5.

The basic plan is as follows. First we make the action principle explicit, so that our remaining task is only to find a stationary point of the action. We then give explicit formulae for the “gradient” and “Hessian” (first and second derivatives, respectively) of the action in terms of the Fourier coefficients of certain auxiliary functions, and discuss how to bound those coefficients. We find that the Hessian is dominated by the kinetic contribution, which is very simple. Treating the potential contribution as a correction to the dominant kinetic term, we show how to bound the inverse of the Hessian. We conclude by combining the bounds to get a condition on when our effective existence theorem, Theorem I.2.1, applies to establish the existence of a nearby zero of the gradient.

Before beginning, let us note just once that all numerical computations in our proof software (with just two exceptions, which we spell out explicitly in §5.1 and §5.4) are done with rigorous interval arithmetic (see §I.3). We talk about a computation being “exact” when

it would be exact given infinite-precision arithmetic, and so the final interval is guaranteed to include the exact answer.

## 4.1 Action principle

The very beginning of our approach is the action principle, the law that stationary points of the action are physical trajectories. (Here and in the sequel, when we say a trajectory is *physical* we just mean that it obeys the equations of motion.) In the case of periodic functions (which is the only case we need), we have thus far defined the action (2.1) but have not formally stated the principle. We fill in that lacuna now.

The principle is very well-known, but in most sources it is either not developed in detail or it is developed in a setting different than ours. Fortunately it is straightforward to prove in our setting, so rather than adapting standard calculus of variations results, we will instead give a self-contained derivation of the action principle, stated in our specific space, from the Newtonian equations of motion. This makes the presentation somewhat unconventional (cf. the classic *Mechanics* by Landau and Lifshitz [62]), but it yields rigorous underpinnings for our analysis relatively quickly.

There is no cost in generalizing to general potential functions, so we do so. Suppose  $n$  point particles with positive masses  $m_1, \dots, m_n$ , positions  $x^{(1)}, \dots, x^{(n)}$ , and velocities  $\dot{x}^{(1)}, \dots, \dot{x}^{(n)}$  interact via a potential function  $V(x^{(1)}, \dots, x^{(n)})$ . For convenience write  $x = (x^{(1)}, \dots, x^{(n)})$ . Suppose that  $V$  is  $C^1$  on an open domain  $\Omega \subset (\mathbb{R}^d)^n$ . If  $x \in \Omega$ , then the kinetic energy is  $T = \sum_{i=1}^n \frac{1}{2} m_i |\dot{x}^{(i)}|^2$ , the potential energy is  $V$ , and the dynamics are governed by the differential equations

$$\ddot{x}_c^{(i)} = -\frac{1}{m_i} \cdot \frac{\partial V}{\partial x_c^{(i)}}. \quad (4.1)$$

If  $x^{(1)}, \dots, x^{(n)}: \mathbb{R} \rightarrow \mathbb{R}^d$  are 1-periodic  $C^1$  functions (not necessarily satisfying (4.1)) such that

$$x(t) := (x^{(1)}(t), \dots, x^{(n)}(t)) \in \Omega \text{ for all } t \in \mathbb{R}, \quad (4.2)$$

then we can define the action  $S = \int_0^1 (T - V) dt$ .

Now recall the Banach space  $\mathcal{X}$  defined in §2.4, consisting of mean-zero 1-periodic functions and endowed with a certain  $\ell^1$  norm in terms of the Fourier coefficients. For use in this section only, define the augmentation  $\widehat{\mathcal{X}} = \mathcal{X} \oplus \mathbb{R}^d$  of  $\mathcal{X}$  that allows periodic functions with nonzero mean. Endow first  $\widehat{\mathcal{X}}$  and then the product space  $\widehat{\mathcal{X}}^n$  with the  $\ell^1$  product norm. Let  $U \subset \widehat{\mathcal{X}}^n$  be the subset (open, by Lemma 2.5) of functions  $(x^{(1)}, \dots, x^{(n)})$  satisfying (4.2).

Then we have the following.

**Proposition 4.1** (Action Principle for  $\widehat{\mathcal{X}}^n$ ). *The action defines a Fréchet differentiable function  $S: U \rightarrow \mathbb{R}$ . If  $x \in U$  satisfies  $DS(x) = 0$ , then the functions  $x^{(1)}(t), \dots, x^{(n)}(t)$  satisfy (4.1), i.e., they describe a physical trajectory of the system.*

*Proof.* Fix  $x \in U$  and, for  $i = 1, \dots, n$  and  $c = 1, \dots, d$ , write the Fourier series for  $x_c^{(i)}(t)$  as

$$x_c^{(i)}(t) = a_0^{(i,c)} + 2 \sum_{k=1}^{\infty} \left( a_k^{(i,c)} \cos 2\pi kt + b_k^{(i,c)} \sin 2\pi kt \right).$$

As in §2.5, the kinetic part of the action can be written in terms of these Fourier coefficients:

$$S = 4\pi^2 \sum_{i=1}^n m_i \sum_{k=1}^{\infty} k^2 \sum_{c=1}^d \left( (a_k^{(i,c)})^2 + (b_k^{(i,c)})^2 \right) - \int_0^1 V(x(t)) dt.$$

The partial derivatives of  $S$  with respect to the constant terms  $a_0^{(i,c)}$  exist and are continuous. For  $k \geq 1$ , the partial derivatives of  $S$  with respect to the scaled coefficients  $ka_k^{(i,c)}, kb_k^{(i,c)}$  exist (continuously) and are given by, e.g.,

$$\frac{\partial S}{\partial (ka_k^{(i,c)})}(x) = 8\pi^2 m_i (ka_k^{(i,c)}) - \frac{2}{k} \int_0^1 \frac{\partial V}{\partial x_c^{(i)}} \cos 2\pi kt dt. \quad (4.3)$$

The remaining integral above is a Fourier coefficient of a continuous function which depends continuously on the position  $x_c^{(i)}$ . Thus the Fourier coefficients are uniformly bounded in  $k$  and in a neighborhood of  $x \in U$  (even before being divided by  $k$ ). The first term is also uniformly bounded in a neighborhood, as  $ka_k^{(i,c)}$  is a summand in the definition of  $\|x_c^{(i)}\|$ . But  $\widehat{\mathcal{X}}^n$  is an  $\ell^1$  space with respect to these coefficients, so uniform boundedness of the continuous partial derivatives implies Fréchet differentiability.

Now suppose  $DS(x) = 0$ . Then, in particular, each partial derivative of  $S$  vanishes. Reading from (4.3), combining the two equations for  $a_k^{(i,c)}$  and  $b_k^{(i,c)}$  into one complex equation, and rewriting the Fourier coefficients as integrals, this means that, for all  $k \geq 0$ ,

$$\int_0^1 4\pi^2 m_i k^2 x_c^{(i)}(t) e^{2\pi ikt} dt = \int_0^1 \frac{\partial V}{\partial x_c^{(i)}} e^{2\pi ikt} dt.$$

The left-hand side is essentially a Fourier coefficient of  $\ddot{x}_c^{(i)}$ , but we do not yet know that  $x_c^{(i)}$  is twice-differentiable. We could skirt this issue by working in  $L^2(\mathbb{R})$ , but instead we will spend a moment to complete the analysis without leaving the realm of continuous functions.

Recalling that  $x_c^{(i)}$  is  $C^1$  (because it is in  $\widehat{\mathcal{X}}$ ), we can integrate by parts on the left, differentiating  $x_c^{(i)}(t)$  and integrating  $\exp(2\pi ikt)$ . We also integrate by parts on the right side, but this time differentiate  $\exp(2\pi ikt)$  and integrate the partial derivative. Defining

$$W_c^{(i)}(t) = \int_0^t \frac{\partial V}{\partial x_c^{(i)}}(x(\tau)) d\tau,$$

the result of these two integration by parts operations is

$$\int_0^1 2\pi i k m_i \dot{x}_c^{(i)}(t) e^{2\pi ikt} dt = - \int_0^1 2\pi i k W_c^{(i)}(t) e^{2\pi ikt} dt.$$

Note in particular that the extra terms generated by the integration by parts all vanish: the term on the left vanishes by periodicity, and the term on the right vanishes because  $W_c^{(i)}(1) = 0$  is exactly the  $\partial S / \partial a_0^{(i,c)} = 0$  condition. Collecting terms, we have

$$\int_0^1 \left( m_i \dot{x}_c^{(i)}(t) + W_c^{(i)}(t) \right) e^{2\pi ikt} dt = 0$$

for all  $k \geq 1$ . Hence  $m_i \dot{x}_c^{(i)}(t) = -W_c^{(i)}(t) + C_c^{(i)}$  for some constant  $C_c^{(i)}$ . The right-hand side

is continuously differentiable, so the left-hand side is as well. Taking the time derivative, we conclude, as desired, that

$$m_i \ddot{x}_c^{(i)} = -\frac{\partial V}{\partial x_c^{(i)}}. \quad \square$$

Next we show that symmetries of the system can be incorporated into the action principle to limit the directions in which  $S$  must be stationary. The following two results may be viewed as simple instantiations of the Palais “principle of symmetric criticality” [76]; the first concerns continuous symmetry and the second concerns discrete symmetry. They are really independent applications of the Palais principle, but to avoid some mild notational complications we will state the second in a way dependent on the first.

The first of the two results is an easy consequence of the following general symmetry principle.

**Lemma 4.2.** *Let  $G$  be a Lie group acting on a Banach space  $X$ . Let  $D \subset X$  be an open set and let  $\psi: D \rightarrow \mathbb{R}$  be a Fréchet differentiable function which is  $G$ -invariant, i.e., which satisfies  $\psi(g \cdot d) = \psi(d)$  for all  $d \in D$  and  $g \in G$  such that  $g \cdot d \in D$ . Let  $\pi: X \rightarrow F$  be a projection of  $X$  onto a finite-dimensional space  $F$ . Let  $\text{Lie}(G)$  be the Lie algebra of  $G$ , and suppose  $x \in D$  satisfies  $\pi(\text{Lie}(G) \cdot x) = F$ . If the restriction of  $D\psi(x)$  to  $\ker \pi$  vanishes, then  $D\psi(x) = 0$ .*

*Proof.* Fixing  $\gamma \in \text{Lie}(G)$ , for all sufficiently small  $t$  we have  $\exp(t\gamma) \cdot x \in D$  and thus  $\psi(\exp(t\gamma) \cdot x) = \psi(x)$ . It follows that  $D\psi(x)(\gamma \cdot x) = 0$ , so  $\text{Lie}(G) \cdot x \subset \ker D\psi(x)$ . But also  $\ker \pi \subset \ker D\psi(x)$ . The restriction of  $\pi$  to  $\text{Lie}(G) \cdot x$  is surjective, by assumption, so  $(\text{Lie}(G) \cdot x) + (\ker \pi) = X$ . This proves that  $\ker D\psi(x) = X$ , as desired.  $\square$

To apply this to our present scenario, let  $(\widehat{\mathcal{X}}^n)_0$  denote the “mean-zero” subspace of  $\widehat{\mathcal{X}}^n$ , i.e., the space of functions  $(x^{(1)}, \dots, x^{(n)})$  such that  $\sum_{i=1}^n \int_0^1 x^{(i)}(t) dt = 0$ . (Note how this differs from  $\mathcal{X}^n$ : that space consists of periodic trajectories in which each body has mean zero, while  $(\widehat{\mathcal{X}}^n)_0$  consists of periodic trajectories in which the mean (over the  $n$  bodies) of the mean positions (of each body separately) is zero.) This is a closed subspace of  $\widehat{\mathcal{X}}^n$  and so is itself a Banach space. Let  $U_0 = U \cap (\widehat{\mathcal{X}}^n)_0$ ; then  $S$  defines a Fréchet differentiable function  $S_0: U_0 \rightarrow \mathbb{R}$ .

**Corollary 4.3.** *Suppose the potential  $V$  is translation-invariant. If  $(x^{(1)}, \dots, x^{(n)}) \in U_0$  satisfies  $DS_0(x^{(1)}, \dots, x^{(n)}) = 0$ , then the functions  $x^{(1)}(t), \dots, x^{(n)}(t)$  satisfy (4.1), i.e., they describe a physical trajectory of the system.*

*Proof.* Let  $\pi: \widehat{\mathcal{X}}^n \rightarrow \mathbb{R}^d$  be the projection onto the mean, i.e.,

$$\pi(x^{(1)}, \dots, x^{(n)}) = \sum_{i=1}^n \int_0^1 x^{(i)}(t) dt.$$

Then  $\ker \pi = (\widehat{\mathcal{X}}^n)_0$ , and the restriction of  $DS(x)$  thereto vanishes by assumption. Let  $G$  be the  $d$ -dimensional translation group on  $\mathbb{R}^d$ . Then  $\text{Lie}(G) \cdot x$  is the space of constant functions which surjects by  $\pi$  onto  $\mathbb{R}^d$ . Thus  $DS(x) = 0$  by Lemma 4.2, whence Proposition 4.1 applies.  $\square$

Now let  $G$  be a finite group and let  $\rho$  be a representation of  $G$  on  $(\widehat{\mathcal{X}}^n)_0$  which fixes the action, i.e., for all  $g \in G$  the map  $\rho(g): (\widehat{\mathcal{X}}^n)_0 \rightarrow (\widehat{\mathcal{X}}^n)_0$  is a linear isomorphism fixing  $U_0$  and satisfying  $S_0(\rho(g) \cdot x) = S_0(x)$  for all  $i = 1, \dots, r$  and all  $x \in U_0$ . Several sources (e.g.,



Chenciner [17, p. 77]) work in a Hilbert space and assume that  $\rho$  is a unitary representation, but this is not necessary. (And indeed, we will apply this theory with representations which fix the action but do not act as isometries on  $\mathcal{X}^n$ : because we use the  $\ell^1$  norm, time translation is not an isometry.) Let  $(\widehat{\mathcal{X}}^n)_0^\rho$  be the subspace of  $(\widehat{\mathcal{X}}^n)_0$  fixed under  $\rho$ ; this is a closed subspace and so it is a Banach space. Set  $U_0^\rho = U \cap (\widehat{\mathcal{X}}^n)_0^\rho$ . Then the action defines a Fréchet differentiable function  $S_0^\rho: U_0^\rho \rightarrow \mathbb{R}$ .

**Corollary 4.4.** *Suppose the potential  $V$  is translation-invariant. If  $x \in U_0^\rho$  satisfies  $DS_0^\rho(x) = 0$ , then the functions  $x^{(1)}(t), \dots, x^{(n)}(t)$  satisfy (4.1), i.e., they describe a physical trajectory of the system.*

*Proof.* Fix for the moment  $g \in G$ . For all  $\delta x \in (\widehat{\mathcal{X}}^n)_0$ , sufficiently small so that  $x + \delta x \in U_0$ , we have  $S_0(\rho(g) \cdot (x + \delta x)) = S_0(x + \delta x)$ . It follows that  $D(S_0(\rho(g) \cdot x)) = DS_0(x)$ , or  $DS_0(\rho(g) \cdot x) \circ \rho(g) = DS_0(x)$ . But  $x$  is fixed by  $\rho(g)$ , so the above equation reads  $DS_0(x) \circ \rho(g) = DS_0(x)$ .

Now, averaging over  $g \in G$ , we have

$$DS_0(x) = DS_0(x) \circ \left( \frac{1}{|G|} \sum_{g \in G} \rho(g) \right).$$

But the image of  $\frac{1}{|G|} \sum_{g \in G} \rho(g)$  is  $(\widehat{\mathcal{X}}^n)_0^\rho$ , and the restriction of  $DS_0(x)$  to this subspace is  $DS_0^\rho(x) = 0$ . Thus the right-hand side is identically zero, so the left is as well, whence Corollary 4.3 applies.  $\square$

We close this section by applying these principles to our case. Recall the setup of §2.5. We no longer use  $x$  to refer to  $(x^{(1)}, \dots, x^{(n)}) \in \widehat{\mathcal{X}}^n$ , but rather we have made the mean-zero choreographic assumption, so that the entire system is described by one function  $x \in \mathcal{X}$ . We also restore the letter  $S$  as denoting the choreographic action in (2.2). Finally, we recall the definition of  $U_{cf}$  as the “collision-free” subset of  $\mathcal{X}$ .

**Corollary 4.5** (Action Principle for  $\mathcal{X}$ ). *The action defines a Fréchet differentiable function  $S: U_{cf} \rightarrow \mathbb{R}$ . If  $x \in U_{cf}$  satisfies  $DS(x) = 0$ , then the path determined by  $x(t)$  is a physical choreography, i.e., it obeys the equations of motion.*

*Proof.* The Newtonian potential  $V$  is translation-invariant and  $C^1$  (in fact, smooth) on the domain  $\Omega = \{(x^{(1)}, \dots, x^{(n)}) : x^{(i)} \neq x^{(j)} \text{ for all } i \neq j\}$ , so we may apply the above theory to it. Let  $\rho$  be the representation of  $\mathbb{Z}/n\mathbb{Z}$  on  $(\widehat{\mathcal{X}}^n)_0$  given by cyclic shift of the bodies and cyclic rotation in time:

$$(\rho(1) \cdot (x^{(1)}, \dots, x^{(n)}))(t) = (x^{(n)}(t + 1/n), x^{(1)}(t + 1/n), \dots, x^{(n-1)}(t + 1/n)).$$

Then  $(\widehat{\mathcal{X}}^n)_0^\rho$  is the space of mean-zero choreographies, i.e., it is isomorphic to  $\mathcal{X}$ . The conclusion now follows from Corollary 4.4.  $\square$

We could have stated Corollary 4.5 in a manner that would allow for additional symmetries, at the mild cost of some notational burden. While this could be useful (see §3.5), in this document we will not apply the result in any such way, so we leave the straightforward generalization to the reader.

## 4.2 Expressions for the gradient and Hessian

As stated in Corollary 4.5, the action is Fréchet differentiable. The derivative is a linear functional  $DS(x): \mathcal{X} \rightarrow \mathbb{R}$ , and the natural norm on such functionals is the operator norm (also known as the dual norm). Because the norm on  $\mathcal{X}$  is just the  $\ell^1$  norm in the variables  $ka_k^{(c)}, kb_k^{(c)}$ , the dual norm of  $DS(x)$  is the  $\ell^\infty$  norm of the partial derivatives of  $S$  with respect to said variables.

However, the norm we want to use is much stronger; as defined in (2.3), it is the  $\ell^1$  norm of those partial derivatives instead of the  $\ell^\infty$  norm. The motivation for this is follows. Consider only the kinetic part of the action; reading from (2.2), this is  $S_T = \frac{1}{2} \sum_{k=1}^{\infty} k^2 \sum_{c=1}^d |A_k^{(c)}|^2$ . The partial derivative of this with respect to (e.g.)  $ka_k^{(c)}$  is just  $ka_k^{(c)}$ , and so the (infinite) matrix of second partial derivatives is just the identity. To state this more formally, equip  $\mathcal{X}$  with the basis corresponding to the variables  $\{ka_k^{(c)}, kb_k^{(c)}\}$ , and equip  $\mathcal{X}^*$  with the basis corresponding to the coordinate functions giving the coefficients of the same variables. The former space has the  $\ell^1$  norm and the latter has the  $\ell^\infty$  norm. Using the natural identification between these bases,  $D^2S_T: \mathcal{X} \rightarrow \mathcal{X}^*$  is just the inclusion of  $\ell^1$  in  $\ell^\infty$ . This map is certainly continuous, but it is *not* continuously invertible;  $\ell^\infty$  does not embed in  $\ell^1$ .

This is critical because we need a bounded inverse in order to apply Theorem I.2.1. Consider replacing  $\mathcal{X}^*$ , the space of bounded linear functionals on  $\mathcal{X}$ , with

$$\mathcal{X}_1^* := \{T: \mathcal{X} \rightarrow \mathbb{R} : \|T\|_1 < \infty\},$$

where  $\|T\|_1$  is the  $\ell^1$  norm defined in (2.3). As a function from  $\mathcal{X}$  to  $\mathcal{X}_1^*$ ,  $D^2S_T$  is still defined, and moreover it is a linear isomorphism identifying the two  $\ell^1$  spaces. Up to equivalence, this is the only choice of norm on the dual space that would let us apply our effective existence theorems.

Of course, not every Fréchet differentiable function  $F: \mathcal{X} \rightarrow \mathbb{R}$  has the property that  $\|DF\|_1 < \infty$ . For want of a better term, we refer to such functions as  $\ell^1$ -differentiable.

**Proposition 4.6.** *The action is an  $\ell^1$ -differentiable function  $S: U_{cf} \rightarrow \mathbb{R}$ . Its partial derivatives are given by*

$$\frac{\partial S}{\partial(ka_k^{(c)})}(x) + i \frac{\partial S}{\partial(kb_k^{(c)})}(x) = kA_k^{(c)} + \frac{1}{k} \sum_{j=1}^{n-1} \int_0^1 \frac{x_c(t+j/n) - x_c(t)}{|x(t+j/n) - x(t)|^3} e^{2\pi ikt} dt. \quad (4.4)$$

*Proof.* Consider first the partial derivative with respect to  $ka_k^{(c)}$ . Reading (2.2), we find

$$\frac{\partial S}{\partial(ka_k^{(c)})}(x) = ka_k^{(c)} - \frac{1}{2k} \sum_{j=1}^{n-1} \int_0^1 \frac{x_c(t+j/n) - x_c(t)}{|x(t+j/n) - x(t)|^3} \cdot (\cos 2\pi k(t+j/n) - \cos 2\pi kt) dt.$$

Break the integrals above each into two terms and then time-shift the first of each pair, i.e., the integrals with  $\cos 2\pi k(t+j/n)$  in the integrand. Recombining the two integrals, we have

$$ka_k^{(c)} + \frac{1}{2k} \sum_{j=1}^{n-1} \int_0^1 \left( \frac{x_c(t-j/n) - x_c(t)}{|x(t-j/n) - x(t)|^3} + \frac{x_c(t+j/n) - x_c(t)}{|x(t+j/n) - x(t)|^3} \right) \cdot \cos 2\pi kt dt.$$

By periodicity, the sums over  $j$  of the two terms in the integrand are equal. This proves equality in the real part of (4.4). Computing the partial derivative with respect to  $kb_k^{(c)}$

similarly verifies the imaginary part of the assertion.

It remains to check that  $S$  is  $\ell^1$ -differentiable, i.e., that the partial derivatives in (4.4) are summable. Summability of the first (kinetic) term follows from the definition of  $\mathcal{X}$ . Consider now the second (potential) term. Define

$$F_c(t) = \sum_{j=1}^{n-1} \frac{x_c(t+j/n) - x_c(t)}{|x(t+j/n) - x(t)|^3}. \quad (4.5)$$

Being in  $\mathcal{X}$ , the function  $x$  is  $C^1$ ; being in  $U_{cf}$ , the denominators in the definition of  $F_c(t)$  are nonzero and so it too is  $C^1$ . Thus its  $k^{\text{th}}$  Fourier coefficient is  $O(1/k)$  by the Riemann-Lebesgue lemma. But then the corresponding term in  $\partial S/\partial(ka_k^{(c)})$ , i.e., the  $k^{\text{th}}$  Fourier coefficient of  $F_c(t)$  divided by  $k$ , is  $O(1/k^2)$ . This is indeed summable.  $\square$

**Definition 4.7.** To disambiguate the space of definition, when we think of it as an element of  $\mathcal{X}_1^*$  we denote the derivative  $DS(x)$  by  $\nabla S(x)$ .

We will sometimes identify the space  $\mathcal{X}_1^*$  with  $\ell^1$ , using the basis of projections onto the scaled Fourier coefficients  $ka_k^{(c)}, kb_k^{(c)}$ . Using the identifications of both with  $\ell^1$ , we can also identify  $\mathcal{X}_1^*$  with  $\mathcal{X}$ . The map  $D^2S_T: \mathcal{X} \rightarrow \mathcal{X}_1^*$  discussed above is the identity with respect to this identification.

We now consider the second derivative, i.e., the derivative of the gradient. Fix  $x \in U_{cf}$ , and for all  $c, c' \in \{1, \dots, d\}$  and  $j \in \{1, \dots, n-1\}$ , define

$$G_j^{(c,c')}(t) = -3 \cdot \frac{(x_{c'}(t+j/n) - x_{c'}(t))(x_c(t+j/n) - x_c(t))}{|x(t+j/n) - x(t)|^5} + \delta_{cc'} \cdot \frac{1}{|x(t+j/n) - x(t)|^3}, \quad (4.6)$$

where  $\delta_{cc'}$  is the Kronecker delta. Let  $\widehat{G}_j^{(c,c')}(k) = \int_0^1 G_j^{(c,c')}(t)e^{2\pi ikt} dt$  denote the  $k^{\text{th}}$  Fourier coefficient of  $G_j^{(c,c')}(t)$ .

**Proposition 4.8.** *The gradient of the action,  $\nabla S: U_{cf} \rightarrow \mathcal{X}_1^*$ , is continuously Fréchet differentiable. Its partial derivatives are computed as follows. Fix  $c, c' \in \{1, \dots, d\}$  and  $k, k' \geq 1$  and set*

$$\begin{aligned} \Sigma_+ &= (\Sigma_+)_{k,k'}^{c,c'} = \sum_{j=1}^{n-1} \left( e^{2\pi i k' j/n} - 1 \right) \widehat{G}_j^{(c,c')}(k' + k), \\ \Sigma_- &= (\Sigma_-)_{k,k'}^{c,c'} = \sum_{j=1}^{n-1} \left( e^{2\pi i k' j/n} - 1 \right) \widehat{G}_j^{(c,c')}(k' - k). \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial^2 S}{\partial(k'a_{k'}^{(c')})\partial(ka_k^{(c)})} + i \frac{\partial^2 S}{\partial(k'b_{k'}^{(c')})\partial(ka_k^{(c)})} &= \delta_{cc'} \delta_{kk'} + \frac{1}{kk'} (\Sigma_- + \Sigma_+), \\ -i \frac{\partial^2 S}{\partial(k'a_{k'}^{(c')})\partial(kb_k^{(c)})} + \frac{\partial^2 S}{\partial(k'b_{k'}^{(c')})\partial(kb_k^{(c)})} &= \delta_{cc'} \delta_{kk'} + \frac{1}{kk'} (\Sigma_- - \Sigma_+). \end{aligned} \quad (4.7)$$

*Proof.* By differentiating (4.4), a straightforward (albeit tedious) computation verifies the veracity of (4.7). The derivative  $D(\nabla S)$  should be a linear operator from  $\mathcal{X}$ , an  $\ell^1$  space, to  $\mathcal{X}_1^*$ , another  $\ell^1$  space. The partial derivatives exist and are continuous, so to check that  $\nabla S$

is  $C^1$ , it suffices to show that, in some neighborhood of any  $x \in U_{\text{cf}}$ , the linear functions defined by the partial derivatives are uniformly continuous (i.e., bounded). Recall that the  $\ell^1 \rightarrow \ell^1$  operator norm of a matrix is given by taking the maximum over columns of the  $\ell^1$  norm of each column. Thus it suffices to show that (in a neighborhood) there is some uniform bound  $B$  such that, if we fix a variable  $k'a_{k'}^{(c')}$  or  $k'b_{k'}^{(c')}$  (fixing a “column”), then the  $\ell^1$  norm of the partial derivatives over all  $c, k$  and both parts (summing over a “row”) is bounded by  $B$ . But each auxiliary function  $G_j^{(c,c')}$  is a  $C^1$  function of time and its time derivative varies continuously in  $U_{\text{cf}}$ . Thus the Fourier coefficients  $\widehat{G}_j^{(c,c')}(\ell)$  are uniformly  $O(1/\ell)$ ; uniform boundedness follows easily therefrom.  $\square$

In the same way as we abuse notation by calling  $\nabla S$  the gradient, we often call its derivative  $D(\nabla S)$  the “Hessian” and denote it by  $HS$ . Having identified both the domain and codomain with  $\ell^1$ , we can think of  $HS(x)$  as a matrix, with rows and columns indexed by the coordinates  $ka_k^{(c)}, kb_k^{(c)}$ . (More specifically, the columns and rows are indexed by the Fourier modes and the projections onto those modes, respectively.) To be even more explicit, we now choose an ordering for these bases: we order first by mode (ascending), and then by coordinate. This fixes an ordering to the entries in  $HS(x)$ , so that we may talk, e.g., about the “upper-left” submatrix.

The operator norm on the Hessian is the  $\ell^1/\ell^1$  matrix norm, which is computed by taking the supremum (over columns) of the  $\ell^1$  norm of each column. This is easy and efficient to compute given any explicit (finite) matrix.

### 4.3 Computing bounds on functions and their Fourier coefficients

The formulae in Propositions 4.6 and 4.8 show that there are finitely many auxiliary functions such that every term in the gradient and Hessian (coming from the potential energy) can be understood as a Fourier coefficient of one of them. Rigorously bounding these coefficients is essential to the proof. While the computation of such bounds can be viewed as implementation-specific, it is important for the proof to understand what bounds are available to us. We will give an abbreviated treatment now.

The fundamental building block for everything we do (in the course of our proof technique) is the periodic function. As input we take truncated Fourier series, which we treat as representing trigonometric polynomials. We then define various other periodic functions in terms of these exactly-representable trigonometric polynomials using basic arithmetic and square-root extraction; let us call any periodic function which is defined in this way *expressible*.

We first and foremost need to be able to control the values of the expressible functions. With some foresight we consider them not as functions of a real variable, but rather as functions on a strip of the complex plane. We shall have occasion to evaluate the functions not on  $\mathbb{R}$ , but instead on some horizontal line  $\mathbb{R} + ih$ ; we refer to  $h$  as the “height” in such cases.

The following lemma is simple-minded but effective.

**Lemma 4.9.** *Let  $f(\xi) = \sum_{k=-K}^K A_k e^{-2\pi i k \xi}$  be a 1-periodic trigonometric polynomial. Fix real numbers  $a \leq b$ . On the strip  $\Omega_a^b := \{\xi \in \mathbb{C} : a \leq \text{Im } \xi \leq b\}$ ,*

$$|f'(\xi)| \leq B_f(a, b) := 2\pi \sum_{k=1}^K k(|A_k|e^{-2\pi k a} + |A_{-k}|e^{2\pi k b}).$$

Thus, for any  $z \in \Omega_a^b$  and all  $\xi \in \Omega_a^b$  such that  $|\operatorname{Re}(\xi - z)| \leq \frac{\varepsilon}{2}$ ,

$$|f(\xi) - f(z)| \leq B_f(a, b) \cdot \sqrt{\left(\frac{\varepsilon}{2}\right)^2 + (\max\{b - \operatorname{Im} z, \operatorname{Im} z - a\})^2}. \quad (4.8)$$

The proof of the lemma is immediate. Using it, we can state the first *bound* of this section; by a “bound” here we mean an outline of a computational algorithm for computing mathematical bounds. More specifically, these algorithms compute rigorous bounds that can be made arbitrarily close to the true value at the cost of increased runtime.

**Bound 1.** *If a given expressible function is analytic on a given strip, then we can rigorously verify its analyticity and compute bounds for the function values on that strip.*

*Method.* First consider the case of trigonometric polynomials. Given such a function, represented by its Fourier coefficients, use a discrete Fourier transform to get (exact) function values at  $N$  regularly-spaced sample points. This can be done at any desired height  $h$  by first scaling the Fourier coefficients by  $e^{-2\pi kh}$ . Then apply (4.8) with  $\varepsilon = 1/N$  to get rigorous bounds on rectangles which together cover the whole strip  $\Omega_a^b$ . Note that, by increasing  $N$  and/or subdividing the interval  $[a, b]$  and bounding the strip associated to each subinterval separately, these bounds can achieve any desired accuracy.

For a general expressible function, first divide the strip into rectangles and compute bounds on the underlying trigonometric polynomials as above; this produces complex balls containing the image of each rectangle. Now compute the values of the expressible function on each rectangle by using “complex interval arithmetic” to find, after each arithmetic operation, a complex ball guaranteed to contain the true values. If this cannot be done analytically, for instance if we need to divide and the ball representing the denominator contains zero, then subdivide the rectangles and try again.  $\square$

*Remark.* Here and throughout, when we talk about a function being analytic on a closed set, we mean that it is analytic on some neighborhood thereof.

**Bound 2.** *We can compute rigorous bounds for any  $L^p$  norm of any expressible function at any height (at which the function is analytic).*

*Method.* Immediate from the previous bound.  $\square$

*Aside.* Analyticity of division fails if and only if the denominator is zero, so there is no question of how to interpret that. There is some question in the treatment of square-root extraction, though. Because strips are simply connected, a function has an analytic square root as long as it is never zero. However, since we want to actually compute the square root, it is useful to commit to a branch instead of having to determine a valid square root at runtime. We use the principal square root, and thus to verify analyticity we have to check that the function never touches the negative real axis (in addition to never touching zero).

Our real goal is to bound the Fourier coefficients of expressible functions. The most obvious way is to use Parseval’s identity: writing  $\widehat{f}(k) := \int_0^1 f(t)e^{2\pi ikt} dt$  for the  $k^{\text{th}}$  Fourier coefficient of a 1-periodic function  $f$ , we have  $\sum_{k \in \mathbb{Z}} |\widehat{f}(k)|^2 = \int_0^1 |f(t)|^2 dt$ .

**Bound 3.** *If  $f$  is an expressible function with Fourier coefficients  $\widehat{f}(k)$ , then we can compute a bound on  $\sum |\widehat{f}(k)|^2 = \int_0^1 |f(t)|^2 dt$ . Moreover, this is true even if  $f$  (or perhaps the functions it is defined in terms of) is (are) only known up to some error in the uniform norm.*

*Method.* Any starting error in the function values can be added to the error in (4.8) and then propagated through the computation. Doing so, rigorously upper bound  $\int_0^1 |f(t)|^2 dt$  using the above method and then apply Parseval.  $\square$

The addendum concerning additional error applies in particular if the underlying trigonometric polynomials approximate some element of  $\mathcal{X}$  closely (as measured by the norm thereon). We will use this to bound the change in the Hessian within an  $\varepsilon$  ball of our starting point.

We could in principle compute all of the bounds we need using Parseval's identity applied either to our functions or their time derivatives (see §5.3). In fact, if we started with arbitrary coordinate functions in  $\mathcal{X}$ , then we would only know  $C^1$  regularity and so Parseval would be essentially the only bound available to us. However, our initial coordinate functions have significantly more structure to them, and in particular there is a much better bound available for them. Namely, we have the following simple-but-powerful result (compare with Theorem 8.2.2 in Brass and Petras [15, p. 267]).

**Lemma 4.10.** *Let  $f(t)$  be a 1-periodic function which has analytic continuation to a strip  $\Omega_0^h$  (as defined in Lemma 4.9). Let  $\hat{f}(k) = \int_0^1 f(t)e^{2\pi ikt} dt$  be the  $k^{\text{th}}$  Fourier coefficient. Then*

$$|\hat{f}(k)| \leq \left( \int_0^1 |f(t+ih)| dt \right) e^{-2\pi hk}.$$

*Proof.* Consider integrating the analytic function  $f(t)e^{2\pi ikt}$  around the perimeter of the rectangle  $[0, 1] \times [0, h] \subset \mathbb{R}^2 = \mathbb{C}$ . By Cauchy's integral theorem, this line integral vanishes. That is, we have

$$\int_0^1 f(x)e^{2\pi ikx} dx + \int_0^h f(1+iy)e^{2\pi ik(1+iy)} dy + \int_1^0 f(x+ih)e^{2\pi ik(x+ih)} dx + \int_h^0 f(iy)e^{2\pi ik(iy)} dy$$

equals 0. The second and fourth integrals cancel. That leaves us with

$$\hat{f}(k) = \int_0^1 f(x)e^{2\pi ikx} dx = \int_0^1 f(x+ih)e^{2\pi ik(x+ih)} dx = e^{-2\pi hk} \int_0^1 f(x+ih)e^{2\pi ikx} dx,$$

whence the triangle inequality completes the proof.  $\square$

*Remark.* We work with real-valued functions, so the negative-mode Fourier coefficients are just conjugates of the positive-mode coefficients. Thus Lemma 4.10 actually gives exponential decay bounds for all of the coefficients. If we did not have this conjugate symmetry, then we could just bound the negative-mode coefficients separately by considering  $\Omega_{-h}^0$ .

**Bound 4.** *Suppose  $f$  is an expressible function with analytic continuation in a neighborhood of the real line. Then we can compute any desired number of explicit Fourier coefficients, with rigorous error bounds, as well as an exponentially decaying bound on the remaining coefficients.*

*Method.* Lemma 4.10 gives the asserted exponentially decaying bound. As for the computation of explicit Fourier coefficients, compute  $N$  regularly-spaced samples of  $f$  on the real line and take their discrete Fourier transform. This gives approximations

$$\tilde{f}(k) = \frac{1}{N} \sum_{j=0}^{N-1} f\left(\frac{j}{N}\right) e^{2\pi ik(j/N)}$$

for the Fourier coefficients  $\widehat{f}(k) := \int_0^1 f(t)e^{2\pi ikt} dt$  of  $f$ . More specifically, the fast Fourier transform computes  $\widetilde{f}(k)$  for  $-\lfloor \frac{N-1}{2} \rfloor \leq k \leq \lceil \frac{N-1}{2} \rceil$ . Now  $f(t) = \sum_{k \in \mathbb{Z}} \widehat{f}(k)e^{-2\pi ikt}$ , so the output of the discrete Fourier transform is exactly

$$\left\{ \widetilde{f}(k) = \sum_{m \in \mathbb{Z}} \widehat{f}(k + Nm) \mid k = -\left\lfloor \frac{N-1}{2} \right\rfloor, \dots, \left\lceil \frac{N-1}{2} \right\rceil \right\}.$$

With the exponentially decaying (in particular, summable) bound on the higher Fourier coefficients, we can bound  $\sum_{m \neq 0} \widehat{f}(k + Nm)$ , which is the error  $\widetilde{f}(k) - \widehat{f}(k)$  in our approximation of the  $k^{\text{th}}$  Fourier coefficient.  $\square$

Note that we start with trigonometric polynomials, which are entire, and the auxiliary functions we study are holomorphic away from collisions. Thus the analytic continuation property is satisfied in all of the computations we make at our starting point (i.e., all of our computations except those in §4.7).

#### 4.4 Step 1: bounding the gradient

For the remainder of the section, fix a mean-zero, 1-periodic curve  $x_0: \mathbb{R} \rightarrow \mathbb{R}^d$  which we intend to prove is close to a choreography. (We do not use 0 as a coordinate subscript, so this notation does not technically collide with that of the coordinate functions  $x_c$ .) We assume that  $x_0$  is a trigonometric polynomial, i.e., it is represented by a finite Fourier series. In the notation of §2.4,  $A_k^{(c)} = 0$  for all  $c$  and all  $|k| > K$ . Note that this certainly implies  $x_0 \in \mathcal{X}$ . We also fix  $\varepsilon > 0$ , the radius in which we are going to try to prove existence of a choreography.

The first step towards our goal of being able to apply Theorem I.2.1 to prove existence of a choreography near  $x_0$  is bounding the norm of the gradient, i.e.,  $\|\nabla S(x_0)\|$ . This is easy.

First, following Bound 1, we compute a height  $h$  such that the distance functions  $|x_0(t + j/n) - x_0(t)|$  admit analytic continuation to the strip  $\Omega_0^h$ . This is always possible, unless  $x_0$  contains a collision. This process only needs to be done once, because all of the functions we consider have analytic continuation wherever the distance functions do. Next, using Bound 4, we compute bounds on the Fourier coefficients of the auxiliary functions  $F_c$  defined in (4.5). We take care to compute explicit values, with error, for at least the coefficients with modes  $-K, \dots, K$ . Then, using Proposition 4.6, we directly compute the entries of the gradient in the modes for which we have explicit Fourier coefficients. We bound the remaining entries (the “tail”) using the exponential decay estimate. Combining the explicit entries, errors on those explicit entries, and the tail bound gives a rigorous upper bound on  $\|\nabla S(x)\|$ .

#### 4.5 Step 2: accounting for symmetries

The next step is to bound the inverse of the Hessian, but in this we encounter a serious problem: at a stationary point of the action, the Hessian is not actually invertible. The reason for this is that there is a positive-dimensional group acting on  $\mathcal{X}$  via symmetries of the action:  $O(2) \times O(d)$ , acting via time translation and spatial isometry.

Indeed, if  $x_* \in \mathcal{X}$  is a choreography, then  $\nabla S(x_*) = 0$ . For all  $g \in O(2) \times O(d)$ ,  $g \cdot x_*$  is also a choreography, so  $\nabla S(g \cdot x_*) = 0$ . But this implies  $D(\nabla S)(x_*)(\text{Lie}(O(2) \times O(d)) \cdot x_*) = 0$ .

In particular, unless  $x_*$  has positive-dimensional stabilizer (i.e., unless  $x_*$  is the circular orbit), the Hessian has a null space of dimension  $\dim(\text{Lie}(\text{O}(2) \times \text{O}(d))) = 1 + \binom{d}{2}$ .

Generally speaking, the solution to this problem is to break symmetry. There are several specific ways to handle it (see §6.3), but we will use a particularly simple one: we just drop a few variables.

More specifically, consider a finite subset  $\mathcal{I}$  of the variables  $\{ka_k^{(c)}, kb_k^{(c)}\}$  and let  $\pi: \mathcal{X} \rightarrow \mathbb{R}^{|\mathcal{I}|}$  be the projection onto them. We seek a subset of size  $|\mathcal{I}| = 1 + \binom{d}{2}$  such that, for all  $x \in \mathcal{X}$  satisfying  $\|x - x_0\| < \varepsilon$ , the restriction of  $\pi$  to the subspace  $\text{Lie}(\text{O}(2) \times \text{O}(d)) \cdot x$  is surjective. Surjectivity is an open condition, so this can be readily checked by computer.

*Remark.* If we allow ourselves to decrease  $\varepsilon$  as needed, then it is possible to find such a set if and only if  $x_0$  has zero-dimensional stabilizer in  $\text{O}(2) \times \text{O}(d)$ . In particular, as long as we are willing to discard the circular orbit, this search will succeed. If we did want to handle degenerate cases (namely, the circular orbit), then we would just need to detect the dimension of the stabilizer numerically and adjust the size of the set  $\mathcal{I}$  accordingly.

Having identified such a set of variables, in principle we delete them from consideration in all future endeavors. (In practice, though, we just hack around them (see §5.2).) If our computer-assisted proof succeeds, then we apply Lemma 4.2 to the resulting solution to see that it is indeed a choreography.

*Aside.* In the context of Theorem I.2.1, we can think of variable dropping in two different ways. One way is as described, by removing the variables from consideration altogether (mathematically, we are quotienting  $\mathcal{X}$  at the start by the corresponding finite-dimensional subspace). Another way is to project out the variables from  $\mathcal{X}_1^*$  only, utilizing the flexibility in Theorem I.2.1 that we only need a *right* inverse, not a two-sided inverse. From this perspective, the first approach corresponds to choosing the right inverse in which the rows corresponding to the deleted variables are all zero. In principle one might wish to choose a different right inverse, and so the second perspective is more flexible. On the other hand, the approach we use is more conducive to applying Proposition I.2.4, which is useful in principle (see §7.5). The reader is free to adopt whichever perspective feels more natural.

*Remark.* We are not able to rule out the possibility that the Hessian could still be singular after accounting for symmetries, but no such cases are known (cf. Problem 7.4). A hypothetical choreography with a singular Hessian would spell trouble for our computer-assisted proof technique.

## 4.6 Step 3: bounding the Hessian

To apply Theorem I.2.1 we need to (1) find a linear operator  $T: \mathcal{X}_1^* \rightarrow \mathcal{X}$ , which should approximate  $(HS(x_0))^{-1}$ ; (2) bound  $\|T\|$ ; and (3) bound  $\|HS(x) \circ T - \text{id}\|$  for all  $x \in \mathcal{X}$  such that  $\|x - x_0\| < \varepsilon$ , where  $\text{id}$  is the identity map on  $\mathcal{X}_1^*$ . We will break the last step into two parts: (3a) bound  $\|HS(x_0) \circ T - \text{id}\|$  and (3b) bound  $\|HS(x) - HS(x_0)\|$ . Steps (1), (2), and (3a) — the computations confined to our current point  $x_0$  — are the topic of this subsection.

Calculating the bounds described in this subsection is by far the most time-consuming part of the computation, and so we have made several crucial changes and optimizations in our implementation. The description of those is deferred to §5.1; here we describe the simplest, unoptimized, proof-of-concept approach.

As we worked out in §4.2, the contribution of kinetic energy to the Hessian is just the identity matrix. The contribution of potential energy is less ruly, but it decays with



the row and column mode. Thus we expect the Hessian to be eventually (ignoring a prefix of rows/columns) *diagonally dominant*. The plan is to exploit this to control the infinite-dimensional Hessian by a finite-dimensional approximation.

Having already verified analyticity on a strip in §4.4, we use Bound 4 to compute the Fourier coefficients of the auxiliary functions  $G_j^{(c,c')}$  defined in (4.6). More precisely, we compute (with error bounds) some explicit coefficients and an exponentially decaying bound for the rest. With this done, we can compute and/or bound any term of the Hessian matrix according to Proposition 4.8. Choose a cutoff  $\kappa \geq 1$  and let  $M$  denote the upper-left  $(2d\kappa) \times (2d\kappa)$  submatrix of  $HS(x_0)$ , i.e., the submatrix consisting of the modes  $k = 1, \dots, \kappa$ . We compute  $M$  explicitly and then invert it numerically.

Define  $T: \ell^1 \rightarrow \ell^1$  to be the matrix with  $M^{-1}$  in the upper-left block, a unit diagonal in the remaining rows/columns, and zero everywhere else. It is easy to compute  $\|T\|$ ; it is just  $\max\{\|M^{-1}\|, 1\}$ .

Our remaining task is to bound  $\|HS(x_0) \circ T - \text{id}\|$ . Let  $R$  denote the matrix obtained by taking  $HS(x_0)$ , subtracting the identity matrix, and then zeroing out the top-left  $(2d\kappa) \times (2d\kappa)$  submatrix. Then  $\|HS(x_0) \circ T - \text{id}\| \leq \|R\| \cdot \|T\|$ , so we can accomplish our goals by bounding the norm of  $R$ .

For each pair of modes  $k, k'$ , there is a corresponding  $2d \times 2d$  submatrix in  $R$  coming from the variables  $\{ka_k^{(c)}, kb_k^{(c)} : c = 1, \dots, d\}$  and  $\{k'a_{k'}^{(c)}, k'b_{k'}^{(c)} : c = 1, \dots, d\}$ . To simplify the exposition we ignore this here and just think of  $k, k'$  as specifying one entry in  $R$ . The persnickety reader should apply our discussion separately to the  $(2d)^2$  choices of coordinate and combine them by summing over the  $2d$  rows and taking the maximum over the  $2d$  columns.

We see in (4.7) that each entry in  $R$  (except those in the zeroed upper-left submatrix) is a combination of two parts: the  $\Sigma_-$  term and the  $\Sigma_+$  term. Summing over  $j$  the bounds on the Fourier coefficients of  $G_j^{(c,c')}$ , we find exponentially decaying bounds  $\mathcal{G}(k)$  such that the total contribution towards the operator norm of  $R$  coming from the entries with row mode index  $k$  and column mode index  $k'$  is bounded by  $(\mathcal{G}(|k' - k|) + \mathcal{G}(k' + k))/(kk')$ . We then sum the exponential decay bounds to get an estimate  $B$  such that  $\sum_{k=0}^{\infty} \mathcal{G}(k) \leq B$ . Now, for any  $k'$ , the sum over  $k$  of  $\mathcal{G}(|k' - k|) + \mathcal{G}(k' + k)$  is certainly bounded by  $3B$ . Moreover, each nonzero entry of  $R$  satisfies  $\max\{k, k'\} \geq \kappa + 1$ , and hence  $\|R\| \leq 3B/(\kappa + 1)$ . This bound is very weak but it does go to zero as the cutoff  $\kappa$  increases.

#### 4.7 Step 4: bounding the change in the Hessian

Our present task is (3b) from the previous subsection, bounding  $\|HS(x) - HS(x_0)\|$  for  $x \in B(x_0, \varepsilon) := \{x \in \mathcal{X} : \|x - x_0\| < \varepsilon\}$ . This is the most mathematically intricate step of the whole method, so we shall proceed through it with more detail than we used in the other steps.

Recall the definition (4.6) of the auxiliary functions  $G_j^{(c,c')}(t)$ . By (4.7) the entries of the Hessian are linear combinations, over  $j$ , of the Fourier coefficients of  $G_j^{(c,c')}(t)$ . We will bound the change coming from each  $j$  separately and sum the final operator norm bounds. We fix  $j \in \{1, \dots, n-1\}$  for the rest of the section. In addition to simplifying the discussion, this lets us reduce notational overload by dropping the subscript and writing  $G^{(c,c')} = G_j^{(c,c')}$ .

The functions  $G^{(c,c')}$  depend on the point  $x \in \mathcal{X}$ . Having just freed  $G$  of its subscript, we now write  $G_x^{(c,c')}$  to emphasize this dependence. Taking partial derivatives, for each  $m \geq 1$

and  $e = 1, \dots, d$  we have

$$\frac{\partial G_x^{(c,c')}(t)}{\partial(ma_m^{(e)})} + i \frac{\partial G_x^{(c,c')}(t)}{\partial(mb_m^{(e)})} = \frac{2(e^{2\pi imj/n} - 1)}{m} \cdot P_x^{(c,c',e)}(t) e^{2\pi imt}, \quad (4.9)$$

where the auxiliary functions  $P_x^{(c,c',e)}$  are defined by

$$P_x^{(c,c',e)}(t) = \frac{15\Delta x_{c'}\Delta x_c\Delta x_e}{|\Delta x|^7} - \delta_{ce} \cdot \frac{3\Delta x_{c'}}{|\Delta x|^5} - \delta_{c'e} \cdot \frac{3\Delta x_c}{|\Delta x|^5} - \delta_{cc'} \cdot \frac{3\Delta x_e}{|\Delta x|^5}$$

using the shorthand

$$\Delta x_c = x(t + j/n) - x(t).$$

In the language of §4.3,  $P_x^{(c,c',e)}(t)$  is an expressible function defined in terms of  $x$ . Using Lemma 2.5 we can bound the uniform norm of  $x - x_0$  for all  $x \in B(x_0, \varepsilon)$ . Then we can apply Bound 3 to get a bound  $N^{(c,c',e)}$  such that  $\int_0^1 (P_x^{(c,c',e)}(t))^2 dt \leq N^{(c,c',e)}$  for all  $x \in B(x_0, \varepsilon)$ .

*Aside.* In invoking Lemma 2.5, we were using the fact that the supremum norm of the function  $x(t)$  is continuous with respect to the norm on  $\mathcal{X}$ , i.e., elements of  $\mathcal{X}$  which are close with respect to the norm are also uniformly close pointwise. This is a crucial property for the norm to possess. If only, say, the  $L^2$  norm of  $x(t)$  were continuous, it would be impossible to compute such a bound  $N^{(c,c',e)}$ . Indeed, we would not even know if every  $x \in B(x_0, \varepsilon)$  is collision-free.

Taking the  $L^2$  norm of (4.9) and summing over  $m$ , we find the bound

$$\begin{aligned} \int_0^1 \left[ \sum_{m=1}^{\infty} \left( \left( \frac{\partial G_x^{(c,c')}(t)}{\partial(ma_m^{(e)})} \right)^2 + \left( \frac{\partial G_x^{(c,c')}(t)}{\partial(mb_m^{(e)})} \right)^2 \right) \right] dt &\leq \sum_{m=1}^{\infty} \frac{16}{m^2} \int_0^1 (P_x^{(c,c',e)}(t))^2 dt \\ &\leq \frac{8\pi^2}{3} N^{(c,c',e)}. \end{aligned}$$

For temporary use, define the  $\ell^2$  norm on  $\mathcal{X}$  by

$$\|x\|_2^2 = \sum_{k=1}^{\infty} \sum_{c=1}^d k^2 |A_k^{(c)}|^2.$$

Clearly  $\|x\|_2 \leq \|x\|$ . Now, by the mean value theorem and the Cauchy-Schwarz inequality, for any  $x \in B(x_0, \varepsilon)$  we have

$$(G_x^{(c,c')}(t) - G_{x_0}^{(c,c')}(t))^2 \leq \left[ \sum_{e=1}^d \sum_{m=1}^{\infty} \left( \left( \frac{\partial G_y^{(c,c')}(t)}{\partial(ma_m^{(e)})} \right)^2 + \left( \frac{\partial G_y^{(c,c')}(t)}{\partial(mb_m^{(e)})} \right)^2 \right) \right] \cdot \|x - x_0\|_2^2$$

for some  $y \in B(x_0, \varepsilon)$  (in fact, for some  $y$  on the line segment between  $x$  and  $x_0$ ). Integrating,

$$\int_0^1 (G_x^{(c,c')}(t) - G_{x_0}^{(c,c')}(t))^2 dt \leq \varepsilon^2 \cdot \sum_{e=1}^d \frac{8\pi^2}{3} N^{(c,c',e)}.$$

In summary, we have found an upper bound on the  $L^2$  norm of  $\Delta G^{(c,c')} := G_x^{(c,c')} - G_{x_0}^{(c,c')}$ .

We are nearly done. Recalling (4.7), the entries of the matrices  $HS(x)$  and  $HS(x_0)$  are linear combinations of the Fourier coefficients of  $G_x$  and  $G_{x_0}$ , respectively. By linearity, the entries  $HS(x) - HS(x_0)$  are the corresponding linear combinations of the Fourier coefficients of  $\Delta G^{(c,c')}$ . As in §4.6, let us conveniently forget that each pair of modes is associated to a  $2d \times 2d$  matrix, with the understanding that the bounds we write down should be summed over the  $2d$  rows and then a maximum should be taken over the  $2d$ ; so we drop coordinate labels and ignore the distinction between real and imaginary parts. The entry of the matrix  $HS(x) - HS(x_0)$  in row indexed by  $k$  and column indexed by  $k'$  is (a signed real or imaginary part of)

$$\frac{1}{kk'} \left( \widehat{\Delta G}(k' - k) \pm \widehat{\Delta G}(k' + k) \right).$$

As  $k$  varies, the mode in the first term ranges in absolute value from 0 to  $\infty$ . It covers the modes once if  $k' = 1$  and otherwise covers each term at most twice; thus, absorbing the  $1/k'$  factor, we may treat it as covering the range once. As  $k$  varies, the mode in the second term never repeats.

Putting this together, we see that the sum over  $k$  of the absolute values of the entries of  $HS(x) - HS(x_0)$  is bounded by two inner products of the vectors  $(\widehat{\Delta G}(k) : k \geq 0)$  and  $(1/k : k \geq 1)$ , possibly with reordering. By Cauchy-Schwarz, any such inner product is bounded by the square root of

$$\left( \sum_{k=1}^{\infty} \frac{1}{k^2} \right) \left( \sum_{k=0}^{\infty} |\widehat{\Delta G}(k)|^2 \right) \leq \frac{\pi^2}{6} \cdot \int_0^1 |\Delta G(t)|^2 dt \leq \frac{\pi^2}{6} \cdot \varepsilon^2 \cdot \sum_{\varepsilon=1}^d \frac{8\pi^2}{3} N^{(c,c',\varepsilon)}.$$

This gives the desired bound on  $\|HS(x) - HS(x_0)\|$ .

*Remark.* While this analysis was fairly loose, the resulting bound is  $O(\varepsilon)$  and so can be made as small as needed. The scaling with  $\varepsilon$  is what makes this step practical (cf. §5.3).

## 4.8 Final bounds

In §4.6 we computed bounds on  $\|T\|$  and  $\|HS(x_0) \circ T - \text{id}\|$ , and in §4.7 we computed a bound on  $\|HS(x) - HS(x_0)\|$  for  $x \in \mathcal{X}$  such that  $\|x - x_0\| < \varepsilon$ . Using

$$\|HS(x) \circ T - \text{id}\| \leq \|HS(x_0) \circ T - \text{id}\| + \|HS(x) - HS(x_0)\| \cdot \|T\|,$$

we obtain a bound on  $\|HS(x) \circ T - \text{id}\|$ . Combining this with the bound on  $\|\nabla S(x_0)\|$  from §4.4, we can attempt to check the hypothesis of Theorem I.2.1. More specifically, if  $\|\nabla S(x_0)\| \leq \alpha$ ,  $\|T\| \leq \beta$ , and  $\|HS(x) \circ T - \text{id}\| \leq \gamma$  and are the bounds we computed, then we test if  $\gamma < 1 - \alpha\beta/\varepsilon$ . If that hypothesis is satisfied, then (after invoking Corollary 4.5 and the comments of §4.5) Theorem I.2.1 rigorously proves the existence of an  $n$ -body choreography within  $\varepsilon$  of the starting path  $x_0$ . As a reminder, “within  $\varepsilon$ ” is measured by the norm on  $\mathcal{X}$ ; it implies closeness in the supremum norm by Lemma 2.5.

We close this section with the remark that all of the bounds we made here can be improved arbitrarily by using suitably large parameters. In particular, given any choreography at which the Hessian is invertible (after accounting for symmetries), there exist parameters for which the algorithm described here could prove existence. Of course, said parameters may not be practical; and indeed, without significant optimization (especially in §4.6), they are not.

## 5 Implementation Details

We implemented the proof technique just outlined in a C++ program named `PROVER`. However, in order to handle interesting cases, several changes and optimizations were necessary. Some of the most important (practically) and interesting (theoretically) are discussed in this section.

### 5.1 Optimizations in Hessian computations

The Hessian bounds in §4.6 are the most costly computations by a wide margin. In `PROVER` we implemented several improvements, both theoretically (so as to obtain a proof with more modest parameters) and computationally (so as to speed up the calculations we do make).

The first and most important change is that, as in the previous chapter (see §II.6.1), we perform matrix inversion in native double-precision arithmetic. That is, instead of inverting the Hessian exactly using interval arithmetic, we cast each entry down to (machine-native) double-precision, invoke a floating-point library to compute the inverse, and then cast the results back up to intervals (of zero width). We use this as the matrix  $T$ . We have to account for the fact that  $T$  is not the exact inverse with another error term (see §5.4), but in practice the precision of the numerically-computed inverse is very good, so the extra error is negligible. In Chapter II we used for  $T$  the least-squares pseudo-inverse. Here we are inverting symmetric square matrices, so we instead used a Cholesky  $LDL^t$  decomposition (with pivoting). Specifically, we used the `DSPTRF` and `DSPTRI` functions in `LAPACK` [3]. They save memory by operating in-place on packed representations of symmetric matrices. That turns out to be advantageous, as for complicated choreographies RAM usage has been a limiting factor in the proof computations.

In the analysis given in §4.6 we noted that by increasing  $\kappa$ , i.e., by looking at a larger submatrix  $M$ , we could improve the quality of our bounds. This is a crude solution, though, because the runtime of the matrix inversion step is cubic in  $\kappa$ . This makes it impractical to handle complicated choreographies by simply increasing  $\kappa$ . Improvements in the analysis are needed.

In the exposition we gave in §4.6, we computed the explicit upper-left  $(2d\kappa) \times (2d\kappa)$  submatrix  $M$  and let  $R$  be the “remainder” matrix obtained by zeroing out those entries. We then bounded  $\|HS(x_0) \circ T - \text{id}\|$  by  $\|R\| \cdot \|T\|$ . In practice the norm of  $T$  can be somewhat large; in some cases it is on the order of  $10^3$ . This makes the bound  $\|R\| \cdot \|T\|$  very crude.

To improve it, we first note that, after the first  $2d\kappa$  rows/columns,  $T$  is the identity matrix. Thus  $\|RT\|$  is bounded by  $\max\{\|R_{\text{left}}M^{-1}\|, \|R_{\text{right}}\|\}$ , where  $R_{\text{left}}$  is the matrix containing the first  $2d\kappa$  columns of  $R$  and  $R_{\text{right}}$  is the matrix containing the remaining columns. This observation means that, while we need to be careful about bounding the first  $2d\kappa$  columns of  $R$ , we can be much more loose in bounding the remaining columns.

*Aside.* Accepting a worst-case cost of a factor of 2, we actually compute the sum  $\|R_{\text{left}}M^{-1}\| + \|R_{\text{right}}\|$  instead of the maximum. The one exception is that when we have some error term that applies equally to both the left and right parts of  $R$ , we only consider it on the left. This is because of the above observation that, since  $R_{\text{left}}$  is amplified by  $M^{-1}$  in our bounds, its norm counts more than that of  $R_{\text{right}}$ . Technically this is only correct if  $\|M^{-1}\| \geq 1$ . We do enforce this in our code, but it could only be violated in the most contrived circumstances. In particular,  $M$  acts on the large Fourier modes like the identity, so the norm of both  $M$  and  $M^{-1}$  are at least on the order of 1.

Now we turn our attention to improving the bounds on  $R$ . It is especially important to bound  $R_{\text{left}}$  well, as its effect in the final operator norm bound is magnified by  $M^{-1}$ . We do this quite simply by computing some of its nonzero terms explicitly. In the first  $2d\kappa$  columns, we compute a user-selectable number of rows of  $R$  below the first  $2d\kappa$  (say, those in which the mode index is between  $\kappa + 1$  and  $\kappa + r$ ). That is, we compute an explicit submatrix below the entries corresponding to  $M$ . Let  $R_{\text{left},1}$  be this submatrix and let  $R_{\text{left},2}$  be the (infinitely tall) submatrix containing the remaining rows. We compute  $R_{\text{left},1}M^{-1}$  explicitly and take its operator norm; we then bound  $\|R_{\text{left},2}\|$  using the basic approach in §4.6. This bound is reasonable because  $k - k'$  (the row mode index minus the column mode index) in  $R_{\text{left},2}$  is at least  $r$ , so we only need to add up the tail of the Fourier coefficients. We then combine these two bounds by noting that  $\|R_{\text{left}}M^{-1}\| \leq \|R_{\text{left},1}M^{-1}\| + \|R_{\text{left},2}\| \cdot \|M^{-1}\|$ .

It remains to bound  $\|R_{\text{right}}\|$ . We start by computing a user-selectable number of columns of it, say up to mode  $c$ . We evaluate enough entries in each column that all of the remaining entries have  $k - k'$  greater than a user-determined cutoff, say  $m$ . After taking the operator norm of the explicitly-computed part, we need to bound the operator norm coming from the columns with mode greater than  $c$  and the norm from the remaining unevaluated terms in the columns we did compute. The latter is easily bounded as in the previous paragraph, using the fact that sum of the Fourier coefficients with mode greater than  $m$  (i.e., the sum of the tail) is small. For the uncomputed columns, we consider separately the terms with  $|k - k'| \leq m$  and those with  $|k - k'| > m$ . The terms in which  $|k - k'|$  is large are bounded as above. For the terms in which  $|k - k'|$  is small, we are forced to use the sum of the corresponding Fourier coefficients; this sum is typically not negligible. However, instead of only being able to divide by  $\max\{k, k'\} \geq \kappa + 1$  in computing the operator norm bound, we can now divide by  $\lambda := (c + 1)(c + 1 - m)$ . This is a big improvement for two reasons:  $\lambda$  depends on  $c$  instead of  $\kappa$ , and the runtime is quadratic in  $c$  as opposed to cubic in  $\kappa$ ; and, if we view  $m$  as fixed, then  $\lambda$  grows quadratically instead of linearly.

These improvements can be summarized by the following four parameters, which are tweakable in `PROVER`:

- (a) `maxmode`, i.e.,  $\kappa$ ;
- (b) `cutoff`, i.e.,  $m$ ;
- (c) `columns`, i.e.,  $c$ ; and
- (d) `extraheight`, i.e.,  $r$ .

## 5.2 Dropping variables

In §4.5 we discussed the approach of accounting for symmetries by dropping variables. There are two implementation issues associated with this: the first is determining a set of variables to drop and verifying the legitimacy of the choice, and the second is actually dropping the variables from our computations.

The second problem is neatly resolved by simply zeroing out the corresponding rows and columns of the explicit matrices we compute ( $M$  and  $M^{-1}$ , in the notation of §4.6). More precisely, in  $M$  we zero out the rows and columns and then set the diagonal entries to 1; no more care that that needs to be taken in the implementation. This is because the  $\ell^1$  operator norm of a matrix can only decrease when rows and/or columns are deleted. Thus the bounds of, e.g., §4.7, still apply to the post-variable-dropping matrices. Moreover, by taking this action in  $M$ , the matrix  $T$  we use will be zero in the rows and columns that

were supposed to be erased (except for the ones on the diagonal). Thus  $HS(x_0) \circ T$  has an extra column for each variable that was supposed to be deleted, but the other columns are unaffected. This extra column can only increase the operator norm (and in practice it does not increase it at all).

We could also have dropped the variables in our calculation of the gradient norm, but we did not bother as it would have given negligible savings.

*Aside.* Above we asserted that the  $\ell^1$  operator norm of a matrix can only decrease when rows and columns are deleted. This is obvious for the  $\ell^1$  and  $\ell^\infty$  norms, as these norms are the suprema of the  $\ell^1$  norms of the columns and rows, respectively. It is also clear for the  $\ell^2$  norm. However, it is not true in general. Endow  $\mathbb{R}^2$  with the norm  $|(x, y)| = \max\{|x + y|, |x|\}$ , take the map  $M(x, y) = (x + y, -2(x + y))$ , and consider “dropping” (i.e., projecting out) the first variable. The operator norm of  $M$  is 1, but  $M(0, 1) = (1, -2)$ , projecting out the first variable gives  $(0, -2)$ , and  $|(0, -2)| = 2 > 1 = |(0, 1)|$ .

Consider now the first problem, that of determining a set of variables to drop and verifying its legitimacy. We do this as follows. Let  $L_0$  be the matrix whose columns are a basis for  $\text{Lie}(\text{O}(2) \times \text{O}(d)) \cdot x_0$ . There are only finitely many nonzero modes in  $x_0$ , as it is given by trigonometric polynomials, so  $L_0$  is a finite matrix. We need to identify a square submatrix of  $L_0$  which is invertible. For this we use a LU decomposition with partial (row) pivoting. This is a standard algorithm in numerical analysis [79, §2.3]. We iteratively do the following procedure for the  $c^{\text{th}}$  column: identify the row with the largest entry (in magnitude), permute it to the  $c^{\text{th}}$  row, and then subtract an appropriate multiple of this column from the subsequent columns so as to make their entries in the  $c^{\text{th}}$  row all equal to zero. The exact numerical behavior of this algorithm is not critical to us; the important thing is that at the end, the top submatrix is invertible. The rows which we permuted into the top in the course of the algorithm are the rows we select for dropping.

It remains to verify that this choice is valid for every nearby point  $x$  (in particular, for the choreography we intend to find). For this we need to pay some attention to the definition of  $L_0$ . For its columns, we take a natural basis in which the first column corresponds to time translation and the remaining columns correspond to spatial rotation in coordinate planes. In particular, each column consists of a permutation, in each Fourier mode, of the  $2d$  coordinates of  $x_0$  (multiplied by the Fourier mode, in the case of time translation). Thus, if  $L_*$  is the corresponding matrix defined for another point  $x_* \in \mathcal{X}$ , then the  $\ell^1$  norm of each column of  $L_* - L_0$  is either bounded by (for the spatial rotations) or equal to (for time translation) the norm  $\|x_* - x_0\|$ . In particular, if  $\|x_* - x_0\| < \epsilon$ , then the operator norm of  $L_* - L_0$  is also bounded by  $\epsilon$ . This is of course also true if we limit our attention to the square submatrix we previously selected. Now we just use the following observation: if  $A, B, C$  are square matrices with  $\|AC - \text{id}\| < \delta$ ,  $\|A - B\| < \epsilon$ , and  $\delta + \epsilon \cdot \|C\| < 1$ , then  $B$  is invertible. (In particular,  $B^{-1} = C(\text{id} - (\text{id} - (AC - (A - B)C)))^{-1}$ , and the last term is invertible because  $\|\text{id} - AC - (A - B)C\| < \delta + \epsilon \cdot \|C\|$ .) Thus, to complete the argument, we just numerically invert the square submatrix of  $L_0$ , compute the norm of the approximate inverse, and compute how far the product of the matrix and its approximate inverse is from the identity.

*Aside.* While simple and arguably elegant, this approach is not very aesthetically pleasing. In particular, there is not a canonical choice, and the choice of variables we make will vary by choreography. In a previous version of our software, we used an approach which is more principled, but which unfortunately is limited to  $d = 2$ . Identify  $\mathbb{R}^2$  with  $\mathbb{C}$  and consider the Fourier series expansion of  $x^{(1)}(t) + ix^{(2)}(t)$ . Pick a particular mode  $k$

and consider the  $+k$  and  $-k$  Fourier coefficients,  $A_k$  and  $A_{-k}$ . The combined action of spatial rotation by  $\theta_1$  radians and time translation by  $\theta_2/(2\pi k)$  acts on these coefficients by  $(A_k, A_{-k}) \mapsto (e^{i(\theta_1 - \theta_2)} A_k, e^{i(\theta_1 + \theta_2)} A_{-k})$ . As long as  $A_k$  and  $A_{-k}$  are both nonzero, there is a canonical representative for the orbit of  $\text{SO}(2) \times \text{SO}(2)$  given by taking both  $A_k$  and  $A_{-k}$  to be positive reals. Thus we can account for symmetries by fixing the imaginary parts to zero (and dropping the corresponding variables).

### 5.3 Computing bounds on Fourier coefficients

We described the approach for bounding Fourier coefficients in §4.3; in this section we just add a few additional remarks to the discussion.

Most of our bounds on Fourier coefficients are in terms of Bound 4, which uses analytic continuation into the complex plane. These computations have two parameters: the height  $h$  that we lift to, and the length of the Fourier transform. These are both adjustable in PROVER.

While we leave them as user-selectable, it would have been possible to choose both of these parameters in a more systematic manner. In particular, we could choose a lift height  $h$  by repeatedly computing the largest height for which Lemma 4.9 is able to prove analyticity, lifting to that height, and recomputing. We implemented this in a previous version of the software, but found it to have an unfavorable tradeoff between usage value and code clarity. In particular, it is not always better to increase the height; as we see from Lemma 4.10, a larger height proves a better rate of exponential decay, but it may also increase the coefficient. Thus it is not clear that maximizing  $h$  is desirable.

From a theoretical perspective, though, increasing the length  $N$  of the Fourier transform is always advantageous. Moreover, the runtime is proportional to  $N \log N$ , so it is relatively cheap to increase this parameter.

Increasing  $N$  allows one to offset a limited lift height. This is important not just because we are not using an optimal choice of  $h$ , but rather because some choreographies simply do not admit analytic continuation to a large strip. Increasing  $N$  helps in the following ways. The error on individual Fourier coefficients comes from bounding the sum of Fourier coefficients of modes differing by a multiple of  $N$ ; if we increase  $N$  then we increase how far out the modes we need to bound are, and thus we can tolerate a bound on those modes that decays more slowly. Also, when we compute the “tail bounds” on sums of Fourier coefficients, we explicitly sum up all of the terms we have available from the Fourier transform, and use the exponentially decaying bound on the remainder. If we compute more terms, then the smallest term to which we apply the exponential decay bound is pushed farther out. Again, this lets us tolerate a slower decay.

On a broader note, recall that we commented in §4.3 that it is possible to determine bounds using only Parseval-type bounds applied to time derivatives. In particular and for contrast with the approach we chose, consider that in §4.6 we could have bounded the large-mode coefficients in a uniform way that applies to the whole ball  $B(x_0, \varepsilon)$ . This is a pleasant idea in principle, and it would essentially avoid the need for the arguments of §4.7. However, in practice the bounds that one gets from such an approach are insufficient to handle even moderately complicated choreographies. To make our software widely applicable, we instead adopted the roundabout approach of (1) exploiting analytic continuation — a special property possessed by  $x_0$  but not a general element of  $B(x_0, \varepsilon)$  — to produce much stronger bounds on the large-mode coefficients, and then (2) using the closeness of  $x$  to  $x_0$  to deduce that  $HS(x)$  cannot have changed much from  $HS(x_0)$ .

## 5.4 Faster matrix multiplication

To compute the matrix  $T$ , we need to perform a matrix inversion. As discussed above, we do this with an optimized library in machine-native arithmetic; the inversion step becomes practical by dint of this optimization. Once we have computed the inverse, we need to multiply it by the original matrix (to compute the error introduced by our approximate inverse). The results of this multiplication need to be rigorous, so one is inclined to use interval arithmetic. However, somewhat surprisingly, merely multiplying the two matrices turns out to be unpleasantly slow.

There has been a lot of work on efficient implementations of matrix multiplication, and in particular there are optimized libraries for multiplication in machine arithmetic (see, e.g., the ATLAS project [96]). To make the most of what we have on hand [4], we decided to use double-precision arithmetic to compute the matrix product and use floating-point error analysis to rigorously bound the error in this computation. This error analysis is standard; here we repeat just enough of the theory for our needs.

The IEEE 754 floating-point standard [52], which is obeyed by nearly every modern processor (including Intel processors since 1987 [51]), specifies that the results of floating-point operations should be exactly rounded. That is, the result should be the same as if the computation had been performed with infinite precision and then rounded (in whatever rounding direction is chosen — most commonly, rounded-to-nearest). The double-precision floating-point type we use has 52 bits of precision (actually 53, including an implied leading 1 in the mantissa). Thus, given any arithmetic operation  $x * y$ , we have

$$|(x * y)_{\text{computed}} - (x * y)_{\text{exact}}| \leq 2^{-52} \cdot (x * y)_{\text{exact}}. \quad (5.1)$$

*Aside.* While there is some nonuniformity in the implementation of trigonometric and other special functions, our analysis only depends on the implementation of addition and multiplication. Thus it is valid even on the infamous early P5 Pentium processors [26].

In particular, we have  $|(x * y)_{\text{computed}}| \leq (1 + 2^{-52}) \cdot |(x * y)_{\text{exact}}|$ . Suppose we are multiplying  $m \times k$  and  $k \times n$  matrices  $A$  and  $B$ , respectively. If  $A_{\max}$  and  $B_{\max}$  are supremum bounds on the entries of  $A$  and  $B$ , respectively, then the computed product of any two terms has magnitude at most  $(1 + 2^{-52})A_{\max}B_{\max}$  and differs from the exact value by at most  $2^{-52}A_{\max}B_{\max}$ . The computed sum of two such terms has magnitude at most  $2(1 + 2^{-52})^2A_{\max}B_{\max}$  and the sum introduces an additional error of at most  $2^{-52} \cdot 2(1 + 2^{-52})A_{\max}B_{\max}$ . Continuing this analysis to a sum of  $k$  such terms, we get a bound on the final error of each entry of the computed matrix product.

This simple analysis does not depend on the order in which terms are added, but it does depend on the following assumption:

$$\begin{aligned} &\text{each term of } A \cdot B, \text{ say } (A \cdot B)_{ij}, \text{ is computed by evaluating the } k \text{ products} \\ &\quad \{A_{ih}B_{hj} : 1 \leq h \leq k\} \\ &\quad \text{and summing them, in some order.} \end{aligned} \quad (5.2)$$

In particular, we assume that no asymptotically fast matrix multiplication techniques, like Strassen multiplication [89], are used.

Our code uses the DGEMM function in BLAS [31] and gives error bounds that are correct whenever (5.1) and (5.2) are satisfied. The specification of the BLAS library does not guarantee that these, and in particular (5.2), are obeyed, but in practice they are.



To accommodate hardware and/or software in which (5.1) and/or (5.2) (respectively) are not satisfied, we also provide a fixed-point implementation of matrix multiplication. This rounds each term to a (representable) integer multiple of a fixed power of 2 and then computes the product using integer arithmetic. It gives simple error bounds in terms of the original rounding operations. The code is somewhat complicated by computing an exponent that ensures no overflows in the integer arithmetic operations. Since assumptions (5.1) and (5.2) are obeyed on most systems, we expect that the fixed-point code will not be needed, and thus we shall not go into any more detail here.

To motivate this discussion, we tested computing  $\|AB - \text{id}\|$ , where  $A$  and  $B$  are  $1000 \times 1000$  matrices, on our workstation. The timings were as follow: 72 seconds with interval arithmetic multiplication, 1.36 seconds with fixed-point multiplication, and 0.81 seconds with floating-point multiplication.

Finally, while in this document we have not heretofore discussed parallelizability, we do note that the matrix multiplication step is embarrassingly parallel. We only care about the operator norm of the result, so we could compute each column of the product separately, take their  $\ell^1$  norms, and then take the maximum of the results.

## 6 Comparison With Previous Work

The technique of computer-assisted proof has previously been applied to a variety of problems, including  $n$ -body choreographies. We are aware of two previous such bodies of work: that of Kapela et al. [58, 55, 56] and that of Arioli et al. [7]. In this section we give some remarks concerning the similarities and differences with the work presented here.

### 6.1 Kapela et al.'s computer-assisted proofs

In three successive papers, Kapela and Piotr Zgliczyński [58], Kapela [55], and then Kapela and Simó [56] gave computer-assisted proofs for the existence of a few choreographies. The fundamental difference in the approaches taken in these papers lies in the treatment of symmetries (see §4.5); we discuss these differences in §6.3.

The approach taken by Kapela et al. differs significantly from ours. While we directly handle the infinite-dimensional problem of action criticality, they instead considered the finite-dimensional dynamics problem. One can define the *time-evolution operator* (an operator on phase space) which takes initial conditions in the form of positions and velocities for each body and returns the same data at a later point in time. Using an algorithm for rigorously computing the Jacobian of this operator (specifically, the  $C^1$ -Lohner algorithm [100]), one can apply effective existence theorems like our Theorem I.2.1. They phrased their existence result in terms of the Krawczyk method, which *a priori* applies only in finite dimensions (but that, of course, is sufficient for their needs).

Their work utilized a general-purpose library, CAPD, which performs rigorous calculations for ordinary differential equations, including evaluation of the time-evolution map and its derivatives [27]. This library has been applied to a range of problems [39, 38, 97, 61].

Assuming an algorithm for computing the Jacobian of the time-evolution operator, this approach is conceptually simpler than ours. It is also easy to adapt to any dynamical system. However, our method is in principle applicable to a wider range of problems other than dynamical systems, such as two-dimensional partial differential equations. These problems cannot, in general, be recast as finite-dimensional; an approach that copes directly with infinite-dimensional spaces is needed.

The approach of Kapela et al. also suffers from the instability of most choreographies; the time-evolution operator is ill-behaved and is quite sensitive to initial conditions. This necessitates the use of extra techniques like parallel shooting to handle nontrivial cases [56, §4.1]. Even with such techniques, they were not able to handle choreographies of the complexity that we treat here (see Figure 8.4). Instability of the dynamical system is not, *a priori*, an issue for our action-based approach.

Our approach depends on the nonsingularity of the Hessian of the action (after accounting for symmetries); their approach depends on the nonsingularity of the time-evolution operator minus the identity (after accounting for invariants). We do not know of any theoretical reason ruling out the possibility that, for some choreography, one of these nonsingularity conditions could be satisfied but not the other. However, we have never seen such a case.

Their dynamical approach also allows one to get a handle on the eigenvalues of the monodromy matrix. When these eigenvalues lie on the unit circle, the orbit is linearly stable. Kapela and Simó used this approach to prove the stability of the figure-eight [56]. They then went on to prove KAM stability of the figure-eight as well as some members of a family of relative choreographies [57]. We have not investigated techniques for proving stability, but propose to pursue them in the future (see §7).

## 6.2 Arioli et al.’s computer-assisted proofs

While early work of Gianni Arioli [5, 6] studied periodic orbits of the  $n$ -body problem using similar finite-dimensional techniques, a later paper with Barutello and Terracini applied a different approach to existence proofs for choreographies [7]. Their approach is much closer to ours, in that it directly considers the infinite-dimensional problem of action criticality. We became aware of this work after developing our own methods.

Like our work, theirs relies on an existence theorem in Banach spaces. They explicitly phrased their existence theorem as a fixed point result and cited it as a consequence of the contraction mapping principle [7, Lemma 2]; while their result is superficially different from ours, it is similar in substance.

Unlike us, Arioli et al. consider a Banach space of functions with analytic continuation to a fixed strip. Recall from §4.3 that we did exploit analytic continuation to improve our bounds; working exclusively with such functions is arguably a cleaner approach. However, because analytic continuation was not technically required for the theoretical analysis and we did not impose it *a priori* on our solution, our approach is in principle applicable to non-smooth problems with, say,  $C^1$  functions rather than holomorphic functions.

Another substantial difference is that we explicitly handle non-polynomial functions, whereas Arioli et al. approximate every function that arises by a polynomial. They do this because they are building on a body of work [9, 8] that only handles polynomials. This is also how they are able to work in a space of analytic functions. To account for the non-polynomial nature of the actual functions, they have to compute error bounds for the quality of the approximation. These error bounds depend on the distances, i.e., the arguments of the approximated functions, remaining in a certain range. This is relegated to a two-sentence remark [7, p. 459], so we are not sure of the precise implementation.

In summary, we understand there to be two major high-level differences between the approach of Arioli et al. and the approach presented in this document. The first is concreteness: they streamlined their analysis by using the language of compact operators on Banach algebras, whereas we have tried to keep our presentation as explicit as possible. The second difference is reliance on computation: by restricting to polynomials, they could use

intricate manual bounds to reduce the need for computer assistance. By contrast, we have not hesitated to offload more of the work onto the computer.

There is also a substantial difference in the implementation of our methods. Arioli et al. do use interval arithmetic, but their base floating-point type is the native double, whereas we use arbitrary-precision floating-point types. This costs us a performance penalty, but it allows us to handle more delicate cases. The difference is reflected in the number of Fourier coefficients needed for the starting choreography approximation: the cases they studied are well-approximated by 60 coefficients, whereas the choreographies we handled routinely require approximations with a number of coefficients on the order of  $10^3$ .

More than just proving existence of individual choreographies, though, Arioli et al. also proved the existence of a continuous family of relative choreographies. We have not extended our method to handle families.

### 6.3 Treatment of symmetries

Recall from §4.5 that, in order to apply an existence theorem tailor-made for *locally unique* solutions, we need to handle the problem of symmetries. This problem also arises in both of the aforementioned computer-assisted proof techniques, but the method of resolving it varies.

In the first of the three papers (on existence of choreographies) by Kapela et al. [58], Kapela and Zgliczyński exploited the symmetries of the figure-eight to reduce to a 2-dimensional phase space. This reduction in particular breaks the  $O(2) \times O(2)$  symmetry group of the planar problem. Thus we expect a locally unique solution in the reduced space, and indeed, Kapela and Zgliczyński successfully applied their existence theorem to this 2-dimensional problem [58].

The reduction afforded by the figure-eight symmetries is overkill for breaking the  $O(2) \times O(d)$  symmetry group. In the planar case, all that is needed is to fix a preferred time phase and rotation of the spatial frame. Choreographies with bilateral symmetry, in which the orbit is invariant under the joint operation of time reflection and reflection through one axis, admit a natural choice therefor. This is the approach taken in the second paper of Kapela [55]. It is also the approach taken by Arioli et al. [7, p. 455].

In order to handle asymmetric choreographies, a different approach is needed. In the third Kapela et al. paper, Kapela and Simó developed such an approach [56]. It is the analog in their finite-dimensional setting of the dropping-variables technique we use. Using the fact that the time-evolution operator preserves the invariants of the physical problem — angular momentum and energy — they verified that it is unnecessary to find a fixed point of the full time-evolution operator, but rather two variables can be omitted.

Comparing this approach with ours exemplifies the Noether’s-theorem duality between symmetries and invariants. We propose that this particular problem is more natural in our setting (that of symmetries) rather than Kapela and Simó’s setting (that of invariants), because computations with energy and momentum are more complicated than computations with symmetries. In particular, checking the conditions of Lemma 4.2 is simple, whereas checking local injectivity of energy and momentum is marginally more delicate.

## 7 Further Work

This work has raised many questions which we think are worthy of study. In this section we give brief mention to a few such questions.

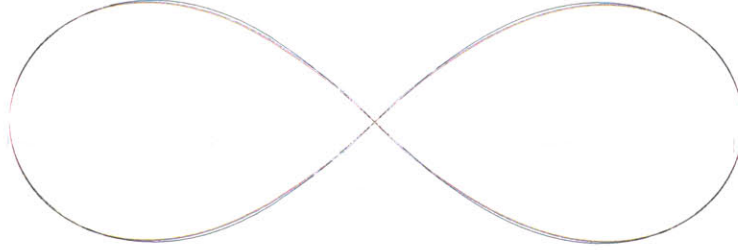


Figure 7.1: “Generalized figure-eight”  $n$ -body choreographies found by Simó [81], rescaled to have the same horizontal range. Depicted are choreographies for  $n = 3$  (black),  $n = 19$  (blue),  $n = 99$  (yellow), and  $n = 199$  (red). Notice that, once rescaled, they are nearly indistinguishable.

The most obvious direction for further work is in extending our results to spatial choreographies. From the perspective of our proof technique, there is no trouble; in fact, PROVER already handles choreographies of arbitrary dimension. The complication arises in finding choreographies; our preliminary results suggest that, without imposing symmetries or other conditions, it is relatively difficult to find non-planar choreographies. We will report on this in greater detail elsewhere.

Another natural generalization is changing the physical model. For instance, in this document we only considered Newtonian gravity, i.e., the inverse-square force law. One could consider relativistic corrections, either to first-order post-Newtonian effects or to the full general relativistic setting. In the setting of general relativity, it is not even clear at the outset whether or not choreographies are possible. There has been some work on this problem [50, 53, 99], but we think it would be interesting to develop this further.

As introduced in §3.6, there is some evidence that the pentagram orbit in Figure 3.8 is (linearly) stable. This begs attention, in a few ways. Firstly, it suggests that there would be value in simulating the trajectory to greater precision and for longer times in order to gain better evidence for its stability. If it does appear stable, then we could attempt to apply computer-assisted proof techniques to prove stability.

**Problem 7.1.** *Prove (or refute) the claim that the pentagram orbit in Figure 3.8 is stable.*

So far we have not developed any techniques for approaching Problem 7.1, but doing so seems worthwhile. Moreover, given simulation and proof technology, it would be interesting to look for other examples of stable choreographies. A good starting point would be to compute high-precision simulations of all of the choreographies we have found. The simple computations we have done so far suggest that there are no other stable examples in our data set; however, it could be the case that there are examples, but they just have small regions of stability.

Another possibility for further work is suggested by the plot in Figure 7.1. This graphic shows four figure-eight-like choreographies, presented on the same plot, and each rescaled to have the same width. The rescaling is logarithmic in the number of bodies.

This figure suggests quite convincingly that the figure-eight orbits are converging, as  $n \rightarrow \infty$ , to some limiting curve.

**Problem 7.2.** *Determine the limiting curve of the family of  $n$ -body (with  $n$  odd) figure-eight orbits, or determine that the curves do not converge.*

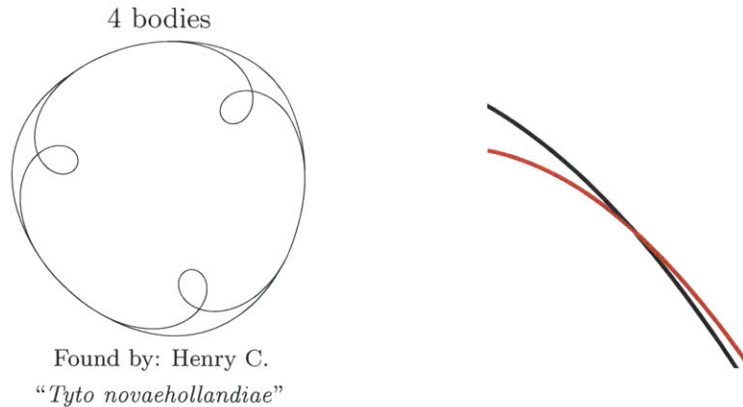


Figure 7.2: A choreography which we have proven to exist, and a magnification of part of its graph. The choreography appears to have a tangency, but actually the curve crosses itself.

The energy must be renormalized in order to define the  $n \rightarrow \infty$  continuum limit of the physical problem, but this renormalization is well-understood [36]. It is appealing to conjecture that the limiting curve has a simple description, e.g., algebraic. The possible algebraic nature of the 3-body figure-eight itself has been investigated, but while it is well-approximated by algebraic curves, it does not itself appear to be algebraic [85, p. 211]. Interestingly, the lemniscate is a valid shape for a choreography if one changes the potential function [37].

Another avenue for research involves relaxing the definition of a choreography. We built several conditions into Definition 1.1 that one might consider removing, such as the equal-mass hypothesis and the equal time spacing hypothesis. Define a *generalized choreography* to be a periodic solution of the  $n$ -body in which the  $n$  bodies follow the same curve.

**Problem 7.3.** *Determine if there exist generalized choreographies which are not choreographies, i.e., generalized choreographies with unequal masses or without equal time spacing.*

A generalized choreography with unequal masses (but equal time spacing) is called a *perverse choreography* by Chenciner; he has shown that there exist no perverse  $n$ -body choreographies with  $n \leq 5$  [20]. As far as we know, the existence of  $n$ -body perverse choreographies with  $n > 5$  and the existence of generalized choreographies with unequal time spacing are both still open questions.

Our next proposal for further work is motivated by Figure 7.2. This shows a particular 4-body choreography and a magnification of part of its graph. The choreography appears to have three points of tangency; however, the magnified plot shows that actually the curve crosses itself. Another type of "near miss" is depicted in Figure 7.3, which shows magnifications of two other choreographies. Both of them appear to have cusps in their graph, but after magnification we see that neither apparent corner of the graph is cuspidal.

Note that, while choreographies are smooth functions of time, it is still possible for the graph of a choreography to have a cusp. This could occur if some body at some point in time is at rest (equivalently, by the choreographic constraint, each individual body is at rest at some point in time).

These observations raise the question of whether there exist choreographies with tangencies and/or cusps, and the related question of whether there exist choreographies in which some body is at rest at some time. There exist periodic orbits of the  $n$ -body problem, and

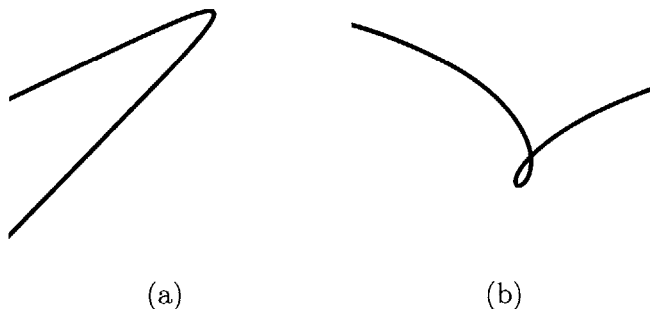


Figure 7.3: Magnifications of near-cusps for two choreographies: (a) is the pentagram in Figure 3.8 and (b) is “deformed heart” in Figure 8.7.

even the 3-body problem, which start with all bodies at rest [86]. A simple argument shows that a choreography cannot start with all bodies at rest, but we know of no argument ruling out the possibility of having an individual body at rest.

We think these questions are interesting on their own, but they also hint at a more fundamental question. A possible explanation for the lack of cusps is that there is no particular reason (as far as we know) for a choreography to prefer a cusp to a near-cusp. That is to say, if we considered a continuous family of choreographies containing one with a near-cusp, say by looking at relative choreographies or by varying the potential function, then the family may well contain one with an actual cusp — but we know of no reason why that representative should correspond to an absolute choreography or should correspond to exactly the inverse-square force law. If, on the other hand, we had such a family within the domain of absolute choreographies with respect to the Newtonian potential, then the same argument suggests that we could find a cuspidal choreography.

**Problem 7.4.** *Find (or rule out the possibility of) a continuous family of  $n$ -body absolute choreographies for the Newtonian potential.*

Members of such a family would necessarily fail our method of computer-assisted proof, and indeed they would fail the other two established methods as well (see §6).

We close this section with a more concrete proposal for further work. We alluded in §3.5 to a possible method for proving that a choreography has a given symmetry group, relying on the uniqueness guarantee of Proposition I.2.4. To elaborate on this idea, notice that a successful application of our computer-assisted proof technique yields more than just existence: it gives existence of a choreography close to our approximation, and it also shows that, in a ball of effectively computable radius, any two choreographies must be related by isometry (in fact, an isometry which is close to the identity). The existence part comes directly from Theorem I.2.1, while the uniqueness part combines Proposition I.2.4 with our approach in §4.5 for handling the symmetries of the problem.

Suppose now that we have a choreography  $x$  which, to reasonably high accuracy, admits a certain finite symmetry group  $G$ . That is, for all  $g \in G$ ,  $g \cdot x \approx x$ . Given suitable bounds, the uniqueness statement could guarantee that  $g \cdot x = \sigma_g \cdot x$ , where  $\sigma_g$  is an isometry that is close to the identity. Assuming  $x$  has no continuous symmetries (i.e., that it is not the circular orbit), this isometry  $\sigma_g$  is unique, so that the transformation  $\sigma_g^{-1}g$  is the unique stabilizer of  $x$  in a small neighborhood of  $g$ . Again assuming suitable bounds, we could deduce from this uniqueness statement that  $(\sigma_g^{-1}g)(\sigma_{g'}^{-1}g') = \sigma_{gg'}^{-1}(gg')$  for all  $g, g' \in G$ . In

other words,  $\{\sigma_g^{-1}g : g \in G\}$  is a finite group of symmetries of  $x$  (and this group is isomorphic to  $G$ ). This discussion is the motivation for the following loosely-stated principle, which for a suitable definition of “appears” can be applied to any choreography whose existence can be proven using our techniques.

**WYSIWYG Principle.** *If a choreography appears to have a certain symmetry group, then it actually does have that symmetry group (up to a small perturbation).*

The name of this principle comes from the computing notion of “what you see is what you get.” Because of the WYSIWYG Principle, symmetry and cusps are qualitatively different; whereas choreographies do not prefer cusps to near-cusps (repeating our language from above), they do prefer symmetries to near-symmetries.

**Problem 7.5.** *Extend PROVER to implement the WYSIWYG Principle, i.e., to rigorously determine the symmetry group of a given choreography.*

The above discussion can be made fully explicit, so there should be no great obstruction in completing Problem 7.5; we just need to make it so [88].

## 8 Gravitational Gallery

Figures 8.1–8.3, together with 3.5 and 7.1, show depictions of the 45  $n$ -body choreographies found by Simó, using data from his website [81]. (There are two more orbits in his data set — circular orbits with 3 and 11 bodies — but we have suppressed them here.)

Figure 8.4 presents the choreographies whose existence was established by the computer-assisted proof techniques developed in the three papers by Kapela et al. [58, 55, 56]. We have proven existence for all of these with our own software as well, and the graphs we show come from that data.

The remaining figures show some more choreographies handled by PROVER, i.e., choreographies for which our methods were able to prove existence.

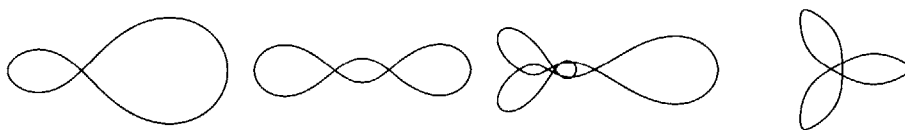


Figure 8.1: 4-body choreographies found by Simó [81].

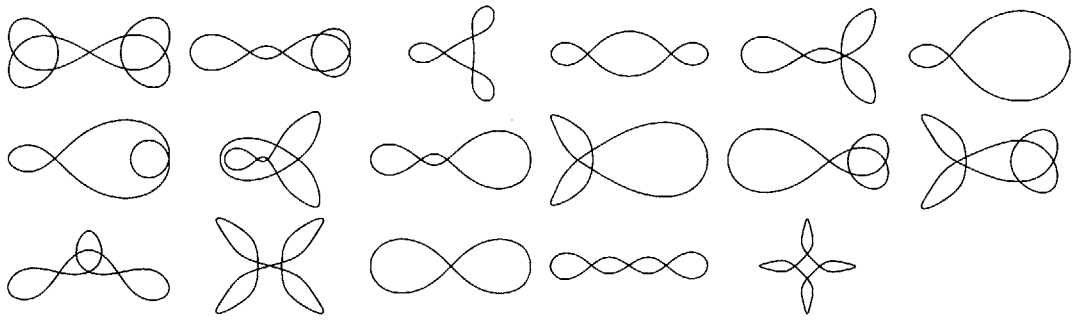


Figure 8.2: 5-body choreographies found by Simó [81].

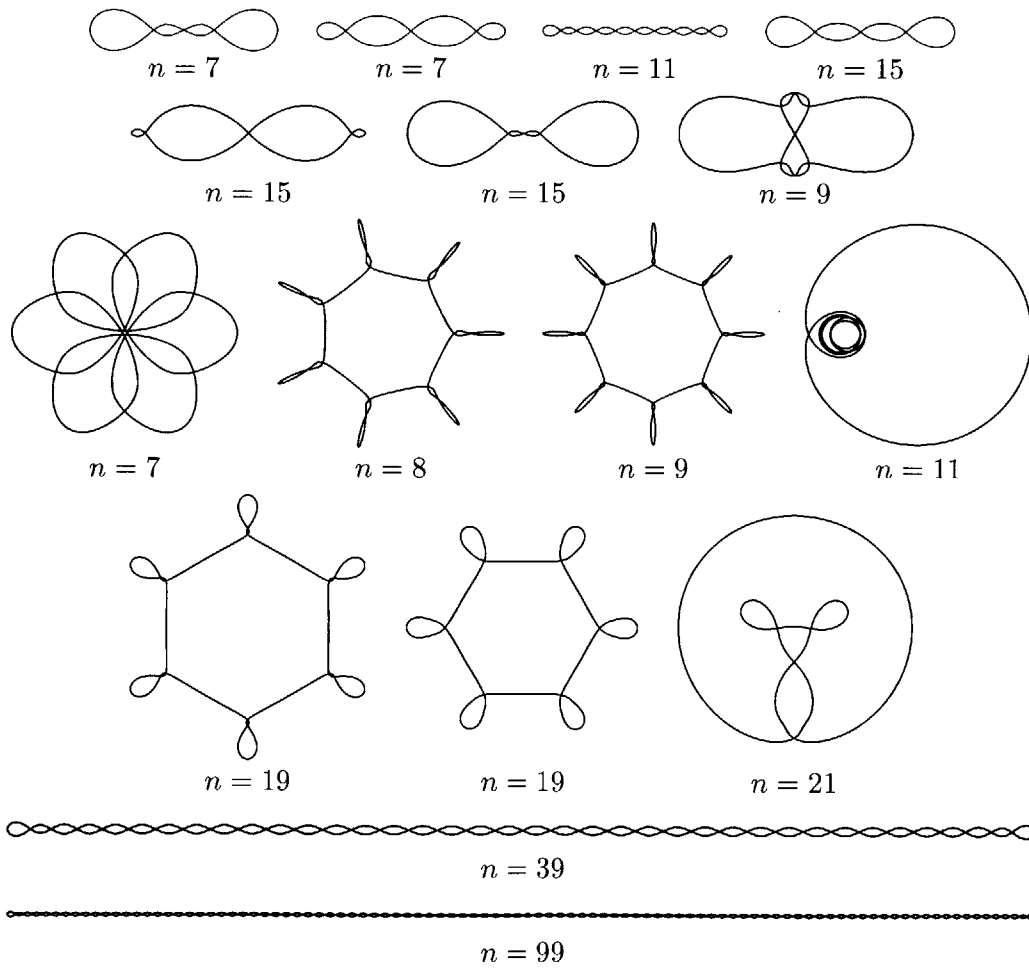


Figure 8.3: 16 additional  $n$ -body choreographies found by Simó [81].



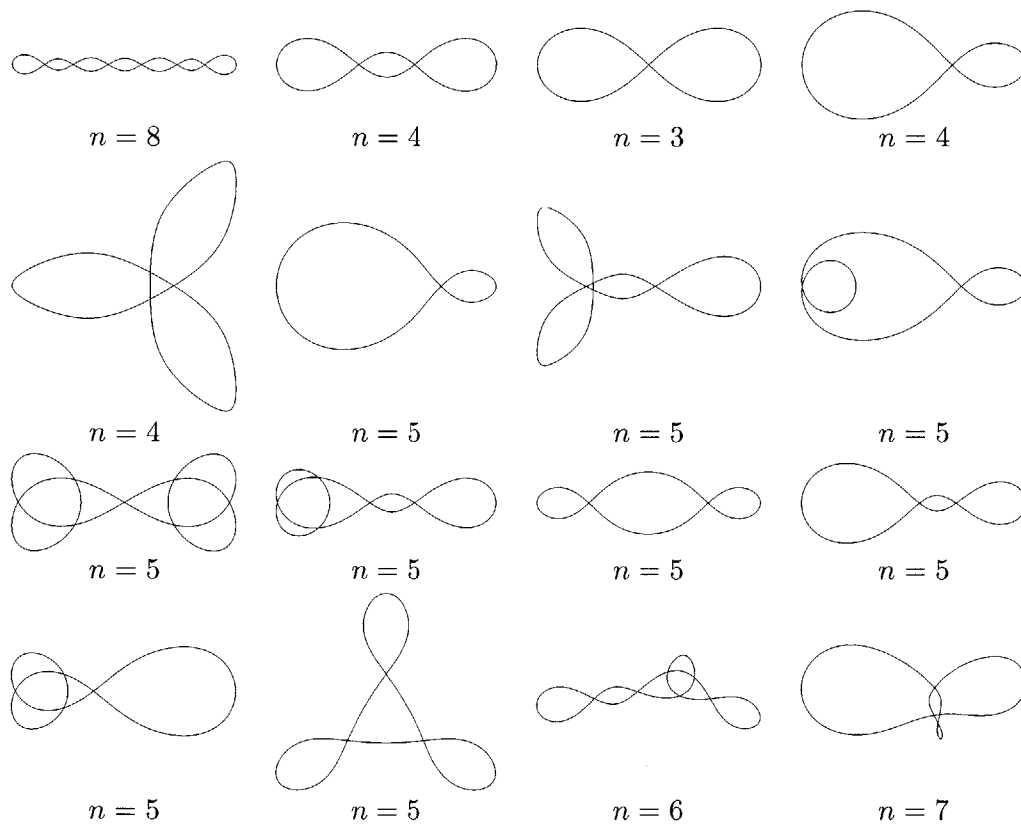


Figure 8.4: The  $n$ -body choreographies whose existence was proved with Kapela's computer-assisted methods [58, 55, 56].

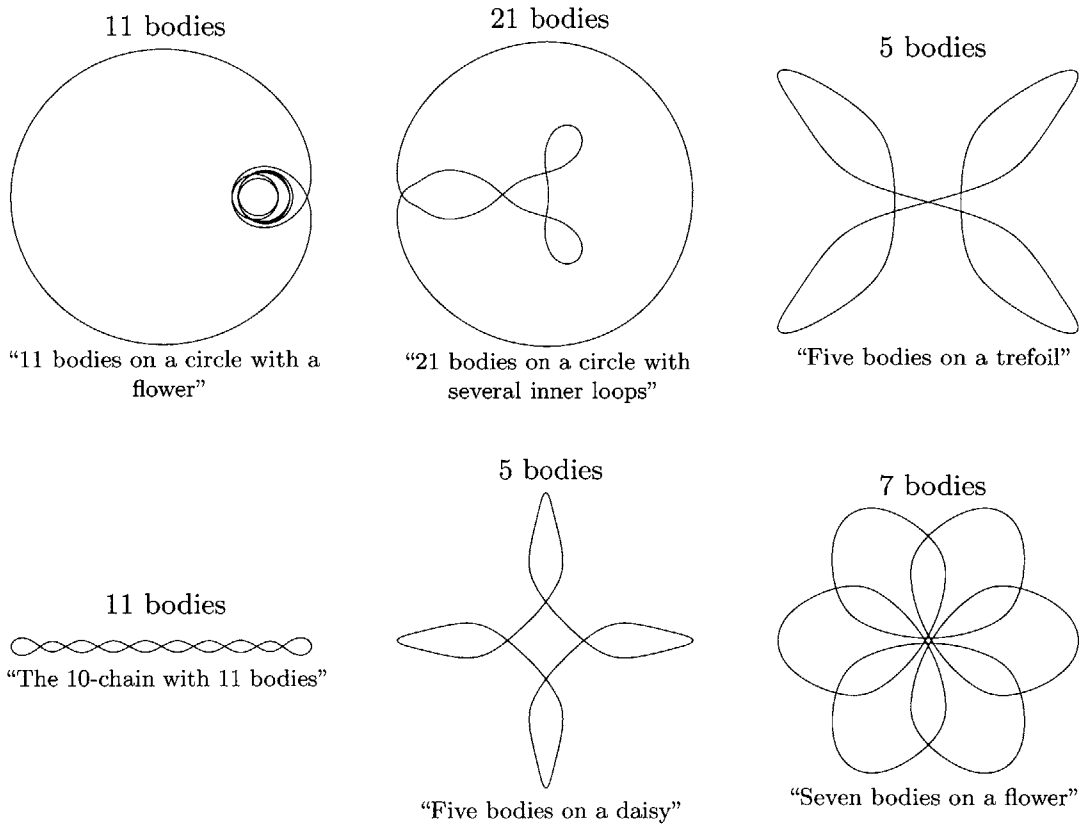


Figure 8.5: 6 of the 33 orbits from Simó's  $n$ -body data set [81] for which we have proven existence.

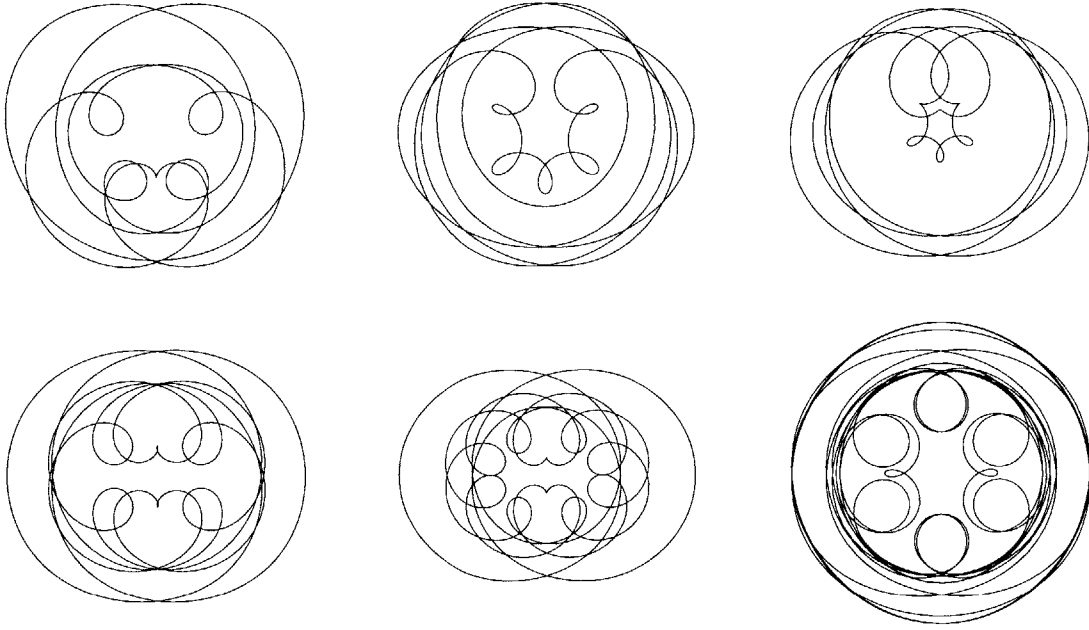


Figure 8.6: 6 of the 19 orbits from Simó's 3-body data set [82] for which we have proven existence.

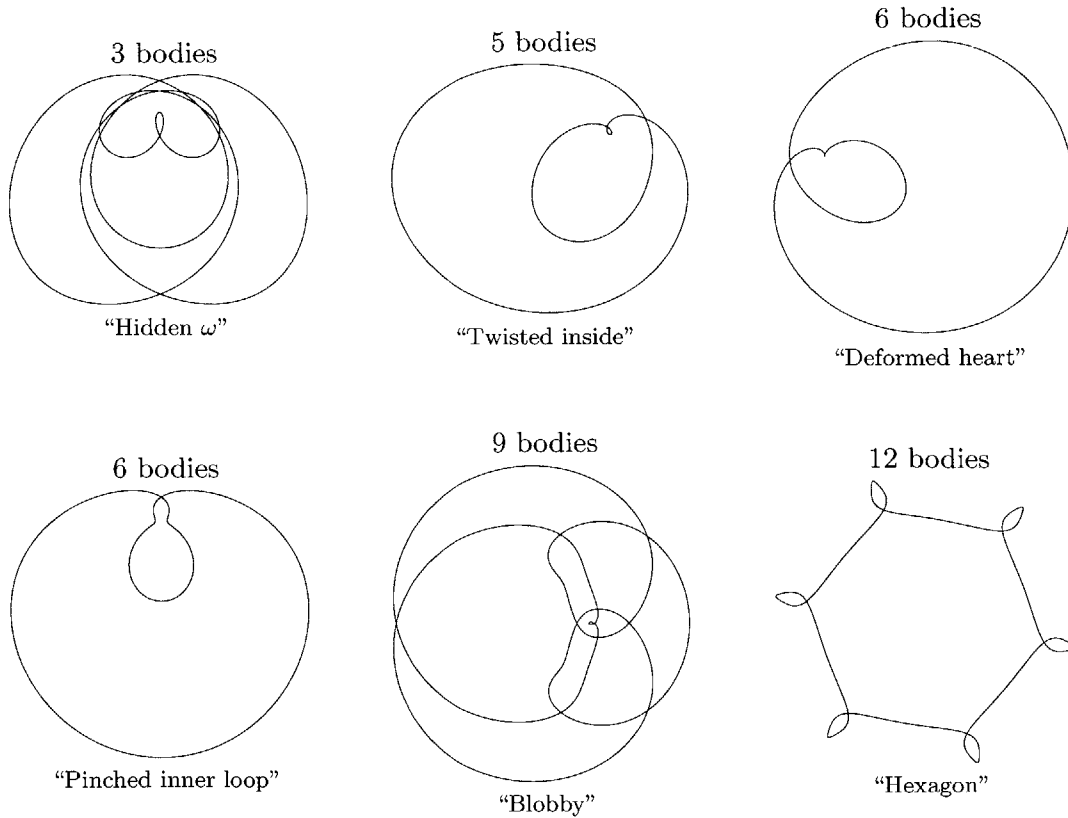


Figure 8.7: Six choreographies which we found using CHOREOGRAPHER and proved to exist.

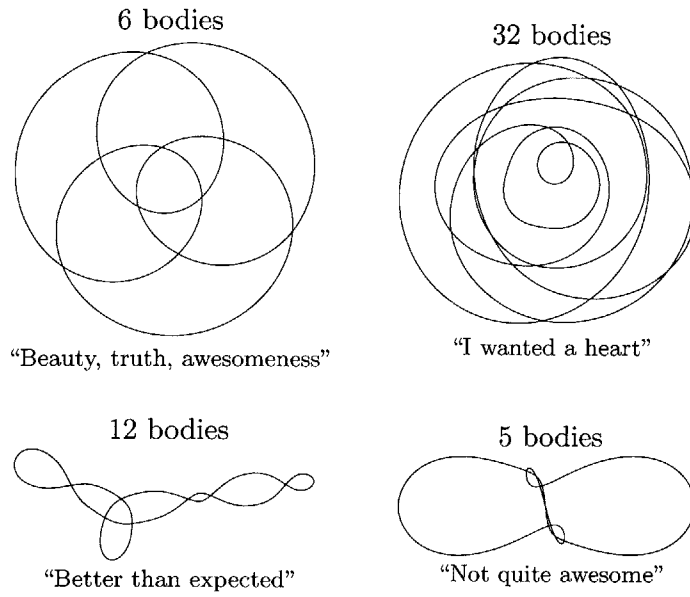


Figure 8.8: Four miscellaneous  $n$ -body choreographies which we found using CHOREO.JS and proved to exist.

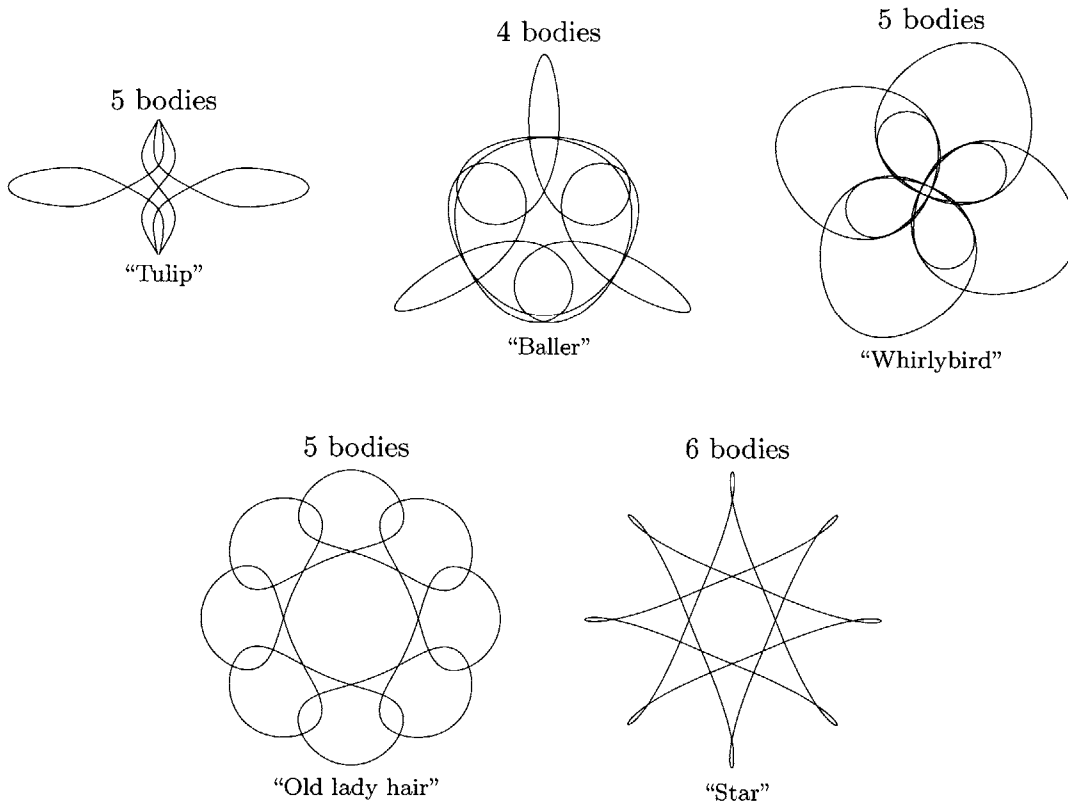


Figure 8.9: Five  $n$ -body choreographies which we found using CHOREO.JS and proved to exist. These choreographies were found using imposed symmetries.

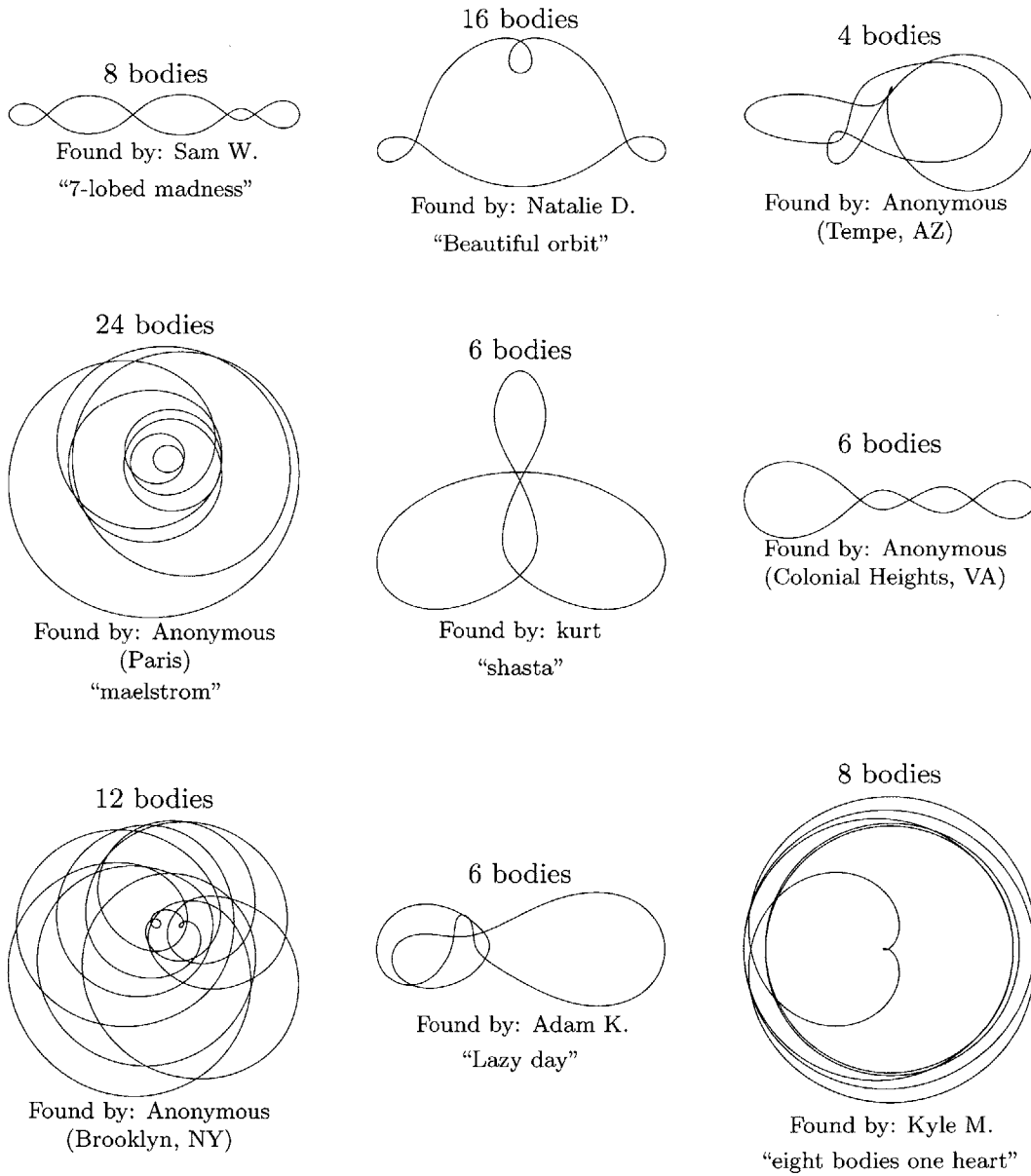


Figure 8.10: Miscellaneous  $n$ -body choreographies which were found by online users of CHOREO.JS and which we proved to exist.

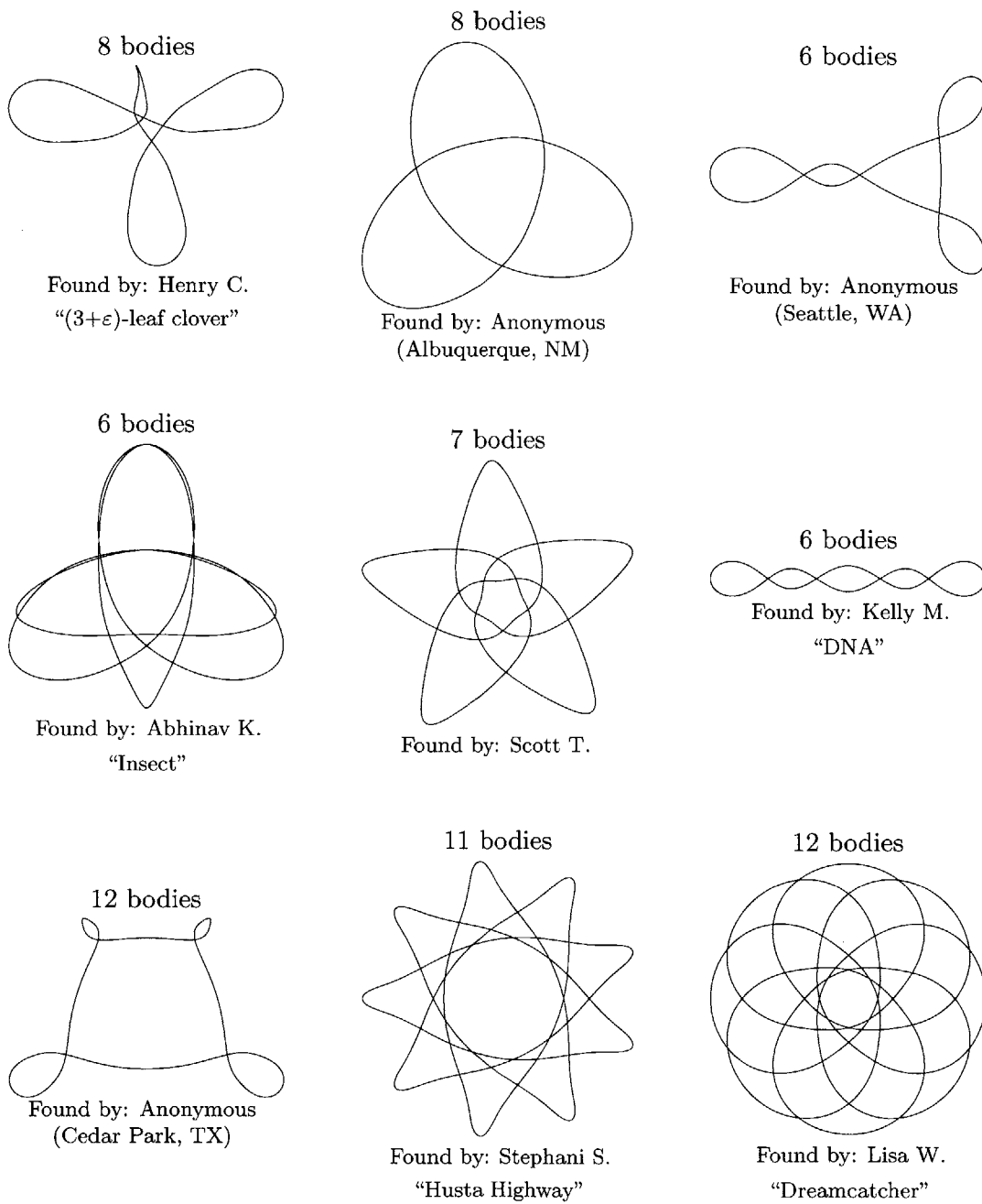


Figure 8.11: Miscellaneous  $n$ -body choreographies which were found by online users of CHOREO.JS and which we proved to exist.

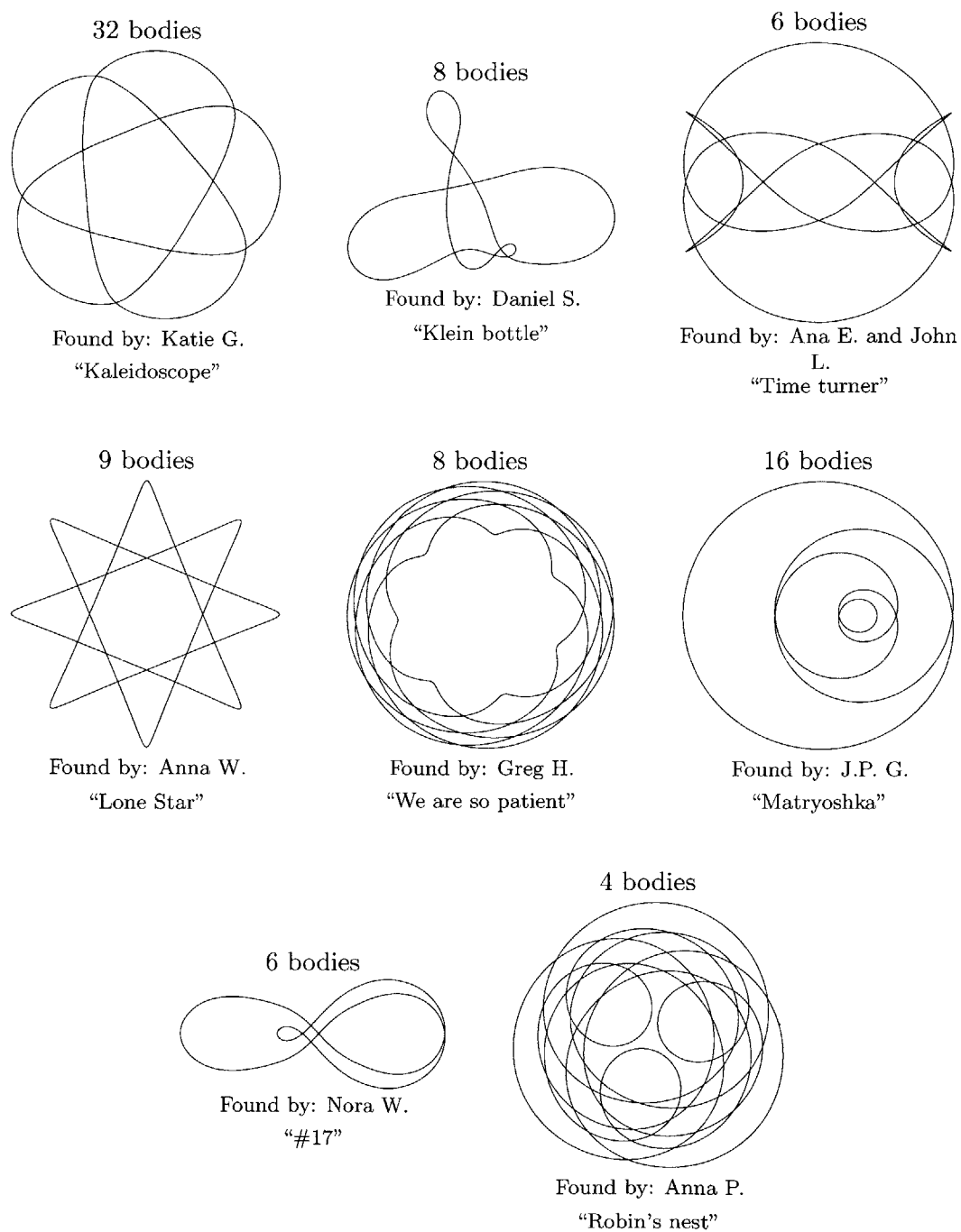


Figure 8.12: Miscellaneous  $n$ -body choreographies which were found by online users of CHOREO.JS and which we proved to exist.





# Bibliography

- [1] Alain Albouy, *The symmetric central configurations of four equal masses*, Hamiltonian dynamics and celestial mechanics (Seattle, WA, 1995), Contemp. Math., vol. 198, Amer. Math. Soc., Providence, RI, 1996, pp. 131–135. MR 1409157 (97g:70012)
- [2] Alain Albouy and Alain Chenciner, *Le problème des  $n$  corps et les distances mutuelles*, Invent. Math. **131** (1998), no. 1, 151–184. MR 1489897 (98m:70017)
- [3] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK users' guide*, third ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.
- [4] Richard Dean Anderson, *MacGyver*, 1985–1992, originally aired by ABC.
- [5] Gianni Arioli, *Periodic orbits, symbolic dynamics and topological entropy for the restricted 3-body problem*, Comm. Math. Phys. **231** (2002), no. 1, 1–24. MR 1947690 (2003j:70011)
- [6] ———, *Branches of periodic orbits for the planar restricted 3-body problem*, Discrete Contin. Dyn. Syst. **11** (2004), no. 4, 745–755. MR 2112702 (2005m:70051)
- [7] Gianni Arioli, Vivina Barutello, and Susanna Terracini, *A new branch of Mountain Pass solutions for the choreographical 3-body problem*, Comm. Math. Phys. **268** (2006), no. 2, 439–463. MR 2259202 (2008g:70015)
- [8] Gianni Arioli and Hans Koch, *Computer-assisted methods for the study of stationary solutions in dissipative systems, applied to the Kuramoto-Sivashinski equation*, Arch. Ration. Mech. Anal. **197** (2010), no. 3, 1033–1051. MR 2679365 (2011g:35179)
- [9] Gianni Arioli, Hans Koch, and Susanna Terracini, *Two novel methods and multi-mode periodic solutions for the Fermi-Pasta-Ulam model*, Comm. Math. Phys. **255** (2005), no. 1, 1–19. MR 2123374 (2006b:82083)
- [10] Vladimir I. Arnold, Valery V. Kozlov, and Anatoly I. Neishtadt, *Mathematical aspects of classical and celestial mechanics*, third ed., Encyclopaedia of Mathematical Sciences, vol. 3, Springer-Verlag, Berlin, 2006, [Dynamical systems. III], Translated from the Russian original by E. Khukhro. MR 2269239 (2008a:70001)
- [11] June Barrow-Green, *Poincaré and the three body problem*, History of Mathematics, vol. 11, American Mathematical Society, Providence, RI, 1997. MR 1415387 (97g:01013)
- [12] Vivina Barutello, Davide L. Ferrario, and Susanna Terracini, *Symmetry groups of the planar three-body problem and action-minimizing trajectories*, Arch. Ration. Mech. Anal. **190** (2008), no. 2, 189–226. MR 2448317 (2010g:70018)
- [13] Vivina Barutello and Susanna Terracini, *Action minimizing orbits in the  $n$ -body problem with simple choreography constraint*, Nonlinearity **17** (2004), no. 6, 2015–2039. MR 2097664 (2005k:70029)

- [14] ———, *A bisection algorithm for the numerical mountain pass*, NoDEA Nonlinear Differential Equations Appl. **14** (2007), no. 5-6, 527–539. MR 2374198 (2008j:58010)
- [15] Helmut Brass and Knut Petras, *Quadrature theory*, Mathematical Surveys and Monographs, vol. 178, American Mathematical Society, Providence, RI, 2011, The theory of numerical integration on a compact interval. MR 2840013 (2012h:65047)
- [16] H. Bruns, *Über die Integrale des Vielkörper-Problems*, Acta Math. **11** (1887), no. 1-4, 25–96. MR 1554748
- [17] Alain Chenciner, *Action minimizing periodic orbits in the Newtonian  $n$ -body problem*, Celestial mechanics (Evanston, IL, 1999), Contemp. Math., vol. 292, Amer. Math. Soc., Providence, RI, 2002, pp. 71–90. MR 1884893 (2004b:70023)
- [18] ———, *Action minimizing solutions of the Newtonian  $n$ -body problem: from homology to symmetry*, Proceedings of the International Congress of Mathematicians, Vol. III (Beijing, 2002) (Beijing), Higher Ed. Press, 2002, pp. 279–294. MR 1957539 (2004f:70026a)
- [19] ———, *Simple non-planar periodic solutions of the  $n$ -body problem*, Proceedings of the NDDS Conference, Kyoto, 2002, <http://www.imcce.fr/Equipes/ASD/preprints/prep.2002/Kyoto.2002.pdf>.
- [20] ———, *Are there perverse choreographies?*, New Advances in Celestial Mechanics and Hamiltonian Systems (J. Delgado, E.A. Lacomba, J. Llibre, and E. Prez-Chavela, eds.), Springer US, 2004, pp. 63–76.
- [21] Alain Chenciner, Jacques Féjoz, and Richard Montgomery, *Rotating eights. I. The three  $\Gamma_i$  families*, Nonlinearity **18** (2005), no. 3, 1407–1424. MR 2134901 (2005m:70053)
- [22] Alain Chenciner, Joseph Gerver, Richard Montgomery, and Carles Simó, *Simple choreographic motions of  $N$  bodies: a preliminary study*, Geometry, mechanics, and dynamics, Springer, New York, 2002, pp. 287–308. MR 1919833 (2003f:70019)
- [23] Alain Chenciner and Richard Montgomery, *A remarkable periodic solution of the three-body problem in the case of equal masses*, Ann. of Math. (2) **152** (2000), no. 3, 881–901. MR 1815704 (2001k:70010)
- [24] Alain Chenciner and Andrea Venturelli, *Minima de l'intégrale d'action du problème newtonien de 4 corps de masses égales dans  $\mathbf{R}^3$ : orbites "hip-hop"*, Celestial Mech. Dynam. Astronom. **77** (2000), no. 2, 139–152 (2001). MR 1820355 (2001k:70012)
- [25] T. Chiba, T. Imai, and H. Asada, *Can  $N$ -body systems generate periodic gravitational waves?*, Monthly Notices of the Royal Astronomical Society **377** (2007), no. 1, 269–272.
- [26] Barry Cipra, *How number theory got the best of the Pentium chip*, Science **267** (1995), no. 5195, 175.
- [27] Computer Assisted Proofs in Dynamics group, *CAPD: A C++ package for rigorous numerics*, <http://capd.wsb-nlu.edu.pl>.
- [28] Vittorio Coti Zelati, *Periodic solutions for  $N$ -body type problems*, Ann. Inst. H. Poincaré Anal. Non Linéaire **7** (1990), no. 5, 477–492. MR 1138534 (93a:70009)
- [29] Ian Davies, Aubrey Truman, and David Williams, *Classical periodic solution of the equal-mass  $2n$ -body problem,  $2n$ -ion problem and the  $n$ -electron atom problem*, Phys. Lett. A **99** (1983), no. 1, 15–18. MR 726510 (85a:70020)
- [30] Florin Diacu and Philip Holmes, *Celestial encounters*, Princeton University Press, Princeton, NJ, 1996, The origins of chaos and stability. MR 1435973 (98e:70006)

- [31] J. J. Dongarra, Jeremy Du Croz, Sven Hammarling, and I. S. Duff, *A set of level 3 basic linear algebra subprograms*, ACM Trans. Math. Softw. **16** (1990), no. 1, 1–17.
- [32] Jacques Féjóz, *The N-body problem*, to appear in UNESCO, Celestial Mechanics, 2013.
- [33] Davide L. Ferrario, *Symmetry groups and non-planar collisionless action-minimizing solutions of the three-body problem in three-dimensional space*, Arch. Ration. Mech. Anal. **179** (2006), no. 3, 389–412. MR 2208321 (2006m:70022)
- [34] ———, *Transitive decomposition of symmetry groups for the n-body problem*, Adv. Math. **213** (2007), no. 2, 763–784. MR 2332609 (2008f:70028)
- [35] Davide L. Ferrario and Susanna Terracini, *On the existence of collisionless equivariant minimizers for the classical n-body problem*, Invent. Math. **155** (2004), no. 2, 305–362. MR 2031430 (2005b:70010)
- [36] Michael H. Freedman, Zheng-Xu He, and Zhenghan Wang, *Möbius energy of knots and unknots*, Ann. of Math. (2) **139** (1994), no. 1, 1–50. MR 1259363 (94j:58038)
- [37] Toshiaki Fujiwara, Hiroshi Fukuda, and Hiroshi Ozaki, *Choreographic three bodies on the lemniscate*, J. Phys. A **36** (2003), no. 11, 2791–2800. MR 1965292 (2004b:70020)
- [38] Denis Gaidashev and Tomas Johnson, *Dynamics of the universal area-preserving map associated with period doubling: hyperbolic sets*, Nonlinearity **22** (2009), no. 10, 2487–2520. MR 2539765 (2010m:37066)
- [39] Zbigniew Galias and Piotr Zgliczyński, *Computer assisted proof of chaos in the Lorenz equations*, Phys. D **115** (1998), no. 3-4, 165–188. MR 1626596 (99h:58123)
- [40] Joseph L. Gerver, *The existence of pseudocollisions in the plane*, J. Differential Equations **89** (1991), no. 1, 1–68. MR 1088334 (92a:70008)
- [41] William B. Gordon, *A minimizing property of Keplerian orbits*, Amer. J. Math. **99** (1977), no. 5, 961–971. MR 0502484 (58 #19497)
- [42] C. G. Gray, G. Karl, and V. A. Novikov, *Direct use of variational principles as an approximation technique in classical mechanics*, Amer. J. Phys. **64** (1996), no. 9, 1177–1184. MR 1406152 (97f:70032)
- [43] Stephen Hawking, *The illustrated on the shoulders of giants: The great works of physics and astronomy*, Running Press, 2004.
- [44] Douglas Heggie, *A new outcome of binary–binary scattering*, Monthly Notices of the Royal Astronomical Society **318** (2000), no. 4, L61–L63.
- [45] Douglas Heggie and Piet Hut, *The gravitational million-body problem*, Cambridge University Press, Cambridge, 2003. MR 2025098 (2005j:85003)
- [46] Michael T. Heideman, Don H. Johnson, and C. Sidney Burrus, *Gauss and the history of the fast Fourier transform*, Arch. Hist. Exact Sci. **34** (1985), no. 3, 265–277. MR 815154 (87f:01018)
- [47] Michel Hénon, *Families of periodic orbits in the three-body problem*, Celestial mechanics **10** (1974), no. 3, 375–388.
- [48] Yozo Hida, Xiaoye S. Li, and David H. Bailey, *Algorithms for quad-double precision floating point arithmetic*, Proceedings of the 15th Symposium on Computer Arithmetic, IEEE Computer Society Press, 2001, pp. 155–162.

- [49] Erik Holmberg, *On the Clustering Tendencies among the Nebulae. II. a Study of Encounters Between Laboratory Models of Stellar Systems by a New Integration Procedure.*, *Astrophys. J.* **94** (1941), no. 3, 385–395.
- [50] Takumi Ichita, Kei Yamada, and Hideki Asada, *Post-newtonian effects on lagrange’s equilateral triangular solution for the three-body problem*, *Phys. Rev. D* **83** (2011), 084026.
- [51] *IEEE 754: An interview with William Kahan*, *Computer* **31** (1998), no. 3, 114–115, notes available at <http://www.cs.berkeley.edu/~wkahan/ieee754status/754story.html>.
- [52] *IEEE standard for floating-point arithmetic*, *IEEE Std 754-2008* (2008), 1–58.
- [53] Tatsunori Imai, Takamasa Chiba, and Hideki Asada, *Choreographic solution to the general-relativistic three-body problem*, *Phys. Rev. Lett.* **98** (2007), 201102.
- [54] Emmanuelle Julliard-Tosel, *Brun’s theorem: the proof and some generalizations*, *Celestial Mech. Dynam. Astronom.* **76** (2000), no. 4, 241–281. MR 1800399 (2002k:70020)
- [55] Tomasz Kapela, *N-body choreographies with a reflectional symmetry—computer assisted existence proofs*, *EQUADIFF 2003*, World Sci. Publ., Hackensack, NJ, 2005, pp. 999–1004. MR 2185163
- [56] Tomasz Kapela and Carles Simó, *Computer assisted proofs for nonsymmetric planar choreographies and for stability of the Eight*, *Nonlinearity* **20** (2007), no. 5, 1241–1255, With multimedia enhancements available from the abstract page in the online journal. MR 2312391 (2008f:70023)
- [57] ———, *Rigorous KAM results around arbitrary periodic orbits for Hamiltonian systems*, *arXiv:1105.3235*, 2011.
- [58] Tomasz Kapela and Piotr Zgliczyński, *The existence of simple choreographies for the N-body problem—a computer-assisted proof*, *Nonlinearity* **16** (2003), no. 6, 1899–1918. MR 2012847 (2004h:70019)
- [59] Johannes Kepler, *Harmonices mundi*, 1619, scan available at [http://posner.library.cmu.edu/Posner/books/book.cgi?call=520\\_K38PI](http://posner.library.cmu.edu/Posner/books/book.cgi?call=520_K38PI).
- [60] ———, *Selections from Kepler’s Astronomia nova*, Green Lion Press, 2004, translation of the 1609 original.
- [61] Hiroshi Kokubu, Daniel Wilczak, and Piotr Zgliczyński, *Rigorous verification of cocoon bifurcations in the Michelson system*, *Nonlinearity* **20** (2007), no. 9, 2147–2174. MR 2351028 (2008j:37111)
- [62] L. D. Landau and E. M. Lifshitz, *Mechanics*, Course of Theoretical Physics, Vol. 1. Translated from the Russian by J. B. Bell, Pergamon Press, Oxford, 1960. MR 0120782 (22 #11531)
- [63] Kenneth Levenberg, *A method for the solution of certain non-linear problems in least squares*, *Quart. Appl. Math.* **2** (1944), 164–168. MR 0010666 (6,52a)
- [64] Christian Marchal, *The family  $P_{12}$  of the three-body problem—the simplest family of periodic orbits, with twelve symmetries per period*, *Celestial Mech. Dynam. Astronom.* **78** (2000), no. 1-4, 279–298 (2001), *New developments in the dynamics of planetary systems* (Badhofgastein, 2000). MR 1845981 (2002k:70015)
- [65] ———, *How the method of minimization of action avoids singularities*, *Celestial Mech. Dynam. Astronom.* **83** (2002), no. 1-4, 325–353, *Modern celestial mechanics: from theory to applications* (Rome, 2001). MR 1956531 (2004b:70024)

- [66] Donald W. Marquardt, *An algorithm for least-squares estimation of nonlinear parameters*, J. Soc. Indust. Appl. Math. **11** (1963), 431–441. MR 0153071 (27 #3040)
- [67] Gregory Minton, *Choreographer.js*, <http://gminton.org/choreo.html>, accessed: 08-14-2013.
- [68] James Montaldi, *Planar choreographies*, <http://www.ma.man.ac.uk/~jm/Choreographies/>, accessed: 07-29-2013.
- [69] James Montaldi and Katrina Steckles, *Classification of symmetry groups for planar  $n$ -body choreographies*, arXiv:1305.0470, 2013.
- [70] Richard Montgomery, *Action spectrum and collisions in the planar three-body problem*, Celestial mechanics (Evanston, IL, 1999), Contemp. Math., vol. 292, Amer. Math. Soc., Providence, RI, 2002, pp. 173–184. MR 1884899 (2003a:70012)
- [71] Cristopher Moore, *Braids in classical dynamics*, Phys. Rev. Lett. **70** (1993), no. 24, 3675–3679. MR 1220207 (94d:58055)
- [72] Cristopher Moore and Michael Nauenberg, *New periodic orbits for the  $n$ -body problem*, arXiv:math/0511219, 2005.
- [73] Michael Nauenberg, *Periodic orbits for three particles with finite angular momentum*, Phys. Lett. A **292** (2001), no. 1-2, 93–99. MR 1916506 (2004c:70023)
- [74] ———, *Continuity and stability of families of figure eight orbits with finite angular momentum*, Celestial Mech. Dynam. Astronom. **97** (2007), no. 1, 1–15. MR 2289174 (2007k:70013)
- [75] Isaac Newton, *Philosophiæ naturalis principia mathematica*, The Royal Society, 1687, scan available at <http://cudl.lib.cam.ac.uk/view/PR-ADV-B-00039-00001/>.
- [76] Richard S. Palais, *The principle of symmetric criticality*, Comm. Math. Phys. **69** (1979), no. 1, 19–30. MR 547524 (81c:58026)
- [77] Henri Poincaré, *Les méthodes nouvelles de la mécanique céleste*, 1892–1899, 3 tomes.
- [78] ———, *Sur les solutions périodiques et le principe de moindre action*, C. R. Acad. Sci. Paris, Sér. I Math. **123** (1896), 915–918.
- [79] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical recipes*, third ed., Cambridge University Press, Cambridge, 2007, The art of scientific computing. MR 2371990 (2009b:65001)
- [80] Mitsuru Shibayama, *Variational proof of the existence of the super-eight orbit in the four-body problem*, arXiv:1307.2959, 2013.
- [81] Carles Simó, *Choreographies of the  $N$ -body problem*, <http://www.maia.ub.edu/dsg/nbody.html>, accessed: 07-29-2013.
- [82] ———, *Choreographies of the planar three body problem*, <http://www.maia.ub.edu/dsg/3body.html>, accessed: 07-29-2013.
- [83] ———, *Periodic orbits of the planar  $N$ -body problem with equal masses and all bodies on the same path*, The Restless Universe: Applications of Gravitational  $n$ -body Dynamics To Planetary, Stellar, and Galactic Systems (B. A. Steves and A. J. Maciejewski, eds.), Scottish Universities Summer School in Physics, 2000, pp. 265–284.
- [84] ———, *New families of solutions in  $N$ -body problems*, European Congress of Mathematics, Vol. I (Barcelona, 2000), Progr. Math., vol. 201, Birkhäuser, Basel, 2001, pp. 101–115. MR 1905315 (2003g:70012)

- [85] ———, *Dynamical properties of the figure eight solution of the three-body problem*, Celestial mechanics (Evanston, IL, 1999), Contemp. Math., vol. 292, Amer. Math. Soc., Providence, RI, 2002, pp. 209–228. MR 1884902 (2003b:70013)
- [86] E. Myles Standish, Jr., *New periodic orbits in the general problem of three bodies*, Periodic Orbits, Stability and Resonances (G.E.O. Giacaglia, ed.), Springer Netherlands, 1970, pp. 375–381.
- [87] Ian Stewart, *Visions of infinity*, Basic Books, New York, 2013, The great mathematical problems. MR 3025386
- [88] Patrick Stewart, *Star Trek: The Next Generation*, 1987–1994, originally aired by CBS.
- [89] Volker Strassen, *Gaussian elimination is not optimal*, Numer. Math. **13** (1969), 354–356. MR 0248973 (40 #2223)
- [90] Šuvakov, Milovan and Dmitrašinović, V., *Three classes of Newtonian three-body planar periodic orbits*, Phys. Rev. Lett. **110** (2013), 114301.
- [91] Susanna Terracini, *n-body problem and choreographies*, Mathematics of Complexity and Dynamical Systems (Robert A. Meyers, ed.), Springer New York, 2011, pp. 1043–1069.
- [92] Mauri Valtonen and Hannu Karttunen, *The three-body problem*, Cambridge University Press, Cambridge, 2006. MR 2223553 (2007d:70016)
- [93] Robert J. Vanderbei, *New orbits for the n-body problem*, Annals of the New York Academy of Sciences **1017** (2004), no. 1, 422–433.
- [94] Andrea Venturelli, *Une caractérisation variationnelle des solutions de Lagrange du problème plan des trois corps*, C. R. Acad. Sci. Paris Sér. I Math. **332** (2001), no. 7, 641–644. MR 1841900 (2002h:70021)
- [95] H. von Zeipel, *Sur les singularités du problème des n corps*, Ark. Mat. Astr. Fys. **4** (1908), 1–4.
- [96] R. Clint Whaley, Antoine Petitet, and Jack J. Dongarra, *Automated empirical optimization of software and the ATLAS project*, Parallel Computing **27** (2001), no. 1–2, 3–35, Also available as University of Tennessee LAPACK Working Note #147, UT-CS-00-448, 2000, <http://www.netlib.org/lapack/lawns/lawn147.ps>.
- [97] Daniel Wilczak and Piotr Zgliczyński, *Period doubling in the Rössler system—a computer assisted proof*, Found. Comput. Math. **9** (2009), no. 5, 611–649. MR 2534406 (2010h:37194)
- [98] Zhihong Xia, *The existence of noncollision singularities in Newtonian systems*, Ann. of Math. (2) **135** (1992), no. 3, 411–468. MR 1166640 (93h:70005)
- [99] Kei Yamada and Hideki Asada, *Triangular solution to the general relativistic three-body problem for general masses*, Phys. Rev. D **86** (2012), 124029.
- [100] Piotr Zgliczynski,  *$C^1$  Lohner algorithm*, Found. Comput. Math. **2** (2002), no. 4, 429–465. MR 1930946 (2003h:65062)