The Triangular Bi-Pyramid Minimizes a Range of Power Law Potentials

Richard Evan Schwartz *

August 17, 2016

Abstract

Combining a brilliant observation of A. Tumanov with a divide-and-conquer approach to Thomson's 5-electron problem, we give a rigorous computer-assisted proof that the triangular bi-pyramid is the unique minimizer with respect to the potential

$$R_s(r) = \text{sign}(s)\frac{1}{r^s}$$

amongst all spherical 5 point configurations when $s \in (-2, 13] - \{0\}$. The case $s = -1$ corresponds to Polya's problem and the case $s = 1$ corresponds to Thomson's problem. The lower bound of 13 is fairly close to the presumed optimal cutoff of 15.040809....

1 Introduction

1.1 The Energy Minimization Problem

Let S^2 denote the unit sphere in $\boldsymbol{R}^3$ and let $P = \{p_1, ..., p_n\}$ be a finite list of pairwise distinct points on S^2. Given some function $f : (0, 2] \to \boldsymbol{R}$ we can form the energy potential

$$\mathcal{E}_f(P) = \sum_{i<j} f(\|p_i - p_j\|). \tag{1}$$

* Supported by N.S.F. Research Grant DMS-1204471

For fixed f and n, one can ask which configuration(s) minimize $\mathcal{E}_f(P)$.

For this problem, the energy functional $f = R_s$, where

$$R_s(r) = \text{sign}(s)\frac{1}{r^s}, \tag{2}$$

is a natural one to consider. At least for $s > 0$, this is called the *Riesz energy*, and we will also use this name when $s < 0$. The case $s = 1$, where $f(r) = 1/r$, corresponds to the electrostatic potential. This case is known as *Thomson's problem*. See [**Th**]. The case $s = -1$ corresponds to the problem of placing points on the sphere so as to maximize the total sum of the distances between pairs. This is known as *Polya's problem*.

There is a large literature on the energy minimization problem. See [**C**] for some early local results. The online website [**CCD**] is a compilation of experimental results. The paper [**SK**] gives a nice survey in the two dimensional case, with an emphasis on the case when n is large. See also [**RSZ**]. The paper [**BBCGKS**] gives a survey of results, both theoretical and experimental, about highly symmetric configurations in higher dimensions.

When $n = 2, 3, 4, 6, 12$, the most symmetric configurations are the unique minimizers for all R_s with $s \in (-2, \infty) - \{0\}$. For the cases $n = 4, 6$ (tetrahedron, octahedron) see [**Y**]. For the case $n = 12$ (icosahedron) see [**A**]. All these cases are covered by the vast result in [**CK**, Theorem 1.2].

The case $n = 5$ has been notoriously intractable. In this case, there is a general feeling that for a wide range of energy choices, and in particular for the Riesz energies, that the global minimizer is either the triangular bi-pyramid or else some pyramid with square base. The *triangular bi-pyramid* (TBP) is the configuration of 5 points with one point at the north pole, one at the south pole, and three arranged in an equilateral triangle around the equator.

[**HS**] has a rigorous computer-assisted proof that the TBP is the unique minimizer for R_{-1} (Polya's problem) and my paper [**S1**] has a rigorous computer-assisted proof that the TBP is the unique minimizer for R_1 (Thomson's problem) and R_2. The paper [**DLT**] gives a traditional proof that the TBP is the unique minimizer for the logarithmic potential. The TBP is <u>not</u> the minimizer for R_s when $s > 15.040809....$

Define

$$G_k(r) = (4 - r^2)^k, \qquad k = 1, 2, 3, ... \tag{3}$$

In [**T**], A. Tumanov gives a traditional proof of the following result.

Theorem 1.1 (Tumanov) *Let $f = a_1G_1 + a_2G_2$ with $a_1, a_2 > 0$. The TBP is the unique global minimizer with respect to f. Moreover, a critical point of f must be the TBP or a pyramid with square base.*

As an immediate corollary, the TBP is a minimizer for G_1 and G_2. Tumanov points out that these potentials do not have an obvious geometric interpretation, but they are amenable to a tradicional analysis. He points out his result might be a step towards proving the that TBP minimizes a range of power law potentials. He makes the following observation: If the TBP is the unique minimizer for G_3 and G_5, then the TBP is the unique minimizer for R_s provided that $s \in (-2, 2] - \{0\}$.

1.2 Results

Tumanov did not give a proof of his observation in [**T**], but we will prove a more extensive result.

Lemma 1.2 *Suppose the TBP is the unique minimizer for G_3, G_4, G_5, G_6 and also for $G_{10}^{\#} = G_{10}+28G_5+102G_2$. Then the TBP is the unique minimizer for R_s for any $s \in (-2, 13] - \{0\}$.*

§8.1 has a clear description of the idea behind the result. This kind of technique, in some form, is used in many papers on energy minimization. See [**BDHSS**] for very general ideas like this.

The main purpose of this paper is to prove the following result.

Theorem 1.3 (Main) *The TBP is the unique minimizer for G_3, G_4, G_5, G_6 and $G_{10}^{\#}$.*

Combining Theorem 1.3 with Lemma 1.2, we get

Theorem 1.4 *The TBP is the unique minimizer for R_s if $s \in (-2, 13] - \{0\}$.*

The lower bound of -2 in Corollary 1.4 is sharp, because the TBP is not a minimizer for R_s when $s < -2$. We separate this case out for special mention.

Theorem 1.5 *Let $p \in (0, 2)$ be arbitrary. Then the TBP is the unique maximizer, amongst 5-point configurations on the unit sphere, of the sums of the pth powers of the distances between the points.*

In particular, our methods give another solution of Polya's problem. This is the case $p = 1$, which is solved in [**HS**] by a different kind of computer-assisted proof.

1.3 Outline of the Paper

In §2 I will discuss the moduli space of normalized configurations of 5 points on the sphere. The key idea is to use stereographic projection to move the points into $\boldsymbol{R}^2 \cup \infty$. This gives the moduli space a natural Euclidean structure, making a divide-and-conquer algorithm easier to manage, even though the expressions for the energy potentials are more complicated. Our basic object is a *block*, a certain kind of rectangular solid in the moduli space.

In §3 I will prove some elementary facts about spherical geometry, and define the main quantities associated to blocks. These quantities form the basis of the main technical result, the Energy Theorem.

In §4 I will prove the Energy Theorem, Theorem 4.1, which gives an efficient estimate on the minimum energy of all configurations contained within a block B based on the minimum energy of the configurations associated to the vertices of B and an auxiliary error term.

In §5 I will describe a divide-and-conquer algorithm based on the Energy Theorem. Given a potential function F from the Main Theorem and a small neighborhood B_0 of the TBP, the program does a recursive search through the poset of dyadic blocks, eliminating a block B if $B \subset B_0$ or if the Energy Theorem determines that all configurations in B have higher energy than the TBP. (Sometimes we also eliminate B by symmetry considerations.) If B is not eliminated, B is subdivided and the pieces are then tested. If the program runs to completion, it constitutes a proof that some F-minimizer lies in B_0. See Lemmas 5.1 and 5.2.

In §6 I prove that the Hessian of the F energy functional is positive definite throughout B_0 for each relevant choice of F. This combines with Lemmas 5.1 and 5.2 to prove the Main Theorem. Our local analysis is essentially Taylor's Theorem with Remainder.

In §7-8 I prove Lemma 1.2.

In §9 I will explain some details about the computer program. All the calculations which go into Lemmas 5.1 and 5.2 are done with interval arithmetic to avoid roundoff error. The calculations for Lemma 1.2 are all exact integer calculations. The only place where floating point calculations occur without any control is for the extremely loose bound in Equation 91. For convenience, we rely on Mathematica [**W**] to estimate some numbers which involve square roots of integers. Whereas Mathematica can produce decimal expansions of these numbers up to thousands of digits, we just need the answer to be accurate up to a factor of 3.

1.4 Discussion

Our proof of Theorem 1.3 will not work for G_1 because the TBP is not the unique minimizer in this case. However, this case is relatively trivial. Our proof would work for G_2, but the setup in §2 needs to be modified somewhat. We don't pursue this.

One can check that some pyramid with square base has lower G_7 energy than the TBP. I checked by calculating the hessian of the energy, as in §6, that the TBP is not even a local minimizer for $G_8, ..., G_{100}$. This fact led me to search for $G_{10}^{\#}$ and other combinations like it.

The analysis of $G_{10}^{\#}$ is tougher than the analysis of G_3, G_4, G_5, G_6. For the (dis)interested reader, we mention that one can still prove Theorem 1.4 for all $s \in (-2, 6] - \{0\}$ without $G_{10}^{\#}$. Indeed, an earlier version of the paper had this more limited result and then I decided to push harder on the method.

I am starting to think that I can get the definitive result on the positive side – i.e. all the way to the presumed cutoff of 15.040809... Here is a sketch. The same techniques we use in §8 establish the following result.

Lemma 1.6 *Let X_0 denote the TBP and let X be any other 5-point configuration on the unit sphere. Suppose that the energy of X exceeds the energy of X_0 with respect to*

$$G_2^* = 3G_2 - 2G_1, \qquad G_5^* = G_5 - 5G_1, \qquad G_{10}^* = G_{10} + 13G_5 + 55G_2.$$

Then X is not the minimizer for the R_s energy when $s \in [13, 15.05]$.

My program runs to completion on G_2^* and G_5^*. Since the TBP is not the minimizer for G_{10}^*, my program will not run to completion. However, the program eliminates most other configurations. Precisely, my program (coupled with the same local analysis as is done in §6) seems to show that a minimizer for G_{10}^* either is the TBP or else can be rotated so that its stereographic projection has points $z_0, z_1, z_2, z_3, \infty$, where

- $z_0 \in [13/16] \times \{0\}$.
- $z_1 \in [-1/16, 1/16] \times [-5/8, -15/16]$.
- $z_2 \in [-1, -3/4] \times [-1/16, 1/16]$.
- $z_3 \in [-1/16, 1/16] \times [5/8, 15/16]$.

Here is a picture of these regions in the plane. The 4 dots represent the stereographic projection of the nearest copy of the TBP. The grey circle is the unit circle.

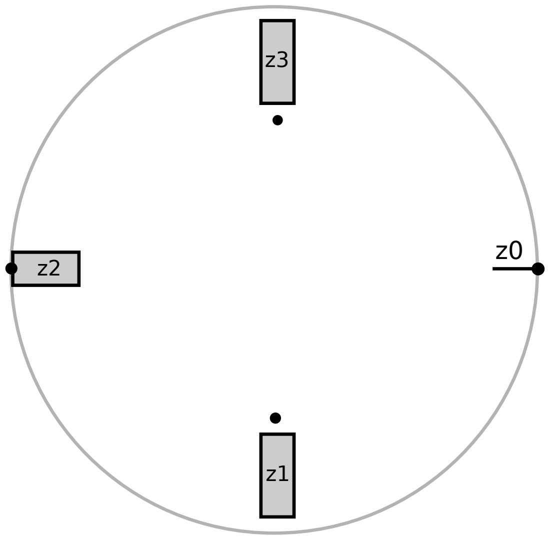

Figure 2: Location of the minimizer for power laws in $[13, 15.05]$.

These configurations are quite close to pyramids with square base, and for all of them it appears (from a billion random trials) that the symmetrization operation

$$(z_0, z_1, z_2, z_3) \to (\rho_0, -i\rho_1, -\rho_0, i\rho_1), \qquad \rho_i = \left|\frac{z_i - z_{i+1}}{2}\right|$$

decreases the energy with respect to G_k for all $k = 2, ..., 14$. This result, combined with another approximation lemma like Lemma 1.6, seems sufficient to show that for any $s \in [13, 15.05]$ the minimizer with respect to R_s is either the TBP or a pyramid with square base. That would show the existence of a cutoff around 15.040809 and allow one to compute its value, say, to a million decimal places.

This paper has many points in common with [**S1**] but the exposition here is simpler and cleaner. Now I have a better understanding of which ingredients are important in the running time of the program, I was able to give a more efficient treatment of the problem. In the end, the estimates and the routine are pretty simple. I think that a competent programmer could reproduce the results.

As in [**S1**], the crucial feature of this paper is a good error estimate, the Energy Theorem, which allows one to control the energies of all configurations within a block of the configuration space just by considering finitely many of those configurations. A good estimate means the difference between a feasible calculation and one that would outlast the universe. I view the Energy Theorem as the main mathematical contribution of the paper.

All the estimates that go into the Energy Theorem are rational functions of the inputs. So, one could in theory give a proof that just uses integer arithmetic. An early version of my program ran integer arithmetic calculations, but they seemed very slow. I hope to eventually optimize the code for integer calculations and try it out seriously. It would be nice to have an exact integer proof of the Main Theorem.

1.5 Companion Computer Programs

The computer code involved in this paper is publicly available by download from my website. See §9.1 for details. There is a small amount of Mathematica code, and then there is are two large Java programs. The main java program runs the divide-and-conquer program which establishes Lemma 5.1 and 5.2. The other java program does the calculations for Lemma 1.2. The Mathematica code is used in the local analysis of the Hessian in §6. The Java programs come with graphical user interfaces, and each one has built in documentation and debuggers. The interfaces let the user watch the programs in action, and check in many ways that they are operating correctly.

1.6 Acknowledgements

I would like to thank Henri Cohn, Abhinav Kumar, Curtis McMullen, Jill Pipher, and Sergei Tabachnikov for discussions related to this paper. I would like to thank Alexander Tumanov for his great observation. Also, I would like to thank I.C.E.R.M. for facilitating this project. My interest in this problem was rekindled during discussions about point configuration problems around the time of the I.C.E.R.M. Science Advisory Board meeting in Nov. 2015.

2 The Configuration Space

2.1 Normalized Configurations

Recall that

$$G_k(r) = (4 - r^2)^k, \qquad G_{10}^{\#} = G_{10} + 28G_5 + 102G_2. \tag{4}$$

The TBP has 6 bonds of length $\sqrt{2}$, and 3 bonds of length $\sqrt{3}$, and one bond of length 2. Hence, the G_k energy of the TBP is $3 + 6 \times 2^k$. In particular,

$$(\mathcal{E}_3, \mathcal{E}_4, \mathcal{E}_5, \mathcal{E}_6, \mathcal{E}_{10}^{\#}) = (51, 99, 195, 387, 14361). \tag{5}$$

Let $\widehat{p}_0, ..., \widehat{p}_4$ be a configuration of 5 points on S^2. We call this configuration *normalized* if

$$\widehat{p}_4 = (0,0,1), \qquad \|\widehat{p}_4 - \widehat{p}_0\| \leq \|\widehat{p}_4 - \widehat{p}_i\|, \qquad \forall i \in \{1,2,3\}. \tag{6}$$

The choice of the bizarre constant in the next lemma will become clearer in the next section.

Lemma 2.1 *For a normalized minimizer, we have* $\|\widehat{p}_4 - \widehat{p}_0\| > 1/2$ *and* $\|\widehat{p}_4 - \widehat{p}_i\| > 4/\sqrt{13}$ *for* $i = 1,2,3$*.*

Proof: The values of $G_3, G_4, G_5, G_6, G_{10}^{\#}$ evaluated at $1/2$ respectively exceed the values in Equation 5, so a normalized minimizer cannot have a point within $1/2$ of $\widehat{p}_4$.

If $\|\widehat{p}_4 - \widehat{p}_i\| \leq 4/\sqrt{13}$ for some $i = 1,2,3$ then $\|\widehat{p}_4 - \widehat{p}_0\| \leq 4/\sqrt{13}$ as well. But then the sum of the energies coming from bonds between $\widehat{p}_4$ and other points is at least $2G_k(4/\sqrt{13}) = 2(36/13)^k$. This exceeds the value given in Equation 5 except in the case of G_3, where all we get is $2(36/13)^3 > 42$.

We need a special argument for G_3. Notice that the same argument as above shows that the distance between any two points for the 4-point minimizer is at least 1, because otherwise the energy of the single bond, 3^k, would exceed the energy $6 \times (4/3)^k$ of the regular tetrahedron. Now observe that the function G_3 is convex decreasing on the interval $[1,2]$. But the regular tetrahedron is the mimimizer for any convex decreasing potential – see [**CK**] for a proof. Hence, the regular tetrahedron is the global minimizer for G_3 amongst all 4 point configurations. Hence the sum of energies in the bonds not involving $\widehat{p}_4$ is at least $6 \times (4/3)^3 > 14$. Since $14 + 42 > 51$, our configuration could not be an minimizer. ♠

2.2 Stereographic Projection

As in [**S1**] we work mainly with $\boldsymbol{R}^2 \cup \infty$ rather than on S^2. Our reason for this is that a configuration space based on points in $\boldsymbol{R}^2$ has a natural flat structure, and lends itself well to divide-and-conquer algorithms.

We map S^2 to $\boldsymbol{R}^2 \cup \infty$ using *stereographic projection*:

$$\Sigma(x,y,z) = \Big(\frac{x}{1-z}, \frac{y}{1-z}\Big). \tag{7}$$

Σ is a conformal diffeomorphism which maps circles in S^2 to lines and circles in $\boldsymbol{R}^2 \cup \infty$. The inverse is:

$$\Sigma^{-1}(x,y) = \Big(\frac{2x}{1+x^2+y^2}, \frac{2y}{1+x^2+y^2}, 1 - \frac{2}{1+x^2+y^2}\Big). \tag{8}$$

We will use the convention that any object S in $\boldsymbol{R}^2 \cup \infty$ corresponds to

$$\widehat{S} = \Sigma^{-1}(S) \tag{9}$$

in S^2. Given our configuration $\widehat{p}_0, ..., \widehat{p}_4$ as in the previous section, we have points $p_0, ..., p_3 \in \boldsymbol{R}^2$ and $p_4 = \infty$. Thus, the points $p_0, ..., p_3$ determine the configation.

Lemma 2.2 *If p_0, p_1, p_2, p_3 correspond to a normalized minimizer for G_k with $k \in \{3,4,5,6\}$, then $|p_0| < 4$ and $|p_i| < 3/2$ for $i = 1,2,3$.*

Proof: We compute that

$$\|\Sigma^{-1}(4,0) - (0,0,1)\| < 1/2, \qquad \|\Sigma^{-1}(3/2,0) - (0,0,1)\| = 4/\sqrt{13}. \tag{10}$$

By symmetry and monotonicity, the first equation holds for any p with $|p| \geq 4$ and the second equation holds for any p with $|p| \geq 3/2$. Lemma 2.1 finishes the proof. ♠

By symmetry we can rotate the picture so that p_0 lies in the x-axis, and has coordinate $(p_{01}, 0)$, with $p_{01} \in (0,4)$. The remaining points must lie in the disk of radius $3/2$ about 0. However, we find it convenient to confine p_1, p_2, p_3 only to the square $[-2,2]^2$ at first. We set our moduli space to be the product

$$\square = [0,4] \times \Big([-2,2]^2\Big)^3 = [0,4] \times [-2,2]^6. \tag{11}$$

Conveniently, $\square$ is a cube of sidelength 4 in $\boldsymbol{R}^7$. We just have to search through this big cube and eliminate everything which is not the TBP!

2.3 The TBP Configurations

The TBP has two kinds of points, the two at the poles and the three at the equator. When ∞ is a polar point, the points p_0, p_1, p_2, p_3 are, after suitable permutation,

$$(1,0), \qquad (-1/2, -\sqrt{3}/2), \qquad (0,0), \qquad (-1/2, +\sqrt{3}/2). \tag{12}$$

We call this the *polar configuration*. When ∞ is an equatorial point, the points p_0, p_1, p_2, p_3 are, after suitable permutation,

$$(1,0), \qquad (0, -1/\sqrt{3}), \qquad (-1,0), \qquad 1/\sqrt{3}). \tag{13}$$

We call this the *equatorial configuration*. We prefer the polar configuration, and we will use a trick to avoid ever having to search very near the equatorial configuration.

We can visualize the two configurations together in relation to the regular 6-sided star. The black points are part of the polar configuration and the white points are part of the equatorial configuration. The grey point belongs to both configurations. The points represented by little squares are polar and the points represented by little disks are equatorial. The beautiful pattern made by these two configurations is not part of our proof, but it is nice to contemplate.

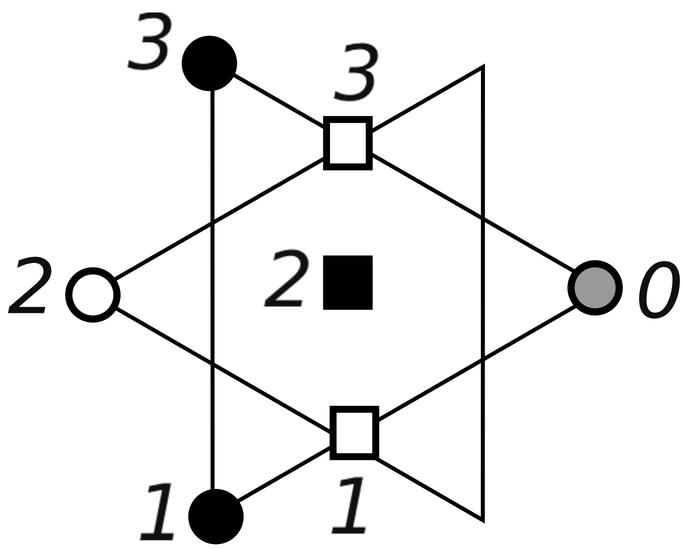

Figure 2.1: Polar and equatorial versions of the TBP.

2.4 Dyadic Blocks

In 1 dimension, the *dyadic subdivision* of a line segment is the union of the two segments obtained by cutting it in half. In 2 dimensions, the *dyadic subdivision* of a square is the union of the 4 quarters that result in cutting the the square in half along both directions. See Figure 2.2.

Figure 2.2: Dyadic subdivision

We say that a *dyadic segment* is any segment obtained from $[0, 4]$ by applying dyadic subdivision recursively. We say that a *dyadic square* is any square obtained from $[-2, 2]^2$ by applying dyadic subdivision recursively. We count $[0, 4]$ as a dyadic segment and $[-2, 2]^2$ as a dyadic square.

Hat and Hull Notation: Since we are going to be switching back and forth between the picture on the sphere and the picture in $\boldsymbol{R}^2$, we want to be clear about when we are talking about solid bodies, so to speak, and when we are talking about finite sets of points. We let $\langle X\rangle$ denote the convex hull of any Euclidean subset. Thus, we think of a dyadic square Q as the set of its 4 vertices and we think of $\langle Q\rangle$ as the solid square having Q as its vertex set. Combining this with our notation for stereographic projection, we get the following notation:

- $\widehat{Q}$ is a set of 4 co-circular points on S^2.
- $\langle\widehat{Q}\rangle$ is a convex quadrilateral whose vertices are $\widehat{Q}$.
- $\widehat{\langle Q\rangle}$ is a "spherical patch" on S^2, bounded by 4 circular arcs.

We will use this notation throughout the paper.

Good Squares: A dyadic square is *good* if it is contained in $[-3/2, 3/2]^2$ and has side length at most $1/2$. Note that a good dyadic square cannot cross the coordinate axes. The only dyadic square which crosses the coordinate axes is $[-2.2]^2$, and this square is not good. Our computer program will only do spherical geometry calculations on good squares.

Dyadic Blocks: We define a *dyadic block* to be a 4-tuple (Q_0, Q_1, Q_2, Q_3), where Q_0 is a dyadic segment and Q_i is a dyadic square for $j = 1, 2, 3$. We say that a block is *good* if each of its 3 component squares is good. By Lemma 2.2, any energy minimizer for G_k is contained in a good block. Our algorithm in §5 quickly chops up the blocks in $\square$ so that only good ones are considered.

The product

$$\langle B \rangle = \langle Q_0 \rangle \times \langle Q_1 \rangle \times \langle Q_2 \rangle \times \langle Q_3 \rangle \tag{14}$$

is a rectangular solid in the configuration space $\square$. On the other hand, the product

$$B = Q_0 \times Q_1 \times Q_2 \times Q_3 \tag{15}$$

is the collection of 128 vertices of $\langle B \rangle$. We call these the *vertex configurations* of the block.

Definition: We say that a configuration p_0, p_1, p_2, p_3 is *in* the block B if $p_i \in \langle Q_i \rangle$ for $i = 0, 1, 2, 3$. In other words, the point in $\square$ representing our configuration is contained in $\langle B \rangle$. Sometimes we will say that this configuration is *associated to* the block.

Sudvidision of Blocks: There are 4 obvious subdivision operations we can perform on a block.

- The operation S_0 divides B into the two blocks (Q_{00}, Q_1, Q_2, Q_3) and (Q_{01}, Q_1, Q_2, Q_3). Here (Q_{00}, Q_{01}) is the dyadic subdivision of Q_0.
- the operation S_1 divides B into the 4 blocks (Q_0, Q_{1ab}, Q_2, Q_3), where $(Q_{100}, Q_{101}, Q_{110}, Q_{111})$ is the dyadic subdivision of Q_1.

The operations S_2 and S_3 are similar to S_1.

The set of dyadic blocks has a natural poset structure to it, and basically our algorithm does a depth-first-search through this poset, eliminating solid blocks either due to symmetry considerations or due to energy considerations. The short chapter §5 describes the algorithm precisely.

2.5 Relevant Configurations

We write $p_i = (p_{i1}, p_{i2})$. Call $(p_0, p_1, p_2, p_3) \in \square$ *relevant* if all 5 of these inequalities hold:

$$p_{01} \geq 1, \quad p_{12} \leq 0, \quad p_{22} \geq 0, \quad p_{22} \leq p_{32}, \quad p_{21} \geq -1/2. \tag{16}$$

and otherwise *irrelevant*. The polar TBP is relevant but the equatorial TBP is irrelevant.

Lemma 2.3 *Relative to any monotone decreasing energy potential, some minimal configuration is relevant.*

Proof: We normalize so that $\widehat{p}_0$ and $\widehat{p}_4$ are the closest pair of points. If Inequality 1 fails then $\widehat{p}_0$ and $\widehat{p}_4$ are more than $\sqrt{2}$ apart, and hence all the points are. But it is impossible to place 5 points on S^2 such that every two of them are more than $\sqrt{2}$ apart. See [**S1**] for a proof.

If p_1, p_2, p_3 lie all inside (or all outside) the upper half plane H then all 5 corresponding points in S^2 lie in the hemisphere $\widehat{H}$ and $\widehat{p}_1, \widehat{p}_2, \widehat{p}_3$ lie in the interior of $\widehat{H}$. But then we can get a configuration of lower energy by replacing $\widehat{p}_1$ with its reflection in $\partial\widehat{H}$. So, reflecting in the x-axis and relabeling if necessary we can assume that p_1 does not lie in the interior of H and p_2, p_3 lie in H. This gives Inequalities 2,3,4.

If Inequality 5 fails, we apply the inversion I_C in the circle C which is centered at p_0 and has the property that $\Sigma^{-1}(C)$ is a great circle. I_C corresponds to the isometric reflection in $\Sigma^{-1}(C)$ which swaps $\widehat{p}_0$ and $\widehat{p}_4$. Let $q_j = I_C(p_j)$ for $j = 1, 2, 3$. Let Δ be the disk bounded by C. Note that I_C fixes C and swaps Δ with the complment of Δ. Let X denote the x-axis. If $p_0 = (1, 0)$, then $C \cap X = \{1 \pm \sqrt{2}\}$. This is the extreme case. When $p_{01} > 1$ both points of $C \cap X$ lies to the right of $(1 - \sqrt{2}, 0)$, which in turn lies to the right of $(-1/2, 0)$. But then $p_2 \not\in \Delta$. Hence $q_2 \in \Delta$. This gives $q_{21} > -1/2$.

Since I_C preserves all the rays through p_0, we still have $q_{12} \leq 0$ and $q_{22}, q_{32} \geq 0$. If $q_3 \in \Delta$ then $q_{31} > -1/2$. If needed, we switch q_2 and q_3 to guarantee that $q_{22} \leq q_{32}$. If $q_3 \not\in \Delta$ then $p_{31} > 1 - \sqrt{2} > p_{21}$ and hence the ray ρ_3 joining p_0 to p_3 lies above the ray ρ_2 joining p_0 to p_2. But then $q_3 \in \rho_3 - \Delta$ and $q_2 \in \rho_2 \cap \Delta$. This forces $q_{22} < q_{32}$. So in all cases, we can retain Inequalities 2,3,4. Since $q_0 = p_0$ we retain Inequality 1. ♠

Call a block *irrelevant* if every configuration in the interior of the block is irrelevant. Call a block *relevant* if it is not irrelevant. Now we give a criterion for a block to be irrelevant.

Given a box Q_j, let $\overline{Q}_{jk}$ and $\underline{Q}_{jk}$ denote the maximum and minimum kth coordinate of a point in Q_j. The good block $B = (Q_0, Q_1, Q_2, Q_3)$ is irrelevant provided that at least one of the following holds.

$$\overline{Q}_{01} \leq 1, \quad \underline{Q}_{12} \geq 0, \quad \overline{Q}_{22} \leq 0, \quad \underline{Q}_{22} \geq \overline{Q}_{32}, \quad \overline{Q}_{21} \leq -1/2. \tag{17}$$

Notice that these 5 inequalities parallel the ones given above for individual configurations.

Every relevant configuration in the boundary of an irrelevant block is also in the boundary of a relevant block. So, to prove Theorem 1.3, we can ignore the irrelevant blocks.

2.6 A Technical Result about Dyadic Boxes

The following result will be useful for the estimates in §3.4. We state it in more generality than we need, to show what hypotheses are required, but we note that good dyadic squares satisfy the hypotheses. We only care about the result for good dyadic squares and for good dyadic segments. In the case of good dyadic segments, the lemma is obviously true.

Lemma 2.4 *Let $\widehat{Q}$ be a rectangle whose sides are parallel to the coordinate axes and do not cross the coordinate axes. Then the points of $\widehat{\langle Q \rangle}$ closest to* $(0,0,1)$ *and farthest from* $(0,0,1)$ *are both vertices.*

Proof: Put the metric on $\boldsymbol{R}^2 \cup \infty$ which makes stereographic projection an isometry. By symmetry, the metric balls about ∞ are the complements of disks centered at 0. The smallest disk centered at 0 and containing $\langle Q \rangle$ must have a vertex on its boundary. Likewise, the largest disk centered at 0 and disjoint from the interior of $\langle Q \rangle$ must have a vertex in its boundary. This latter result uses the fact that $\langle Q \rangle$ does not cross the coordinate axes. These statements are equivalent to the statement of the lemma. ♠

2.7 Very Near the TBP

Our calculations will depend on a a pair (S, ϵ_0), both powers of two. For G_3, G_4, G_5, G_6 we use $S = 2^{25}$ and $\epsilon_0 = 2^{-15}$. For $G_{10}^{\#}$ we use $S = 2^{30}$ and $\epsilon_0 = 2^{-18}$. Our discussion below would work for other similar choices.

Define the *in-radius* of a cube to be half its side length. Let P_0 denote the configuration representing the totally normalized polar TBP and let B_0 denote the cube centered at P_0 ahd having in-radius ϵ_0. Note that B_0 is not a dyadic block. This does not bother us. Here we give a sufficient condition for $B \subset B_0$, where B is some dyadic block.

Let $a = \sqrt{3}/2$. For each choice of S we compute a value a^* such that $Sa^* \in \boldsymbol{Z}$ and $|a - a^*| < 1/S$. There are two such choices, namely

$$\frac{\text{floor}(Sa)}{S}, \qquad \frac{\text{floor}(Sa+1)}{S}. \tag{18}$$

In practice, our program sets a^* to be the first of these two numbers, but we want to state things in a symmetric way that works for either choice.

We define $B_0' = Q_0' \times Q_1' \times Q_2' \times Q_3'$ where

$$Q_0' = [1-\epsilon_0, 1+\epsilon_0],$$

$$Q_1' = [-1/2-\epsilon_0, -1/2+\epsilon_0] \times [-S^{-1} - a^* - \epsilon_0, S^{-1} - a^* + \epsilon_0]$$

$$Q_2' = [-\epsilon_0, \epsilon_0]^2$$

$$Q_3' = [1/2-\epsilon_0, 1/2+\epsilon_0] \times \times [-S^{-1} - a^* - \epsilon_0, S^{-1} - a^* + \epsilon_0]$$

By construction, we have $B_0' \subset B_0$. Given a block $B = Q_0 \times Q_1 \times Q_2 \times Q_3$, the condition $Q_i \subset Q_i'$ for all i implies that $B \subset B_0$. We will see in §9 that this is an exact integer calculation for us.

3 Spherical Geometry Estimates

3.1 Overview

In this chapter we define the basic quantities that go into the Energy Theorem, Theorem 4.1. We will persistently use the hat and hull notation defined in §2.4. Thus:

- When Q is a dyadic square, $\langle\widehat{Q}\rangle$ is a convex quadrilateral in space whose vertices are 4 co-circular points on S^2, and $\widehat{\langle Q\rangle}$ is a subset of S^2 bounded by 4 circular arcs.
- When Q is a dyadic segment, $\langle\widehat{Q}\rangle$ is a segment whose endpoints are on S^2, and $\widehat{\langle Q\rangle}$ is an arc of a great circle.

Here is a summary of the quantities we will define in this chapter. The first three quantities are not rational functions of the inputs, but our estimates only use the squares of these quantities.

Hull Diameter: $d(Q)$ will be the diameter $\langle\widehat{Q}\rangle$.

Edge Length: $d_1(Q)$ will be the length of the longest edge of $\langle\widehat{Q}\rangle$.

Circular Measure: Let $D_Q \subset \boldsymbol{R}^2$ denote the disk containing Q in its boundary. $d_2(Q)$ is the diameter of $\widehat{D}_Q$.

Hull Separation Constant: $\delta(Q)$ will be a constant such that every point in $\widehat{\langle Q\rangle}$ is within $\delta(Q)$ of a point of $\langle\widehat{Q}\rangle$. This quantity is a rational function of the coordinates of Q.

Dot Product Bounds: We will introduce a (finitely computable, rational) quantity $(Q \cdot Q')_{\max}$ which has the property that

$$V \cdot V' \leq (Q \cdot Q')_{\max}$$

for all $V \in \widehat{\langle Q\rangle} \cup \langle\widehat{Q}\rangle$ and $V' \in \widehat{\langle Q'\rangle} \cup \langle\widehat{Q}'\rangle$.

3.2 Some Results about Circles

Here we prove a few geometric facts about circles and stereographic projection.

Lemma 3.1 *Let Q be a good dyadic square or a dyadic segment. The circular arcs bounding $\widehat{\langle Q\rangle}$ lie in circles having diameter at least* 1.

Proof: Let Σ denote stereographic projection. In the dyadic segment case, $\widehat{\langle Q\rangle}$ lies in a great circle. In the good dyadic square case, each edge of $\langle Q\rangle$ lies on a line L which contains a point p at most $3/2$ from the origin. But $\Sigma^{-1}(p)$ is at least 1 unit from $(0,0,1)$. Hence $\Sigma^{-1}(L)$, which limits on $(0,0,1)$ and contains $\Sigma^{-1}(p)$, has diameter at least 1. The set $\Sigma^{-1}(L\cup\infty)$ is precisely the circle extending the relevant edge of $\widehat{\langle Q\rangle}$ ♠

Let $D\subset \boldsymbol{R}^2$ be a disk of radius $r\leq R$ centered at a point which is R units from the origin. Let $\widehat{D}$ denote the corresponding disk on S^2. We consider $\widehat{D}$ as a subset of $\boldsymbol{R}^3$ and compute its diameter in with respect to the Euclidean metric on $\boldsymbol{R}^3$.

Lemma 3.2

$$\text{diam}^2(\widehat{D}) = \frac{16r^2}{1+2r^2+2R^2+(R^2-r^2)^2}. \tag{19}$$

Proof: By symmetry it suffices to consider the case when the center of D is $(R,0)$. The diameter is then achieved by the two points $V=\Sigma^{-1}(R-r,0)$ and $W=\Sigma^{-1}(R+r,0)$. The formula comes from computing $\|V-W\|^2$ and simplifying. ♠

We introduce the functions

$$\chi(D,d)=\frac{d^2}{4D}+\frac{d^4}{2D^3} \qquad \chi^*(D,d)=\frac{1}{2}(D-\sqrt{D^2-d^2}). \tag{20}$$

The second of these is a function closely related to the geometry of circles. This is the function we would use if we had an ideal computing machine. However, since we want our estimates to all be rational functions of the inputs, we will use the first function. We first prove an approximation lemma and then we get to the main point.

Lemma 3.3 *If* $0 \leq d \leq D$ *then* $\chi^*(D,d) \leq \chi(D,d)$.

Proof: If we replace (d,D) by (rd,rD) then both sides scale up by r. Thanks to this homogeneity, it suffices to prove the result when $D=1$. We have $\chi(1,1) = 3/4 > 1/2 = \chi^*(1,1)$. So, if suffices to prove that the equation

$$\chi(1,d) - \chi^*(1,d) = \frac{d^2}{4} + \frac{d^4}{2} - \frac{1}{2}(1 - \sqrt{1-d^2})$$

has no real solutions in $[0,1]$ besides $d=0$. Consider a solution to this equation. Rearranging the equation, we get $A=B$ where

$$A = -\frac{1}{2}\sqrt{1-d^2}, \qquad B = \frac{d^2}{4} + \frac{d^4}{2} - 1/2.$$

An exercise in calculus shows that the only roots of $A^2 - B^2$ in $[0,1]$ are 0 and $\sqrt{1/2(\sqrt{8}-1)} > .95$. On the other hand $A < 0$ and $B > 0$ on $[.95, 1]$. ♠

Now we get to the key result. This result holds in any dimension, but we will apply it once to the 2-sphere, and once to circles contained in suitable planes in $\boldsymbol{R}^3$.

Lemma 3.4 *Let* Γ *be a round sphere of diameter* D*, contained in some Euclidean space. Let* B *be the ball bounded by* Γ*. Let* Π *be a hyperplane which intersects* B *but does not contain the center of* B*. Let* $\gamma = \Pi \cap B$ *and let* γ^* *be the smaller of the two spherical caps on* Γ *bounded by* $\Pi \cap \Gamma$*. Let* $p^* \in \gamma^*$ *be a point. Let* $p \in \gamma$ *be the point so the line* $\overline{pp^*}$ *contains the center of* B*. Then* $\|p - p^*\| \leq \chi(D,d)$.

Proof: The given distance is maximized when p^* is the center of γ^* and p is the center of γ. In this case it suffices by symmetry to consider the situation in $\boldsymbol{R}^2$, where $\overline{pp^*}$ is the perpendicular bisector of γ. Setting $x = \|p - p^*\|$, we have

$$x(D-x) = (d/2)^2. \tag{21}$$

This equation comes from a well-known theorem from high school geometry concerning the lengths of crossing chords inside a circle. When we solve Equation 21 for x, we see that $x = \chi^*(D,d)$. The previous lemma finishes the proof. ♠

3.3 The Hull Approximation Lemma

Circular Measure: When Q is a dyadic square or segment, we define

$$d_2(Q) = \mathrm{diam}(\widehat{D}_Q), \tag{22}$$

Where $D_Q \subset \boldsymbol{R}^2$ is such that $Q \subset D_Q$. So $d_2(Q)$ is the diameter of the small spherical cap which contains $\widehat{Q}$ in its boundary. Note that $\widehat{\langle Q \rangle} \subset \widehat{D}_Q$ by construction. We call $d_2(Q)$ the *circular measure* of Q.

Hull Separation Constant: Recall that $d_1(Q)$ is the maximum side length of $\langle Q \rangle$. When When Q is a dyadic segment, we define $\delta(Q) = \chi(2, d_2)$. When Q is a good dyadic square, We define

$$\delta(Q) = \max\Big(\chi(1, d_1), \chi(2, d_2)\Big). \tag{23}$$

This definition makes sense, because $d_1(Q) \leq 1$ and $d_2(Q) \leq \sqrt{2} < 2$. The point here is that Σ^{-1} is 2-Lipschitz and Q has side length at most $1/2$. We call $\delta(Q)$ the *Hull approximation constant* of Q.

Lemma 3.5 (Hull Approximation) *Let Q be a dyadic segment or a good dyadic square. Every point of the spherical patch $\widehat{\langle Q \rangle}$ is within $\delta(Q)$ of a point of the convex quadrilateral $\langle \widehat{Q} \rangle$.*

Proof: Suppose first that Q is a dyadic segment. $\widehat{\langle Q \rangle}$ is the short arc of a great circle sharing endpoints with $\langle \widehat{Q} \rangle$, a chord of length d_2. By Lemma 3.4 each point of on the circular arc is within $\chi(2, d_2)$ of a point on the chord.

Now suppose that Q is a good dyadic square. Let O be the origin in $\boldsymbol{R}^3$. Let $H \subset S^2$ denote the set of points such that the segment $Op^* \in H$ intesects $\langle \widehat{Q} \rangle$ in a point p. Here H is the itersection with the cone over $\langle \widehat{Q} \rangle$ with S^2.

Case 1: Let $p^* \in \widehat{\langle Q \rangle} \cap H$. Let $p \in \langle Q \rangle$ be such that the segment Op^* contains p. Let B be the unit ball. Let Π be the plane containing $\widehat{Q}$. Note that $\Pi \cap S^2$ bounds the spherical $\widehat{D}_Q$ which contains $\widehat{Q}$ in its boundary. Let

$$\Gamma = S^2, \qquad \gamma^* = \widehat{D}_Q, \qquad \gamma = \Pi \cap B.$$

The diameter of Γ is $D = 2$. Lemma 3.4 now tells us that $\|p - p^*\| \leq \chi(2, d_2)$.

Case 2: Let $p^* \in \widehat{\langle Q \rangle} - H$. The sets $\widehat{\langle Q \rangle}$ and H are both bounded by 4 circular arcs which have the same vertices. H is bounded by arcs of great circles and $\widehat{\langle Q \rangle}$ is bounded by arcs of circles having diameter at least 1. The point p^* lies between an edge-arc α_1 of H and an edge-arc α_2 of $\widehat{\langle Q \rangle}$ which share both endpoints. Let γ be the line segment joining these endpoints. The diameter of γ is at most d_1.

Call an arc of a circle *nice* if it is contained in a semicircle, and if the circle containing it has diameter at least 1. The arcs α_1 and α_2 are both nice. We can foliate the region between α_1 and α_2 by arcs of circles. These circles are all contained in the intersection of S^2 with planes which contain γ. Call this foliation $\mathcal{F}$. We get this foliation by rotating the planes around their common axis, which is the line through γ.

Say that an $\mathcal{F}$-*circle* is a circle containing an arc of $\mathcal{F}$. Let e be the edge of $\langle Q \rangle$ corresponding to γ. Call (e, Q) *normal* if γ is never the diameter of an $\mathcal{F}$-circle. If (e, Q) is normal, then the diameters of the $\mathcal{F}$-circles interpolate monotonically between the diameter of α_1 and the diameter of α_2. Hence, all $\mathcal{F}$-circles have diameter at least 1. At the same time, if (e, Q) is normal, then all arcs of $\mathcal{F}$ are contained in semicircles, by continuity. In short, if (e, Q) is normal, then all arcs of $\mathcal{F}$ are nice. Assuming that (e, Q) is normal, let γ^* be the arc in $\mathcal{F}$ which contains p^*. Let $p \in \Gamma$ be such that the line pp^* contains the center of the circle Γ containing γ^*. Since γ^* is nice, Lemma 3.4 says that $\|p - p^*\| \leq \chi(D, d_1) \leq \chi(1, d_1)$.

To finish the proof, we just have to show that (e, Q) is normal. We enlarge the set of possible pairs we consider, by allowing rectangles in $[-3/2, 3/2]^2$ having sides parallel to the coordinate axes and maximum side length $1/2$. The same arguments as above, Lemma 3.1 and the 2-Lipschitz nature of Σ^{-1}, show that α_1' and α_2' are still nice for any such pair (e', Q').

If e' is the long side of a $1/2 \times 10^{-100}$ rectangle $\langle Q' \rangle$ contained in the 10^{-100}-neighborhood of the coordinate axes, then (e', Q') is normal: The arc α_1' is very nearly the arc of a great circle and the angle between α_1' and α_2' is very small, so all arcs $\mathcal{F}'$ are all nearly arcs of great circles. If some choice (e, Q) is not normal, then by continuity, there is a choice (e'', Q'') in which γ'' is the diameter of one of the two boundary arcs of $\mathcal{F}''$. There is no other way to switch from normal to not normal. But this is absurd because the boundary arcs, α_1'' and α_2'', are nice. ♠

3.4 Dot Product Estimates

Let Q be a dyadic segment or a good dyadic square. Let δ be the hull separation constant of Q. Let $\{q_i\}$ be the points of Q. We make all the same definitions for a second dyadic square Q'. We define

$$(Q \cdot Q')_{\max} = \max_{i,j}(\widehat{q}_i \cdot \widehat{q}'_j) + \delta + \delta' + \delta\delta'. \tag{24}$$

$$(Q \cdot \{\infty\})_{\max} = \max_{i} \ \widehat{q}_i \cdot (0,0,1) \tag{25}$$

Connectors: We say that a *connector* is a line segment connecting a point on $\widehat{\langle Q \rangle}$ to any of its closest points in $\langle \widehat{Q} \rangle$. We let $\Omega(Q)$ denote the set of connectors defined relative to Q. By the Hull Approximation Lemma, each $V \in \Omega(Q)$ has the form $W + \delta U$ where $W \in \langle \widehat{Q} \rangle$ and $\|U\| \leq 1$.

Lemma 3.6 *$V \cdot V' \leq (Q \cdot Q')_{\max}$ for all $(V, V') \in \Omega(Q) \times \Omega(Q')$*

Proof: Suppose $V \in \langle \widehat{Q} \rangle$ and $V' \in \langle \widehat{Q}' \rangle$. Since the dot product is bilinear, the restriction of the dot product to the convex polyhedral set $\langle \widehat{Q} \rangle \times \langle \widehat{Q}' \rangle$ takes on its extrema at vertices. Hence $V \cdot V' \leq \max_{i,j} q_i \cdot q'_i$. In this case, we get the desired inequality whether or not $Q' = \{\infty\}$.

Suppose $Q' \neq \{\infty\}$ and V, V' are arbitrary. We use the decomposition mentioned above:

$$V = W + \delta U, \qquad V' = W' + \delta' U, \qquad W \in \langle \widehat{Q} \rangle, \qquad W' \in \langle \widehat{Q}' \rangle. \tag{26}$$

But then, by the Cauchy-Schwarz inequality,

$$|(V \cdot V') - (W \cdot W')| = |V \cdot \delta' U' + V' \cdot \delta U + \delta U \cdot \delta' U'| \leq \delta + \delta' + \delta\delta'.$$

The lemma now follows immediately from this equation, the previous case applied to W, W', and the triangle inequality.

Suppose that $Q' = \{\infty\}$. We already know the result when $V \in \langle \widehat{Q} \rangle$. When $V \in \widehat{\langle Q \rangle}$ we get the better bound above from Lemma 2.4 and from the fact that the dot product $V \cdot (0,0,1)$ varies monotonically with the distance from V to $(0,0,1)$ and *vice versa*. Now we know the result whenever V is the endpoint of a connector. By the linearity of the dot product, the result holds also when V is an interior point of a connector. ♠

4 The Energy Theorem

4.1 Main Result

We think of the energy potential $G = G_k$ as being a function on $(\boldsymbol{R}^2 \times \infty)^2$, via the identification $p \leftrightarrow \widehat{p}$. Our Energy Theorem below works for any real $k \geq 2$.

Let $\mathcal{Q}$ denote the set of dyadic squares $[-2,2]^2$ together with the dyadic segments in $[0,4]$, together with $\{\infty\}$. When $Q = \{\infty\}$ the constants associated to Q, namely the hull separation constant and the convex hull diameter, are 0.

Now we are going to define a function $\epsilon : \mathcal{Q} \times \mathcal{Q} \to [0,\infty)$. First of all, for notational convenience we set $\epsilon(Q,Q) = 0$ for all Q, When $Q, Q' \in \mathcal{Q}$ are unequal, we define

$$\epsilon(Q,Q') = \frac{1}{2}k(k-1)T^{k-2}d^2 + 2kT^{k-1}\delta \tag{27}$$

Here

- d is the diameter of $\widehat{Q}$.
- $\delta = \delta(Q)$ is the hull approximation constant for Q. See §3.3.
- $T = T(Q,Q') = 2 + 2(Q \cdot Q')_{\max}$. See §3.4.

This is a rational function in the coordinates of Q and Q'. The quantities d^2 and δ are essentially quadratic in the side-lengths of Q and Q'. Note that $\epsilon(\{\infty\}, Q') = 0$ but $\epsilon(Q, \{\infty\})$ is nonzero when $Q \neq \{\infty\}$.

Let $B = (Q_0, Q_1, Q_2, Q_3)$. For notational convenience we set $Q_4 = \{\infty\}$. We define

$$\mathbf{ERR}(B) = \sum_{i=0}^{3} \sum_{j=0}^{4} \epsilon(Q_i, Q_j). \tag{28}$$

Theorem 4.1 (Energy)

$$\min_{v \in \langle B \rangle} \mathcal{E}_k(v) \geq \min_{v \in B} \mathcal{E}_k(v) - \mathbf{ERR}(B).$$

Corollary 4.2 *Suppose that B is a block such that*

$$\min_{v \in B} \mathcal{E}(v) - \mathbf{ERR}(B) > \mathcal{E}_k(\mathrm{TBP}). \tag{29}$$

Then all configurations in B have higher energy than the TBP.

4.2 The Subdivision Recommendation

We can write

$$\mathbf{ERR}(B) = \sum_{i=0}^{3} \mathbf{ERR}_i(B), \qquad \mathbf{ERR}_i(B) = \sum_{j=0}^{4} \epsilon(Q_i, Q_j). \tag{30}$$

We define the *subdivision recommendation* to be the index $i \in \{0,1,2,3\}$ for which $\mathbf{ERR}_i(B)$ is maximal. In the extremely unlike event that two of these terms coincide, we pick the smaller of the two indices to break the tie. The subdivision recommendation feeds into the algorithm described in §5.

The rest of the chapter is devoted to proving Theorem 4.1. The next chapter explains how we use Theorem 4.1 in our proof.

4.3 A Polynomial Inequality

Theorem 4.1 derives from the case $M = 4$ of the following inequality.

Lemma 4.3 *Let $M \geq 2$ and $k \geq 2$. Suppose*

- $0 \leq x_1 \leq ... \leq x_M$.
- $\sum_{i=1}^{M} \lambda_i = 1$ *and* $\lambda_i \geq 0$ *for all* i.

Then

$$\sum_{i=1}^{M} \lambda_i x_i^k - \Big(\sum_{I=1}^{M} \lambda_i x_i\Big)^k \leq \frac{1}{8} k(k-1) x_M^{k-2}\ (x_M - x_1)^2. \tag{31}$$

I discovered Lemma 4.3 experimentally

Lemma 4.4 *The case $M = 2$ of Lemma 4.3 implies the rest.*

Proof: Suppose that $M \geq 3$. We have one degree of freedom when we keep $\sum \lambda_i x_i$ constant and try to vary $\{\lambda_j\}$ so as to maximize the left hand side of the inequality. The right hand side does not change when we do this, and the left hand side varies linearly. Hence, the left hand size is maximized when $\lambda_i = 0$ for some i. But then any counterexample to the lemma for $M \geq 3$ gives rise to a counter example for $M - 1$. ♠

In the case $M = 2$, we set $a = \lambda_1$. Both sides of the inequality in Lemma 4.3 are homogeneous of degree k, so it suffices to consider the case when $x_2 = 1$. We set $x = x_1$. The inequality of is then $f(x) \leq g(x)$, where

$$f(x) = (ax^k + 1 - a) - (ax + 1 - a)^k; \qquad g(x) = \frac{1}{8}k(k-1)(1-x)^2. \quad (32)$$

This is supposed to hold for all $a, x \in [0, 1]$.

The following argument is due to C. McMullen, who figured it out after I told him about the inequality.

Lemma 4.5 *Equation 32 holds for all $a, x \in [0, 1]$ and all $k \geq 2$.*

Proof: Equation 32. We think of f as a function of x, with a held fixed. Since $f(1) = g(1) = 1$, it suffices to prove that $f'(x) \geq g'(x)$ on $[0, 1]$. Define

$$\phi(x) = akx^{k-1}, \qquad b = (1-a)(1-x). \quad (33)$$

We have

$$-f'(x) = \phi(x + b) - \phi(x). \quad (34)$$

Both x and $x + b$ lie in $[0, 1]$. So, by the mean value theorem there is some $y \in [0, 1]$ so that

$$\frac{\phi(x+b) - \phi(x)}{b} = \phi'(y) = ak(k-1)y^{k-2}. \quad (35)$$

Hence

$$-f'(x) = b\phi'(y) = a(1-a)k(k-1)(1-x)y^{k-2} \quad (36)$$

But $a(1-a) \in [0, 1/4]$ and $y^{k-2} \in [0, 1]$. Hence

$$-f'(x) \leq \frac{1}{4}k(k-1)(1-x) = -g'(x). \quad (37)$$

Hence $f'(x) \geq g'(x)$ for all $x \in [0, 1]$. ♠

Remark: Lemma 4.3 has the following motivation. The idea behind the Energy Theorem is that we want to measure the deviation of the energy function from being linear, and for this we would like a quadratic estimate. Since our energy G_k involves high powers, we want to estimate these high powers by quadratic terms.

4.4 The Local Energy Lemma

Let $Q = \{q_1, q_2, q_3, q_4\}$ be the vertex set of $Q \in \mathcal{Q}$. We allow for the degenerate case that Q is a line segment or $\{\infty\}$. In this case we just list the vertices multiple times, for notational convenience.

Note that every point in the convex quadrilateral $\langle \widehat{Q} \rangle$ is a convex average of the vertices. For each $z \in \langle Q \rangle$, there is a some point $z^* \in \langle \widehat{Q} \rangle$ which is as close as possible to $\widehat{z} \in \widehat{\langle Q \rangle}$. There are constants $\lambda_i(z)$ such that

$$z^* = \sum_{i=1}^{4} \lambda_i(z)\ \widehat{q}_j, \qquad \sum_{i=1}^{4} \lambda_i(z) = 1. \tag{38}$$

We think of the 4 functions $\{\lambda_i\}$ as a partition of unity on $\langle Q \rangle$. The choices above might not be unique, but we make such choices once and for all for each Q. We call the assignment $Q \to \{\lambda_i\}$ the *stereographic weighting system.*

Lemma 4.6 (Local Energy) *Let ϵ be the function defined in the Energy Theorem. Let Q, Q' be distinct members of $\mathcal{Q}$. Given any $z \in Q$ and $z' \in Q'$,*

$$G(z, z') \geq \sum_{i=1}^{4} \lambda_i(z) G(q_i, z') - \epsilon(Q, Q'). \tag{39}$$

Proof: For notational convenience, we set $w = z'$. Let

$$X = (2 + 2z^* \cdot \widehat{w})^k. \tag{40}$$

The Local Energy Lemma follows from adding these two inequalities:

$$\sum_{i=1}^{4} \lambda_i G(q_i, w) - X \leq \frac{1}{2} k(k-1) T^{k-2} d^2 \tag{41}$$

$$X - G(z, w) \leq 2kT^{k-1}\delta. \tag{42}$$

We will establish these inequalities in turn.

Let q_1, q_2, q_3, q_4 be the vertices of Q. Let $\lambda_i = \lambda_i(z)$. We set

$$x_i = 4 - \|\widehat{q}_i - \widehat{w}\|^2 = 2 + 2\widehat{q}_i \cdot \widehat{w}, \qquad i = 1, 2, 3, 4. \tag{43}$$

Note that $x_i \geq 0$ for all i. We order so that $x_1 \leq x_2 \leq x_3 \leq x_4$. We have

$$\sum_{i=1}^{4} \lambda_i(z) G(q_i, w) = \sum_{i=1}^{4} \lambda_i x_i^k, \tag{44}$$

$$X = (2 + 2\widehat{z}^* \cdot \widehat{w})^k = \Big(\sum_{i=1}^{4} \lambda_i (2 + \widehat{q}_i \cdot \widehat{w})\Big)^k = \Big(\sum_{i=1} \lambda_i x_i\Big)^k. \tag{45}$$

Combining Equation 44, Equation 45, and the case $M = 4$ of Lemma 4.3,

$$\sum_{i=1}^{4} \lambda_i G(q_i, w) - X = \sum_{i=1}^{4} \lambda_i x_i^k - \Big(\sum_{i=1}^{4} \lambda_i x_i\Big)^k \leq \frac{1}{8} k(k-1) x_4^{k-2} (x_4 - x_1)^2. \tag{46}$$

By Lemma 3.6, we have

$$x_4 = 2 + 2(\widehat{q}_4 \cdot \widehat{w}) \leq 2 + 2(Q \cdot Q')_{\max} = T. \tag{47}$$

Since d is the diameter of $\langle \widehat{Q} \rangle$ and $\widehat{w}$ is a unit vector,

$$x_4 - x_1 = 2\widehat{w} \cdot (\widehat{q}_4 - \widehat{q}_1) \leq 2\|\widehat{w}\| \|\widehat{q}_4 - \widehat{q}_1\| = 2\|\widehat{q}_4 - \widehat{q}_1\| \leq 2d. \tag{48}$$

Plugging Equations 47 and 48 into Equation 46, we get Equation 41.

Now we establish Equation 42. Let γ denote the unit speed line segment connecting $\widehat{z}$ to z^*. Note that the length L of γ is at most δ, by the Hull Approximation Lemma. Define

$$f(t) = \Big(2 + 2\widehat{w} \cdot \gamma(t)\Big)^k. \tag{49}$$

We have $f(0) = X$. Since $\widehat{w}$ and $\gamma(1) = \widehat{z}$ are unit vectors, $f(L) = G(z, w)$. Hence

$$X - G(z, w) = f(0) - f(L), \qquad L \leq \delta. \tag{50}$$

By the Chain Rule,

$$f'(t) = (2\widehat{w} \cdot \gamma'(t)) \times k\Big(2 + 2\widehat{w} \cdot \gamma(t)\Big)^{k-1}. \tag{51}$$

Note that $|2\widehat{w} \cdot \gamma'(t)| \leq 2$ because both vectors are unit vectors. Note that γ parametrizes one of the connectors from Lemma 3.6, so we have

$$|f'(t)| \leq 2k\Big(2 + 2\widehat{w} \cdot \gamma(t)\Big)^{k-1} \leq 2k\Big(2 + 2(Q \cdot Q')_{\max}\Big)^{k-1} = 2kT^{k-1}. \tag{52}$$

Equation 42 now follows from Equation 50, Equation 52, and integration. ♠

4.5 From Local to Global

Let ϵ be the function from the Energy Theorem. Let $B = (Q_0, ..., Q_N)$ be a list of $N+1$ elements of $\mathcal{Q}$. We care about the case $N = 4$ and $Q_4 = \{\infty\}$, but the added generality makes things clearer. Let $q_{i,1}, q_{i,2}, q_{i,3}, q_{i,4}$ be the vertices of Q_i. The vertices of $\langle B \rangle$ are indexed by a multi-index

$$I = (i_0, ..., i_n) \in \{1, 2, 3, 4\}^{N+1}.$$

Given such a multi-index, which amounts to a choice of vertex of $\langle B \rangle$, we define the energy of the corresponding vertex configuration:

$$\mathcal{E}(I) = \mathcal{E}(q_{0,i_0}, ..., q_{N,i_N}) \tag{53}$$

We will prove the following sharper result.

Theorem 4.7 *Let $z_0, ..., z_N \in \langle B \rangle$. Then*

$$\mathcal{E}(z_0, ..., z_N) \geq \sum_I \lambda_{i_0}(z_0) ... \lambda_{i_N}(z_N) \mathcal{E}(I) - \sum_{i=0}^{N} \sum_{j=0}^{N} \epsilon(Q_i, Q_j). \tag{54}$$

The sum is taken over all multi-indices.

Lemma 4.8 *Theorem 4.7 implies the Energy Theorem.*

Proof: Notice that

$$\sum_I \lambda_{i_0}(z_0) ... \lambda_{i_N}(z_N) = \prod_{j=0}^{N} \Big(\sum_{a=1}^{4} \lambda_a(z_j) \Big) = 1. \tag{55}$$

Therefore

$$\sum_I \lambda_{i_0}(z_0) ... \lambda_{i_N}(z_N) \mathcal{E}(I) \geq \min_{v \in B} \mathcal{E}(v), \tag{56}$$

because the sum on the left hand side is the convex average of vertex energies and the term on the right is the minimum of the vertex energies.

For any $(z_0, ..., z_N) \in \langle B \rangle$, we now know from Theorem 4.7 that

$$\mathcal{E}(z_0, ..., z_N) \geq \min_{v \in B} \mathcal{E}(v) - \sum_{i=0}^{N} \sum_{j=0}^{N} \epsilon(Q_i, Q_j).$$

When we take $N = 4$ and $Q_4 = \{\infty\}$, the second expression on the right hand side of this last equation is precisely $\mathbf{ERR}(B)$. This establishes the Energy Theorem. ♠

We now prove Theorem 4.7.

4.5.1 A Warmup Case

Consider the case when $N = 1$. Setting $\epsilon_{ij} = \epsilon(Q_i, Q_j)$, the Local Energy Lemma gives us

$$G(z_0, z_1) \geq \sum_{\alpha=1}^{4} \lambda_\alpha(z_0) G(q_{0\alpha}, z_1) - \epsilon_{01}. \tag{57}$$

$$G(q_{0\alpha}, z_1) \geq \sum_{\beta=1}^{4} \lambda_\beta(z_1) G(q_{1\beta}(z_1), q_{0\alpha}) - \epsilon_{10}. \tag{58}$$

Plugging the second equation into the first and using $\sum \lambda_\alpha(z_0) = 1$, we have

$$G(z_0, z_1) \geq \Big(\sum_{\alpha,\beta} \lambda_\alpha(z_0)\lambda_\beta(z_1) G(q_{0\alpha}, q_{1\beta})\Big) - (\epsilon_{01} + \epsilon_{10}). \tag{59}$$

This is precisely Equation 54 when $N = 1$.

4.5.2 The General Case

Now assume that $N \geq 2$. We rewrite Equation 59 as follows:

$$G(z_0, z_1) \geq \sum_A \lambda_{A_0}(z_0)\lambda_{A_1}(z_1)\ G(q_{0A_0}, q_{1A_1}) - (\epsilon_{01} + \epsilon_{10}). \tag{60}$$

The sum is taken over multi-indices A of length 2.

We also observe that

$$\sum_{I'} \lambda_{i_2}(z_2)...\lambda_{i_N}(z_N) = 1. \tag{61}$$

The sum is taken over all multi-indices $I' = (i_2, ..., i_N)$. Therefore, if A is held fixed, we have

$$\lambda_{A_0}(z_0)\lambda_{A_1}(z_1) = \sum_I{}' \lambda_{I_0}(z_0)...\lambda_{I_N}(z_N). \tag{62}$$

The sum is taken over all multi-indices of length $N + 1$ which have $I_0 = A_0$ and $I_1 = A_1$. Combining these equations, we have

$$G(z_0, z_1) \geq \sum_I \lambda_{I_0}(z_0)...\lambda_{I_N}(z_N) G(q_{0I_0}, q_{1I_1}) - (\epsilon_{01} + \epsilon_{10}). \tag{63}$$

The same argument works for other pairs of indices, giving

$$G(z_i, z_j) \geq \sum_I \lambda_{I_0}(z_0)...\lambda_{I_N}(z_N) G(q_{iI_i}, q_{jI_j}) - (\epsilon_{ij} + \epsilon_{ji}). \tag{64}$$

Now we interchange the order of summation and observe that

$$\sum_{i<j} \Big(\sum_I \lambda_{I_0}(z_0)...\lambda_{I_N}(z_N)\ G(q_{iI_i}, q_{jI_j}) \Big) =$$

$$\sum_I \sum_{i<j} \lambda_{I_0}(z_0)...\lambda_{I_N}(z_N)\ G(q_{iI_i}, q_{jI_j}) =$$

$$\sum_I \lambda_{I_0}(z_0)...\lambda_{I_N}(z_N) \bigg(\sum_{i<j} G(q_{iI_i}, q_{jI_j}) \bigg) =$$

$$\sum_I \lambda_{I_0}(z_0)...\lambda_{I_N}(z_N)\ \mathcal{E}(I). \tag{65}$$

Therefore, when we sum Equation 64 over all $i < j$, we get precisely the inequality in Equation 54. This completes the proof.

4.6 A More General Result

Though we have no need for it in this paper, we mention an easy generalization of the Energy Theorem. Suppose we have some energy of the form

$$F = \sum_{k=1}^{N} a_k G_k \tag{66}$$

where $a_1, ..., a_N$ is some sequence of real numbers, not necessarily positive. We define

$$\mathbf{ERR}_F = \sum_{k=1}^{N} |a_k|\ \mathbf{ERR}_k \tag{67}$$

Here $\mathbf{ERR}_k$ is the error term associated to G_k in the Energy Theorem. With this definition in place, we have

Theorem 4.9

$$\min_{v \in \langle B \rangle} \mathcal{F}_k(v) \geq \min_{v \in B} \mathcal{F}_k(v) - \mathbf{ERR}_F(B).$$

5 The Algorithm

5.1 Grading a Block

In this section we describe what we mean by *grading* a block. This step feeds into the computational algorithm presented in the next section.

Let B_0 denote the cube of in-radius $\epsilon_0 = 2^{-15}$ about the configuration of $\square$ representing the normalized TBP. We fix some energy G_k. We perform the following tests on a block $B = (Q_0, Q_1, Q_2, Q_3)$, in the order listed.

1. If the calculations in Equation 17 deem B irrelevant, we pass B.

2. If some component square Q_i of B has side length more than $1/2$ we fail B and recommend that B be subdivided along the first such index.

3. If we compute that $Q_i \not\subset [-3/2, 3/2]^2$ for some $i = 1, 2, 3$, we pass B. Given the previous step, Q_i is disjoint from $(-3/2, 3/2)^2$.

4. If the calculations in §2.7 show that $B \subset B_0$, we pass B. Here we take $S = 2^{25}$ and (as we have already said) $\epsilon_0 = 2^{-15}$.

5. If the calculations in §4.1 show that B satisfies Corollary 4.2, we pass B. Otherwise, we fail B and pass along the recommended subdivision.

5.2 Depth First Search

We plug the grading step into the following algorithm.

1. Begin with a list LIST of blocks in $\square$. Initially LIST consists of a single element, namely $\square$.

2. Let B be the last member of LIST. We delete B from LIST and then we grade B.

3. Suppose B passes. If LIST is empty, we halt and declare success. Otherwise, we return to Step 2.

4. Suppose B fails. In this case, we subdivide B along the subdivision recommendation and we append to LIST the subdivision of B. Then we return to Step 2.

If the algorithm halts with success, it implies that every relevant block B either lies in B_0 or does not contain a minimizer.

5.3 The Results

I will detail the technical implementations of the algorithm in §9.

For each $k = 3, 4, 5, 6$, I ran the programs (most recently) on August 4-5, 2016 on my 2014 Macbook pro.

- For G_3 the program finished in about 1 hour and 25 minutes
- For G_4 the program finished in about 1 hour and 39 minutes
- For G_5 the program finished in about 2 hours and 40 minutes.
- For G_6 the program finished in about 5 hours and 38 minutes.

In each case, the program produces a partition of $\Box$ into N_k smaller blocks, each of which is either irrelevant, contains no minimizer, or lies in B_0. Here

$$(N_3, N_4, N_5, N_6) = (5513537, 6201133, 9771906, 20854602).$$

These calculations rigorously establish the following result.

Lemma 5.1 *Let $B_0 \subset \Box$ denote the cube of in-radius 2^{-15} about P_0. If $P \in \Box$ is a minimizer with respect to any of G_3, G_4, G_5, G_6 then P has the same energy as a configuration in B_0.*

The algorithm runs for $G_{10}^{\#}$ with the following modifications.

- We use $\epsilon_0 = 2^{-18}$ and $S = 2^{30}$.
- We use Theorem 4.9 in place of Theorem 4.1.

I ran the algorithm on $G_{10}^{\#}$ in the last week of July 2016 on my 2014 iMac. The calculation ran to completion after about 51 hours and 13 minutes, producing a partition of size 67899862. The calculation establishes the following result.

Lemma 5.2 *Let $B_0^{\#} \subset \Box$ denote the cube of in-radius 2^{-18} about P_0. If $P \in \Box$ is a minimizer with respect to $G_{10}^{\#}$ then P has the same energy as a configuration in $B_0^{\#}$.*

6 Local Analysis of the Hessian

6.1 Eigenvalues of Symmetric Matrices

Let H be a symmetric $n \times n$ real matrix. H always has an orthonormal basis of eigenvectors, and real eigenvalues. H is *positive definite* if all these eigenvalues are positive. This is equivalent to the condition that $Hv \cdot v > 0$ for all nonzero v. More generally, $Hv \cdot v \geq \lambda \|v\|$, where λ is the lowest eigenvalue of H.

Here is one way to bound the lowest eigenvalue of H.

Lemma 6.1 (Alternating Criterion) *Let $\chi(t)$ be the characteristic polynomial of H. Suppose that the coefficients of $P(t) = \chi(t + \lambda)$ are alternating and nontrivial. Then the lowest eigenvalue of H exceeds λ.*

Proof: An alternating polynomial has no negative roots. So, if $\chi(t + \lambda) = 0$ then $t > 0$ and $t + \lambda > \lambda$. ♠

Let H_0 be some positive definite symmetric matrix and let Δ be some other symmetric matrix of the same size. Recall various definitions of the L_2 matrix norm:

$$\|\Delta\|_2 = \sqrt{\sum_{ij} \Delta_{ij}^2} = \sqrt{\mathrm{Trace}(\Delta\Delta^t)} = \sup_{\|v\|=1} \|\Delta v\|. \tag{68}$$

Lemma 6.2 (Variation Criterion) *Suppose that $\|\Delta\|_2 \leq \lambda$, where λ is some number less than the lowest eigenvalue of H_0. Then $H = H_0 + \Delta$ is also positive definite.*

Proof: H is positive definite if and only of $Hv \cdot v > 0$ for every nonzero unit vector v. Let v be such a vector. Writing v in an orthonormal basis of eigenvectors we see that $H_0 v \cdot v > \lambda$. Hence

$$Hv \cdot v = (H_0 v + \Delta v) \cdot v \geq H_0 v \cdot v - |\Delta v \cdot v| > \lambda - \|\Delta v\| \geq \lambda - \|\Delta\|_2 \geq 0.$$

This completes the proof. ♠

6.2 Taylor's Theorem with Remainder

In this section we are just packaging a special case of Taylor's Theorem with Remainder. Here are some preliminary definitions.

- Let $P_0 \in \boldsymbol{R}^7$ be some point.
- Let B denote some cube of in-radius ϵ centered at P_0.
- $\phi : \boldsymbol{R}^7 \to \boldsymbol{R}$ be some function.
- Let $\partial_I \phi$ be the partial derivative of ϕ w.r.t. a multi-index $I = (i_1, ..., i_7)$.
- Let $|I| = i_1 + \cdots i_7$. This is the *weight* of I.
- Let $I! = i_1! \cdots i_7!$.
- Let $\Delta^I = x^{i_1} ... x^{i_7}$. Here $\Delta = (x_1, ..., x_7)$ is some vector.
- For each positive integer N let

$$M_N(\phi) = \sup_{|I|=N} \sup_{P \in B} |\partial_I \phi(P)|, \qquad \mu_N(\phi) = \sup_{|I|=N} |\partial_I \phi(P_0)|. \tag{69}$$

Let U be some open neighborhood of B. Given $P \in B$, let $\Delta = P - P_0$. Taylor's Theorem with Remainder says that there is some $c \in (0,1)$ such that

$$\phi(P) = \sum_{a=0}^{N} \sum_{|I|=a} \frac{|\partial_I \phi(P_0)|}{I!} \Delta^I + \sum_{|I|=N+1} \frac{\partial_I f(P_0 + c\Delta)}{I!} \Delta^I$$

Using the fact that

$$|\Delta^I| \leq \epsilon^{|I|}, \qquad \sum_{|I|=m} \frac{1}{I!} = \frac{7^m}{m!},$$

and setting $N = 4$ we get

$$\sup_{P \in B} |\phi(P)| \leq |\phi(P_0)| + \sum_{j=1}^{4} \frac{(7\epsilon)^j}{j!} \mu_j(\phi) + \frac{(7\epsilon)^5}{(5)!} M_5(\phi). \tag{70}$$

6.3 The Lowest Eigenvalues

We will first deal with G_k for $k = 3, 4, 5, 6$ and then at the end of the chapter we will explain the modifications needed to handle $G^{\#}_{10}$. Let

$$\epsilon_0 = 2^{-15} \tag{71}$$

and B_0 be as in Lemma 5.1 Unless otherwise stated, we will take $k \in \{2, 3, 4, 5, 6\}$. (We treat the case $k = 2$ because it will be useful to us when we need to deal with $G^{\#}_{10}$.) We have the energy map $E_k : B_0 \to \boldsymbol{R}_+$ given by

$$E_k(x_1, ..., x_7) = \sum_{i<j} G_k(\Sigma^{-1}(p_i) - \Sigma^{-1}(p_j)). \tag{72}$$

Here we have set $p_4 = \infty$, and $p_0 = (x_1, 0)$ and $p_i = (x_{2i}, x_{2i+1})$ for $i = 1, 2, 3$. As usual Σ is stereographic projection.

Let $P_0 \in B_0$ denote the point corresponding to the TBP. It follows from symmetry, and also from a direct calculation, that P_0 is a critical point for E_k in all cases. The Main Theorem for G_k now follows from Lemma 5.1 and from the statement that the Hessian of G_k is positive definite throughout B_0.

Let H_k denote the Hessian of E_k. Define

$$(\lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6) = (1, 10, 24, 36, 43). \tag{73}$$

Lemma 6.3 *For each $k = 2, 3, 4, 5, 6$, the Hessian $H_k(P_0)$ is positive definite and its lowest eigenvalue exceeds λ_k.*

Proof: All the pairs (H_k, λ_k) satisfy the Alternating Criterion from §6.1. ♠

6.4 The Target Bounds

Now we set up the inequalities we need to establish in order to show that the hessians H_k are positive definite throughout the small neighborhood B_0.

Define

$$F_k = \sqrt{\sum_{|J|=3} M^2_{J,k}}, \qquad M_{J,k} = \sup_{P \in B_0} |\partial_J E_k(P)|. \tag{74}$$

The sum is taken over all multi-indices J of weight 3.

Lemma 6.4 *Suppose that* $\sqrt{7}\epsilon_0 F_k \leq \lambda_k$. *Then the Hessian of* E_k *is positive definite throughout the neighborhood* B_0.

Proof: We suppress the value of k from our notation. Let H_0 denote the Hessian of E at P_0. Let H denote the Hessian of E at P. Let $\Delta = H - H_0$. Clearly $H = H_0 + \Delta$.

Let γ be the unit speed line segment connecting P to P_0 in $\boldsymbol{R}^7$. Note that $\gamma \subset B_0$ and γ has length $L \leq \sqrt{7}\epsilon_0$. We set $H_L = H$ and we let H_t be the Hessian of E at the point of γ that is t units from H_0.

We have

$$\Delta = \int_0^L D_t(H_t)\ dt. \tag{75}$$

Here D_t is the unit directional derivative of H_t along γ.

Let $(H_t)_{ij}$ denote the ijth entry of H_t. Let $(\gamma_1, ..., \gamma_7)$ be the components of the unit vector in the direction of γ. Using the fact that $\sum_k \gamma_k^2 = 1$ and the Cauchy-Schwarz inequality, and the fact that mixed partials commute, we have

$$(D_t H_t)_{ij}^2 = \Big(\sum_{k=1}^7 \gamma_k \frac{\partial}{\partial x_k} \frac{\partial^2 H_t}{\partial x_i \partial x_j}\Big)^2 \leq \sum_{k=1}^7 \Big(\frac{\partial^3 H_t}{\partial x_i \partial x_j \partial x_k}\Big)^2. \tag{76}$$

Summing this inequality over i and j we get

$$\|D_t H_t\|_2^2 \leq \sum_{i,j,k} \Big(\frac{\partial^3 H_t}{\partial x_i \partial x_j \partial x_k}\Big)^2 \leq F^2. \tag{77}$$

Hence

$$\|\Delta\|_2 \leq \int_0^L \|D_t(H_t)\|_2\ dt \leq LF \leq \sqrt{7}\epsilon_0 F < \lambda. \tag{78}$$

This lemma now follows immediately from Lemma 6.4. ♠

Referring to Equation 69, and with respect to the neighborhood B_0, define $M_8(E_k)$ and $\mu_j(E_k)$ for $j = 4, 5, 6, 7$. Let J be any multi-index of weight 3. Using the fact that

$$\mu_j(\partial_J E_K) \leq \mu_{j+3}(E_k), \qquad M_5(\partial_J E_K) \leq M_8(E_k),$$

we see that Equation 70 gives us the bound

$$M_{J,k} \leq |\partial_J E_k(P_0)| + \sum_{j=1}^4 \frac{(7\epsilon_0)^j}{j!} \mu_{j+3}(E_k) + \frac{(7\epsilon_0)^5}{5!} M_8(E_k). \tag{79}$$

6.5 The Biggest Term

In this section we will prove that the last term in Equation 79 is at most 1. Recall that Σ is stereographic projection. Define

$$f_k(a,b) = \Big(4 - \|\Sigma^{-1}(a,b) - (0,0,1)\|^2\Big)^k = 4^k\Big(\frac{a^2+b^2}{1+a^2+b^2}\Big)^k. \tag{80}$$

$$g_k(a,b,c,d) = \Big(4 - \|\Sigma^{-1}(a,b) - \Sigma^{-1}(c,d)\|^2\Big)^k =$$
$$4^k\Big(\frac{1+2ac+2bd+(a^2+b^2)(c^2+d^2)}{(1+a^2+b^2)(1+c^2+d^2)}\Big)^k \tag{81}$$

Note that

$$f_k(a,b) = \lim_{c^2+d^2\to\infty} g_k(a,b,c,d). \tag{82}$$

We have

$$E_k(x_1,...,x_7) = f_k(x_1,0) + f_k(x_2,x_3) + f_k(x_4,x_5) + f_k(x_6,x_7) +$$
$$g_k(x_1,0,x_2,x_3) + g_k(x_1,0,x_4,x_5) + g_k(x_1,0,x_6,x_7) +$$
$$g_k(x_2,x_3,x_4,x_5) + g_k(x_2,x_3,x_6,x_7) + g_k(x_4,x_5,x_6,x_7).$$

Each variable appears in at most 4 terms, 3 of which appear in a g-function and 1 of which appears in an f-function. Hence

$$M_8(E_k) \le M_8(f_k) + 3M_8(g_k) \le 4M_8(g_k). \tag{83}$$

The last inequality is a consequence of Equation 82 and we use it so that we can concentrate on just one of the two functions above.

Lemma 6.5 *When r,s,D are non-negative integers and $r+s \le 2D$,*

$$\Big|\frac{x^r y^s}{(1+x^2+y^2)^D}\Big| < 1.$$

Proof: The quantity factors into expressions of the form $|x^\alpha y^\beta/(1+x^2+y^2)|$ where $\alpha+\beta \le 2$. Such quantities are bounded above by 1. ♠

For any polynomial Π, let $|\Pi|$ denote the sum of the absolute values of the coefficience of Π. For each 8th derivative $D_I g_k$, we have

$$D_I g_k = \frac{\Pi(a,b,c,d)}{(1+a^2+a^2)^{k+8}(1+c^2+d^2)^{k+8}}, \tag{84}$$

Where Π_I is a polynomial of maximum (a,b) degree at most $2k+16$ and maximum (c,d) degree at most $2k+16$. Lemma 6.5 then gives

$$\sup_{(a,b,c,d)\in \boldsymbol{R}^4} |D_I g_k(a,b,c,d)| \leq |\Pi_I|. \tag{85}$$

Thus, we compute in Mathematica that

$$M_8(g_k) \leq \sup_{k=2,3,4,5,6} \sup_I |\Pi_I| = 13400293856913653760 < 2^{64}. \tag{86}$$

The max is achieved when $k=6$ and $I=(8,0,0,0)$ or $(0,8,0,0)$, etc.

Combining Equations 83 and 86, we get

$$\frac{\Big(7\times 2^{-15}\Big)^5}{5!} \times M_8(E_k) < 1, \tag{87}$$

$k=2,3,4,5,6$.

6.6 The End of the Proof

Let

$$\mu^*_{j+3,k} = \frac{(7\epsilon_0)^j}{(j)!}\mu_{j+3}(E_k) \tag{88}$$

We now estimate $\mu^*_{j,k}$ for $j=4,5,6,7$ and $k=2,3,4,5,6,10$.

The same considerations as above show that

$$\mu^*_{j+3,k} \leq \frac{(7\epsilon_0)^j}{(j)!}\Big(\mu_{j+3}(f_k)+3\mu_{j+3}(g_k)\Big). \tag{89}$$

Here we are evaluating the $(j+3)$rd partial at all points which arise in the TBP configuration and then taking the maximum. For instance, for g_k, one choice would be $(a,b,c,d) = (1,0,-1/2,\sqrt{3}/2)$. Here is a computed matrix of upper bounds for $\mu^*_{j,k}$. The rows give the fixed exponent k.

$$\begin{array}{rcccc} j: & 4 & 5 & 6 & 7 \\ k=2: & 1 & 1 & 1 & 1 \\ k=3: & 3 & 1 & 1 & 1 \\ k=4: & 9 & 1 & 1 & 1 \\ k=5: & 12 & 1 & 1 & 1 \\ k=6: & 122 & 1 & 1 & 1 \\ k=10: & 47480 & 44 & 1 & 1 \end{array} \tag{90}$$

The first and last rows are needed for the next section.

Given these bounds, Equation 79 gives

$$|M_{J,k}| < 1 + \mu^*_7 + \mu^*_6 + \mu^*_5 + \mu^*_4 < 1+1+1+1+122 < 500. \tag{91}$$

The first 1 comes from the previous section. We have been generous in the last inequality to illustrate that the situation is not delicate here.

For any real vector $V = (V_1, ..., V_{343})$ define

$$\overline{V} = (|V_1| + 500, ..., |V_{343}| + 500). \tag{92}$$

Let V_k denote the vector of third partials of E_k, evaluated at P_0, and ordered (say) lexicographically. In view of Equation 91, we have $F_k \leq \|\overline{V}_k\|$. Hence

$$\sqrt{7}\epsilon_0 F_k \leq \sqrt{7}\epsilon_0 \|\overline{V}_k\|. \tag{93}$$

Rounding up to the nearest integer, we compute the coordinatewise inequality

$$\sqrt{7}\epsilon_0(\|\overline{V}_2\|, \|\overline{V}_3\|, ..., \|\overline{V}_6\|) < (1,1,2,4,12) \leq (\lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6). \tag{94}$$

This completes the proof of the Main Theorem for G_3, G_4, G_5, G_6.

6.7 The Last Case

Now we deal with $G^{\#}_{10}$. Let $\epsilon^{\#}_0 = 2^{-18}$. The pair $(H^{\#}_{10}, 1448)$ satisfies the Alternating Criterion. Hence, the Hessian $H^{\#}_{10}$ has lowest eigenvalue greater than 1448. In view of Lemma 5.2 and the results in this chapter, the Main Theorem for $G^{\#}_{10}$ follows from the inequality

$$\sqrt{7}\epsilon^{\#}_0 F_{10} + 28\sqrt{7}\epsilon^{\#}_0 F_5 + 102\sqrt{7}\epsilon^{\#}_0 F_2 < 1448. \tag{95}$$

Since $\epsilon_0^{\#} < \epsilon_0$, the bounds in Equation 94 remains true with respect to $\epsilon_0^{\#}$. This reduces our goal to showing that

$$\sqrt{7}\epsilon_0 F_{10} < 1234 \qquad (= 1448 - 28 \times 4 - 102 \times 1). \tag{96}$$

We compute that

$$M_8(g_{10}) \leq 162516942801336639946752000 < 2^{88} \tag{97}$$

This leads to

$$\frac{\left(7 \times 2^{-18}\right)^5}{5!} \times M_8(E_{10}) < 19 \tag{98}$$

Since we are using $\epsilon_0^{\#}$ in place of ϵ_0, we can divide the last row of the matrix in Equation 90 by $(2^{-3}, 2^{-6}, 2^{-9}, 2^{-12})$ to get

$$(\mu^*_{4,10}, \mu^*_{5,10}, \mu^*_{6,10}, \mu^*_{7,10}) < (5935, 1, 1, 1). \tag{99}$$

This information combines with Equation 79 and Equation 97 to give

$$|M_{J,10}| < 19 + 5935 + 1 + 1 + 1 < 6000. \tag{100}$$

For any real vector $V = (V_1, ..., V_{343})$ define

$$\overline{V} = (|V_1| + 6000, ..., |V_{343}| + 6000). \tag{101}$$

This gives us

$$\sqrt{7}\epsilon F_{10} < \sqrt{7}\epsilon_0 \|\overline{V}_{10}\| < 1091 < 1234. \tag{102}$$

This completes the proof of the Main Theorem for $G_{10}^{\#}$.

7 Facts about Polynomials

7.1 Intervals and Interval Polynomials

We define a *rational interval* to be an interval of the form $I = [L, R]$ where $L, R \in \boldsymbol{Q}$ and $L \leq R$. For each operation $* \in \{+, -, \times\}$ we define

$$I_1 * I_2 = [\min(S), \max(S)], \qquad S = \{L_1 * L_2, L_1 * R_2, R_1 * L_2, R_2 * L_2\}. \tag{103}$$

This definition is such that $r_j \in I_j$ for $j = 1, 2$ then $r_1 * r_2 \in I_1 * I_2$. Moreover, $I_1 * I_2$ is the minimal interval with this property. The minimality property implies that our laws are both associative and distributive:

- $(I_1 + I_2) \pm I_3 = I_1 + (I_2 \pm I_3)$.
- $I_1 \times (I_2 \pm I_3) = (I_1 \times I_2) \pm (I_1 \times I_3)$.

We also can raise a rational interval to a nonnegative integer power:

$$I^k = I \times ... \times I \qquad \text{k times}. \tag{104}$$

An *inteval polynomial* is an expression of the form

$$I_0 + I_1 t + ... + I_n t^n. \tag{105}$$

in which each of the coefficients are intervals and t is a variable meant to be taken in $[0, 1]$. Given the rules above, interval polynomials may be added, subtracted or multiplied, in the obvious way. The associative and distributive laws above give rise to similar results about the arithmetic operations on interval polynomials.

We think of an ordinary polynomial as an interval polynomial, just by taking the intervals to have 0 width. We think of a constant as an interval polynomial of degree 0. Thus, if we have some expression which appears to involve constants, ordinary polynomials, and interval polynomials, we interpret everything in sight as an interval polynomial and then perform the arithmetic operations needed to simplify the expression.

Let $\mathcal{P}$ be the above interval polynomial. We say that $\mathcal{P}$ *traps* the ordinary polynomial

$$C_0 + C_1 t + ... + C_n t^n \tag{106}$$

of the same degree degree if $C_j \in I_j$ for all j. We define the *min* of an interval polynomial to be the polynomial whose coefficients are the left endpoints of the intervals. We define the *max* similarly. If $\mathcal{P}$ is an interval polynomial which traps an ordinary polynomial, then

- $\mathcal{P}$ traps $\mathcal{P}_{\min}$.
- $\mathcal{P}$ traps $\mathcal{P}_{\max}$.
- For all $t \in [0,1]$ we have $\mathcal{P}_{\min}(t) \leq P(t) \leq \mathcal{P}_{\mathrm{mat}}(t)$.

Our arithmetic operations are such that if the polynomial $\mathcal{P}_j$ traps the polynomial P_j for $j = 1,2$, then $\mathcal{P}_1 * \mathcal{P}_2$ traps $P_1 * P_2$. Here $* \in \{+,-,\times\}$.

7.2 Rational Approximations of Power Combos

Suppose that $Y = (a_2, a_3, a_4, b_2, b_3, b_4)$ is a 6-tuple of rational numbers. We are interested in expressions of the form

$$C_Y(x) = a_2 2^{-s/2} + a_3 3^{-s/2} + a_4 4^{-s/2} + b_2 s 2^{-s/2} + b_3 s 3^{-s/2} + b_4 s 4^{-s/2} \quad (107)$$

evaluated on the interval $[-2, 16]$. We call such expressions *power combos*.

For each even integer $2k = -2, ..., 16$. we will construct rational polynomials $A_{Y,2k,-}$ and $A_{Y,2k,+}$ and $B_{Y,2k,-}$ and $B_{Y,2k,+}$ such that

$$A_{Y,2k,-}(t) \leq C_Y(2k-t) \leq A_{Y,2k,+}(t), \qquad t \in [0,1]. \quad (108)$$

$$B_{Y,2k,-}(t) \leq C_Y(2k+t) \leq B_{Y,2k,+}(t), \qquad t \in [0,1]. \quad (109)$$

We ignore the cases $(2,-)$ and $(16,+)$.

The basic idea is to use Taylor's Theorem with Remainder:

$$m^{-s/2} = \sum_{j=0}^{11} \frac{(-1)^j \log(m)^j}{m^k 2^j j!}(s-2k)^j + \frac{E_s}{12!}(s-2k)^{12}. \quad (110)$$

Here E_s is the 12th derivative of $m^{-s/2}$ evaluated at some point in the interval. Note that the only dependence on k is the term m^k in the denominator, and this is a rational number.

The difficulty with this approach is that the coefficients of the above Taylor series are not rational. We get around this trick by using interval polynomials. We first pick specific intervals which trap $\log(m)$ for $m = 2,3,4$. We choose the intervals

$$L_2 = \left[\frac{25469}{36744}, \frac{7050}{10171}\right], \quad L_3 = \left[\frac{5225}{4756}, \frac{708784}{645163}\right], \quad L_4 = \left[\frac{25469}{18372}, \frac{345197}{249007}\right].$$

Each of these intervals has width about 10^{-10}. I found them using Mathematica's Rationalize function. It is an easy exercise to check that $\log(m) \in L_m$ for $m = 2, 3, 4$. It is also an easy exercise to show that

$$\sup_{s\in[-2,16]} \max_{m=2,3,4} \left| \frac{d^{12}}{ds^{12}} m^{-s/2} \right| < 1. \tag{111}$$

Indeed, the true answer is closer to 1/256. What we are saying is that we always have $|E_s| < 1$ in the series expansion from Equation 110. Fixing k we introduce the interval Taylor series

$$A_m(t) = \sum_{j=0}^{11} \frac{(+1)^j (L_m)^j}{m^k 2^j j!} t^j + \left[-\frac{1}{12!}, \frac{1}{12!}\right] t^{12}. \tag{112}$$

$$B_m(t) = \sum_{j=0}^{11} \frac{(-1)^j (L_m)^j}{m^k 2^j j!} t^j + \left[-\frac{1}{12!}, \frac{1}{12!}\right] t^{12}. \tag{113}$$

By construction A_m traps the Taylor series expansion from Equation 110 when it is evaluated at $t = s - 2k$ and $t \in [0, 1]$. Likewise B_m traps the Taylor series expansion from Equation 110 when it is evaluated at $t = 2k - s$ and $t \in [0, 1]$. Define

$$A_Y(t) = a_2A_2(t) + a_3A_3(t) + a_4A_4(t)+$$
$$b_2(2k - t)A_2(t) + b_3(2k - t)A_3(t) + b_4(2k - t)A_4(t) \tag{114}$$

$$B_Y(t) = a_2B_2(t) + a_3B_3(t) + a_4B_4(t)+$$
$$b_2(2k + t)B_2(t) + b_3(2k + t)B_3(t) + b_4(2k + t)B_4(t) \tag{115}$$

By construction, $A_Y(t)$ traps $C_Y(2k - t)$ when $t \in [0, 1]$ and $B_Y(t)$ traps $C_Y(2k + t)$ when $t \in [0, 1]$.

Finally, we define

$$A_{Y,2k,-} = (A_Y)_{\min}, \quad A_{Y,2k,-} = (A_Y)_{\max},$$
$$B_{Y,2k,-} = (B_Y)_{\min}, \quad B_{Y,2k,-} = (B_Y)_{\max}. \tag{116}$$

By construction these polynomials satisfy Equations 108 and 109 respectively for each $k = -1, ..., 8$. These are our under and over approximations.

Remark: We implemented these polynomials in Java and tested them extensively.

7.3 Weak Positive Dominance

Let

$$P(x) = a_0 + a_1 x + ... + a_n x^n \tag{117}$$

be a polynomial with real coefficients. Here we describe a method for showing that $P \geq 0$ on $[0,1]$,

Define

$$A_k = a_0 + \cdots + a_k. \tag{118}$$

We call P *weak positive dominant* (or *WPD* for short) if $A_k \geq 0$ for all k and $A_n > 0$.

Lemma 7.1 *If P is weak positive dominant, then $P > 0$ on $(0,1]$.*

Proof: The proof goes by induction on the degree of P. The case $\deg(P) = 0$ follows from the fact that $a_0 = A_0 > 0$. Let $x \in (0,1]$. We have

$$P(x) = a_0 + a_1 x + x_2 x^2 + \cdots + a_n x^n \geq$$

$$a_0 x + a_1 x + a_2 x^2 + \cdots + a_n x^n =$$

$$x(A_1 + a_2 x + a_3 x^2 + \cdots a_n x^{n-1}) = xQ(x) > 0$$

Here $Q(x)$ is weak positive dominant and has degree $n-1$. ♠

Given an interval $I = [a,b] \subset \boldsymbol{R}$, let A_I be one of the two affine maps which carries $[0,1]$ to I. We call the pair (P,I) *weak positive dominant* if $P \circ A_I$ is WPD. If (P,I) is WPD then $P \geq 0$ on $(a,b]$, by Lemma 7.1. For instance, if P is WPD on $[0,1/2]$ and $[1/2,1]$ then $P > 0$ on $(0,1)$.

7.4 The Positive Dominance Algorithm

Here I describe a method for certifying that a polynomial (of several variables) is non-negative on a polytope. I will restrict to the case when the polytope is the unit cube. I have used this method extensively in other contexts. See e.g. [**S2**]. I don't know if this method already exists in the literature. It is something I devised myself.

Given a multi-index $I = (i_1, ..., i_k) \in (\boldsymbol{N} \cup \{0\})^k$ we let

$$x^I = x_1^{i_1} ... x_k^{i_k}. \tag{119}$$

Any polynomial $F \in \boldsymbol{R}[x_1, ..., x_k]$ can be written succinctly as

$$F = \sum a_I X^I, \qquad a_I \in \boldsymbol{R}. \tag{120}$$

If $I' = (i'_1, ..., i'_k)$ we write $I' \leq I$ if $i'_j \leq i_j$ for all $j = 1, ..., k$. We call F *positive dominant* (PD) if

$$A_I := \sum_{I' \leq I} a_{I'} > 0 \qquad \forall I, \tag{121}$$

Lemma 7.2 *If P is PD, then $P > 0$ on $[0, 1]^k$.*

Proof: When $k = 1$ the proof is the same as in Lemma 7.1, once we observe that also $P(0) > 0$. Now we prove the general case. Suppose the the coefficients of P are $\{a_I\}$. We write

$$P = f_0 + f_1 x_k + ... + f_m x_k^m, \qquad f_j \in \boldsymbol{R}[x_1, ..., x_{k-1}]. \tag{122}$$

Let $P_j = f_0 + ... + f_j$. A typical coefficient in P_j has the form

$$b_J = \sum_{i=1}^{j} a_{Ji}, \tag{123}$$

where J is a multi-index of length $k-1$ and Ji is the multi-index of length k obtained by appending i to J. From equation 123 and the definition of PD, the fact that P is PD implies that P_j is PD for all j. ♠

The positive dominance criterion is not that useful in itself, but it feeds into a powerful divide-and-conquer algorithm. We define the maps

$$A_{j,1}(x_1, ..., x_k) = (x_1, ..., x_{j-1} \frac{x_j + 0}{2}, x_{i+1}, ..., x_k),$$

$$A_{j,2}(x_1, ..., x_k) = (x_1, ..., x_{j-1} \frac{x_j + 1}{2}, x_{j+1}, ..., x_k), \tag{124}$$

We define the jth *subdivision* of P to be the set

$$\{P_{j1}, P_{j2}\} = \{P \circ A_{j,1}, P \circ A_{j,2}\}. \tag{125}$$

Lemma 7.3 *$P > 0$ on $[0, 1]^k$ if and only if $P_{j1} > 0$ and $P_{j2} > 0$ on $[0, 1]^k$.*

Proof: By symmetry, it suffices to take $j=1$. Define

$$[0,1]^k_1=[0,1/2]\times[0,1]^{k-1}, \qquad [0,1]^k_2=[1/2,1]\times[0,1]^{k-1}. \tag{126}$$

Note that

$$A_1([0,1]^k)=[0,1]^k_1, \qquad B_1\circ A_1([0,1]^k)=[0,1]^k_2. \tag{127}$$

Therefore, $P>0$ on $[0,1]^k_1$ if and only if $P_{j1}>0$ on $[0,1]^k$. Likewise $P>0$ on $[0,1]^k_2$ if and only if if $P_{j2}>0$ on $[0,1]^k$. ♠

Say that a *marker* is a non-negative integer vector in $\boldsymbol{R}^k$. Say that the *youngest entry* in the the marker is the first minimum entry going from left to right. The *successor* of a marker is the marker obtained by adding one to the youngest entry. For instance, the successor of $(2,2,1,1,1)$ is $(2,2,2,1,1)$. Let μ_+ denote the successor of μ.

We say that a *marked polynomial* is a pair (P,μ), where P is a polynomial and μ is a marker. Let j be the position of the youngest entry of μ. We define the *subdivision* of (P,μ) to be the pair

$$\{(P_{j1},\mu_+),(P_{j2},\mu_-)\}. \tag{128}$$

Geometrically, we are cutting the domain in half along the longest side, and using a particular rule to break ties when they occur.

Divide-and-Conquer Algorithm:

1. Start with a list LIST of marked polynomials. Initially, LIST consists only of the marked polynomial $(P,(0,...,0))$.

2. Let (Q,μ) be the last element of LIST. We delete (Q,μ) from LIST and test whether Q is positive dominant.

3. Suppose Q is positive dominant. We go back to Step 2 if LIST is not empty. Otherwise, we halt.

4. Suppose Q is not positive dominant. we append to LIST the two marked polynomials in the subdivision of (Q,μ) and then go to Step 2.

If the algorithm halts, it constitutes a proof that $P>0$ on $[0,1]^k$. Indeed, the algorithm halts if and only if $P>0$ on $[0,1]^k$.

Parallel Version: Here is a variant of the algorithm. Suppose we have a list $\{P_1, ..., P_m\}$ of polynomials and we want to show that at least one of them is positive at each point of $[0,1]^k$. We do the following

1. Start with m lists LIST(j) for $j = 1, ..., m$ of marked polynomials. Initially, LIST(j) consists only of the marked polynomial $(P_j, (0, ..., 0))$.

2. Let (Q_j, μ) be the last element of LIST(j). We delete (Q_j, μ) from LIST(j) and test whether Q_j is positive dominant. We do this for $j = 1, 2, ...$ until we get a success or else reach the last index.

3. Suppose *at least one* Q_j is positive dominant. We go back to Step 2 if LIST(j) is not empty. (All lists have the same length.) Otherwise, we halt.

4. Suppose none of $Q_1, ..., Q_m$ is positive dominant. For each j we append to LIST(j) the two marked polynomials in the subdivision of (Q_j, μ) and then go to Step 2.

If this algorithm halts it constitutes a proof that at least one P_j is positive at each point of $[0,1]^k$.

7.5 Discussion

For polynomials in 1 variable, the method of Sturm sequences counts the roots of a polynomial in any given interval. An early version of this paper used Sturm sequences, but I prefer the positive dominance criterion. The calculations for the positive dominance criterion are much simpler and easier to implement.

There are generalizations of Sturm sequences to higher dimensions, and also other positivity criteria (such as the Handelman decomposition) but I bet they don't work as well as the positive dominance algorithm. Also, I don't see how to do the parallel positive dominance algorithm with these other methods.

The positive dominance algorithm works so well that one can ask why I didn't simply use it to prove the Main Theorem straight away. After all, the Main Theorem does reduce to a positivity theorem about a finite set of polynomials. I tried this. However, the polynomials seem to involve an astronomical number of terms. It is not a feasible calculation.

8 Proof of Lemma 1.2

8.1 Some General Considerations

First we discuss the general principle behind Lemma 1.2. Suppose that the TBP is a minimizer with respect to Γ_1 and Γ_2, and a unique minimizr with respect to $\Gamma_3, ..., \Gamma_m$. Suppose also that R is some other energy function and we want to show that the TBP is the unique minimizer with respect to R. This is true if we can find a combination

$$\Gamma = a_0 + a_1\Gamma_1 + ... + a_m\Gamma_m, \qquad a_1, ..., a_k \geq 0. \tag{129}$$

Here a_0 could be negative. This doesn't bother us. such that

- The constants $a_3, ..., a_k$ do not identically vanish.
- $\Gamma(x) \leq R(x)$ for all $x \in (0, 2]$.
- $\Gamma(x) = R(x)$ for $x = \sqrt{2}, \sqrt{3}, \sqrt{4}$.

The three special values are the distances between pairs of points of the TBP. We find it nice to write $\sqrt{4}$ instead of 2.

Let X_0 be the TBP and let X be any other configuration of 5 distinct points on the sphere. Letting Γ_0 be the function which is identically 1, we have

$$\mathcal{E}_R(X) \geq \mathcal{E}_\Gamma(X) = \sum_{i=0}^{m} a_i \mathcal{E}_{\Gamma_k}(X) > \sum_{i=0}^{m} a_i \mathcal{E}_{\Gamma_k}(X_0) = \mathcal{E}_\Gamma(X_0) = \mathcal{E}_R(X_0).$$

Here is how we find such positive combinations. We set $m = 4$, so that we are looking for 5 coefficients $a_0, ..., a_4$. We impose the 5 conditions

- $\Gamma(x) = R(x)$ for $x = \sqrt{2}, \sqrt{3}, \sqrt{4}$.
- $\Gamma'(x) = R'(x)$ for $x = \sqrt{2}, \sqrt{3}$.

Here $R' = dR/dx$ and $\Gamma' = d\Gamma/dx$. These 5 conditions give us 5 linear equations in 5 unknowns. In the cases described below, the associated matrix is invertible and there is a unique solution. In our situation it will be obvious that the constants a_1, a_2, a_3, a_4 cannot identically vanish. That $\Gamma \leq R$ is far from obvious, but the techniques from the previous chapter will estblish this in each case of interest.

8.2 Finding the Coefficients

Recall that $R_s(r) = r^{-s}$ when $s > 0$ and $R_s(r) = -r^{-s}$ when $s < 0$. We break Lemma 1.2 into 3 cases.

1. When $s \in (-2, 0)$ we use G_1, G_2, G_3, G_5.
2. When $s \in (0, 6]$ we use G_1, G_2, G_4, G_6.
3. When $s \in [6, 13]$ we use $G_1, G_2, G_5, G_{10}^{\#}$.

We will also keep track of the expression

$$\delta = 2\Gamma'(2) - 2R'(2). \tag{130}$$

In the first case we have

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ \delta \end{bmatrix} = \frac{1}{144} \begin{bmatrix} 0 & 0 & -144 & 0 & 0 & 0 \\ -312 & -96 & 408 & 24 & 80 & 0 \\ 684 & -288 & -396 & -54 & -144 & 0 \\ -402 & 264 & 138 & 33 & 68 & 0 \\ 30 & -24 & -6 & -3 & -4 & 0 \\ 2496 & 768 & -3264 & -192 & -640 & -144 \end{bmatrix} \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2}, \\ s4^{-s/2} \end{bmatrix}.$$

In the second case we have

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ \delta \end{bmatrix} = \frac{1}{792} \begin{bmatrix} 0 & 0 & 792 & 0 & 0 & 0 \\ 792 & 1152 & -1944 & -54 & -288 & 0 \\ -1254 & -96 & 1350 & 87 & 376 & 0 \\ 528 & -312 & -216 & -39 & -98 & 0 \\ -66 & 48 & 18 & 6 & 10 & 0 \\ -6336 & -9216 & 15552 & 432 & 2304 & 792 \end{bmatrix} \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2}, \\ s4^{-s/2} \end{bmatrix}.$$

In the third case we have

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \widehat{a}_4 \\ \delta \end{bmatrix} = \frac{1}{268536} \begin{bmatrix} 0 & 0 & 268536 & 0 & 0 & 0 \\ 88440 & 503040 & -591480 & -4254 & -65728 & 0 \\ -77586 & -249648 & 327234 & 2361 & 65896 & 0 \\ 41808 & -19440 & -22368 & -2430 & -9076 & 0 \\ -402 & 264 & 138 & 33 & 68 & 0 \\ -707520 & -4024320 & 4731840 & 34032 & 525824 & 268536 \end{bmatrix} \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2}, \\ s4^{-s/2} \end{bmatrix}.$$

Thus the coefficients are precisely the power combos considered in §7.2. These power combos are functions of the variable s.

8.3 Positivity Proof

Now we explain how we prove that $a_1, a_2, a_3, a_4, \delta > 0$ on the relevant intervals. We will do the three cases one at a time. We consider a_1 on $(-2, 0)$ in detail.

- We set $Y = (-312, -96, 408, 24, 80, 0)$, the row of the relevant matrix corresponding to a_1. By construction, $a_1(s) = C_Y(s)$.
- We verify that the two under-approximations $A_{Y,-2,+}$ and $A_{Y,0,-}$ are WPD on $[0, 1/2]$ and on $[1/2, 1]$. Hence these functions are positive on $(0, 1]$.
- Since $A_{Y,-2,+}(t) \leq a_1(t-2)$ for $t \in [0, 1]$ we see that $a_1 > 0$ on $(-2, -1]$.
- Since $A_{Y,0,-}(t) \leq a_1(-t)$ for $t \in [0, 1]$ we see that $a_1 > 0$ on $[1, 0)$.

The same argument works for a_2, a_3, a_4, δ on $[-2, 0]$. In each case, the relevant under-approximation is either WPD on $[0, 1]$ or WPD on $[0, 1/2]$ and $[1/2, 1]$. We conclude that $a_1, a_2, a_3, a_4, \delta > 0$ on $(-2, 0)$. The statement $\delta > 0$ means that $\Gamma'(2) > R'(2)$ for all $s \in (-2, 0)$. For later use we use the same method to check that

$$\Gamma(0) = c_0 + 4c_1 + 16c_2 + 256c_3 + 1024c_5 < 0$$

for all $s \in (-2, 0)$.

We do the same thing on the interval $[0, 6]$, except that we use the intervals $[0, 1]$, $[1, 2]$, $[2, 3]$, $[3, 4]$, $[4, 5]$ and $[5, 6]$ and the corresponding under-approximations. In this case, every polynomial in sight – all $30 = 5 \times 6$ of them – is WPD on $[0, 1]$. We conclude that that $a_1, a_2, a_3, a_4, \delta > 0$ on $(0, 6]$. To be sure, we check the interval endpoints $s = 1, 2, 3, 4, 5, 6$ by hand.

Finally, we do the same thing on the interval $[6, 13]$, using the intervals $[6, 7], ..., [12, 13]$. Again, every polynomial in sight is WPD on $[0, 1]$, and in fact PD on $[0, 1]$. Since we just check the WPD condition, we also check that our functions are positive at the integer values in $[6, 13]$ by hand. We conclude that $a_1, a_2, a_3, a_4, \delta > 0$ on $[6, 13]$.

Remark: In the third case, we checked additionally that $a_1, a_2, a_3, a_4, \delta$ are positive on $[3, 13 + 1/16]$. Thus, our result is really true for all power law exponents up to $13 + 1/16$. The reader can play with our graphical user interface, see plots of all these functions, and run positivity tests.

8.4 Under Approximation: Case 1

Here we show that $\Gamma_s(r) \leq R_s(r)$ for all $r \in (0,2]$ and all $s \in (-2,0)$. We suppress the dependence on s as much as we can. In particular, we set $R = R_s$, etc. Here $R < 0$ so we want $\Gamma/R > 1$. Define

$$H(r) = \frac{\Gamma}{R} - 1 = -r^s\Gamma - 1. \tag{131}$$

We just have to show that $H \geq 0$ on $(0,2)$. Let $H' = dH/dr$.

Lemma 8.1 *H' has 4 simple roots in $(0,2)$.*

Proof: We count roots with multiplicity. We have

$$H'(r) = -r^{s-1}(s\Gamma(r) + r\Gamma'(r)). \tag{132}$$

Comgining Equation 132 with the general equation

$$rG_k'(r) = 2kG_k(r) - 8kG_{k-1}(r), \tag{133}$$

we see that the positive roots of $H'(r)$ are the same as the positive roots of

$$-r^{s-1}H'(r) = (10+s)c_4G_5(r) - 40c_4G_4(r) + \sum_{k=1}^{3} b_kG_k(r) + b_0. \tag{134}$$

Here $b_0, ..., b_3$ are coefficients we don't care about. Making the substitution $t = 4 - r^2$ we see that the roots of H' in $(0,2)$ are in bijection with the roots in $(0,4)$ of

$$\psi(t) = t^5 - \frac{40}{10+s}t^4 + b_3t^3 + b_2t^2 + b_1t + b_0. \tag{135}$$

Moreover, the change of coordinates is a diffeomorphism from $(0,4)$ to $(0,2)$ and so it carries simple roots to simple roots.

The polynomial ψ has 5 roots counting multiplicities. Let's find 4 of these roots first. Since $H(\sqrt{2}) = H(\sqrt{3}) = H'(\sqrt{2}) = H'(\sqrt{3}) = 0$, we see that H' has at least 4 roots in $(0,2)$. Besides the roots at $\sqrt{2}$ and $\sqrt{3}$, there is a root in $(\sqrt{2}, \sqrt{3})$ and a root in $(\sqrt{3}, 2)$. This means that ψ has 4 corresponding roots in $(0,4)$. We claim that ψ has an even number of roots in $(0,2)$, counting multiplicity. Once we know this, we can say that the 4 roots we have found are simple. But then the corresponding 4 roots of H' in $(0,2)$ are simple and there are no others.

Now for the parity argument. Since $R<0$ on $(0,2)$ and $R(0)=0$ and $\Gamma(0)<0$ we see that $H(r)\to\infty$ as $r\to 0$. Hence $\psi(t)\to\infty$ as $t\to 4$. Hence, there are arbitrarily small values of $\delta>0$ such that $\psi(4-\delta)>0$.

Since $\Gamma'(2)>R'(2)$ and $R(2)<0$ we see that $H'(2)<0$. Since our change of coordinates is an orientation reversing diffeomorphism we have $\psi(0)>0$ by the chain rule.

Now we know that there are arbitrarily small values $\delta>0$ so that $\psi(\delta)>0$ and $\psi(4-\delta)>0$. But then the number of roots of ψ in $(\delta,4-\delta)$ is even. Since δ is arbitrary, the number roots of ψ in $(0,4)$ is even. ♠

Lemma 8.2 *$H''(\sqrt{2})>0$ and $H''(\sqrt{3})>0$ for all $s\in(-2,0)$.*

Proof: How we mention the explicit dependence on s and remember that we are taking about H_s. We check directly that $H''_{-1}(\sqrt{2})>0$. It cannot happen that $H''_s(\sqrt{2})=0$ for other $s\in(-2,0)$ because then H'_s does not have only simple roots in $(0,2)$. Hence $H''_s(\sqrt{2})>0$ for all $s\in(-2,0)$. The same argument shows that $H''_s(\sqrt{3})>0$ for all $s\in(-2,0)$. ♠

Now we set $H=H_s$ again.

Lemma 8.3 *For all sufficiently small $\delta>0$ the quantities*

$$H(0+\delta),\qquad H(\sqrt{2}\pm\delta),\qquad H(\sqrt{3}\pm\delta),\qquad H(2-\delta)$$

are positive.

Proof: We have already seen that $H(\delta)\to+\infty$ as $\delta\to 0$. Likewise, we have seen that $H'(2)<0$ and $H(2)=0$. So $H(2-\delta)>0$ for all sufficiently small δ. Finally, the case of $\sqrt{2}$ and $\sqrt{3}$ follows from the previous lemma and the second derivative test. ♠

We already know that H' has exactly 4 simple roots in $(0,2)$. In particular, the interval $(0,\sqrt{2})$ has no roots of H' and the intervals $(\sqrt{2},\sqrt{3})$ and $(\sqrt{3},2)$ have 1 root each. Finally, we know that $H>0$ sufficiently near the endpoints of all these intervals. If $H(x)<0$ for some $x\in(0,2)$, then x must be in one of the 3 intervals just mentioned, and this interval contains at least 2 roots of H'. This is a contradiction.

8.5 Under Approximation: Case 2

This time we have $s \in (0,6]$. We have $R > 0$ so we set

$$H = \frac{\Gamma}{R} - 1. \tag{136}$$

It suffices to prove that $H \geq 0$ on $(0,2]$.

Everything in Case 1 works here, word for word, provided that Lemma 8.1 holds for H, when $s \in (0,6]$. The proof is exactly the same, except for a global sign, and the fact that this time we have

$$\psi(t) = t^6 - \frac{48}{12+s}t^4 + b_4 t^4 b_3 t^3 + b_2 t^2 + b_1 t + b_0. \tag{137}$$

We just have to show that ψ has 4 simple roots in $(0,4)$. Note that the sum of the 6 roots of ψ is $48/(12+s) < 4$. This works because $s > 0$ here. The 4 roots of ψ we already know about are 1 and 2 and some number in $(0,1)$ and some number in $(1,2)$. The sum of these roots exceeds 4 and so the remaining two roots cannot also be positive. Hence ψ has at most 5 roots in $(0,4)$.

The parity argument works the same way. This time $H(\delta) \to \infty$ as $\delta \to 0$ because $R \to \infty$ and Γ is bounded. The parity argument shows that ψ has an even number of roots in $(0,4)$. Hence, ψ has exactly 4 such roots and they are all simple. Hence H' has exactly 4 simple roots in $(0,2)$. This completes the proof in Case 2.

8.6 Under Approximation: Case 3

This time we have $s \in [6,16]$ and everything is as in Case 2. All we have to do is show that the polynomial ψ, as Equation 137, has exactly 4 simple roots in $(0,2)$. This time we have

$$\psi(t) = t^{10} - \frac{80}{20+s}t^9 + b_8 t^8 + ... + b_0. \tag{138}$$

Again these coefficients depend on s.

Remark: ψ only has 7 nonzero terms and hence can only have 6 positive real roots. The number of positive real roots is bounded above by Descartes' rule of signs. Unfortunately, ψ turns out to be alternating and so Descartes' rule of signs does not eliminate the case of 6 roots. This approach seems useless. The sum of the roots of ψ is less than 4, so it might seem as if we could proceed as in Case 2. Unfortunately, there are 10 such roots and this approach also seems useless. We will take another approach to proving what we want.

Lemma 8.4 *When $s = 6$ the polynomial ψ has 4 simple roots in $(0, 4)$.*

Proof: We compute explicitly that

$$\psi(t) = t^{10} - \frac{40}{13}t^9 + \frac{830304}{5785}t^5 - \frac{415152}{1157}t^4 + \frac{789255}{1157}t^2 - \frac{3264104}{5785}t + \frac{115060}{1157}.$$

This polynomial only has 4 real roots – the ones we know about. The remaining roots are all at least $1/2$ away from the interval $(0, 4)$ and so even a very crude analysis would show that these roots do not lie in $(0, 4)$. We omit the details.

Lemma 8.5 *Suppose, for all $s \in [6, 13]$ that ψ only has simple roots in $(0, 4)$. Then in all cases ψ has exactly 4 such roots.*

Proof: Let N_s denote the number of simple roots of ψ at the parameter s. The same argument as in Cases 1 and 2 shows that N_s is always even. Suppose s is not constant. Consider the infimal $u \in (6, 13]$ such that $N_u > 4$. The roots of ψ vary continuously with s. How could more roots move into $(0, 4)$ as $s \to u$?

One possibility is that such a root r_s approaches from the upper half plane or from the lower half plane. That is, r_s is not real for $s < u$. Since ψ is a real polynomial, the conjugate $overliner_s$ is also a root. The two roots r_s and $\overline{r}_s$ are approaching $(0, 4)$ from either side. But then the the limit

$$\lim_{s \to u} r_s$$

is a double root of ψ in $(0, 4)$. This is a contradiction.

The only other possibility is that the roots approach along the real line. Hence, there must be some $s < u$ such that both 0 and 4 are roots of ψ. But the same parity argument as in Case 1 shows $\psi(0) > 0$ for all $s \in [6, 13]$. ♠

To finish the proof we just have to show that ψ only has simple roots in $(0,4)$ for all $s \in [6,13]$. We bring the dependence on s back into our notation and write ψ_s. It suffices to show that that ψ_s and $\psi_s' = d\psi_s/dr$ do not simultaneously vanish on the rectangular domain $(s,r) \in [6,13] \times [0,4]$. This is a job for our method of positive dominance.

We will explain in detail what do on the smaller domain

$$(s,r) = [6,7] \times [0,4].$$

The proof works the same for the remining 1×4 rectangles. The coefficients of ψ_s and ψ_s' are power combos in the sense of Equation 7.2.

We have rational vectors $Y_0, ..., Y_9$ such that

$$\psi_s(r) = \sum_{j=0}^{9} C_j r^j, \qquad C_j = C_{Y_j}. \tag{139}$$

We have the under- and over-approximations:

$$A_j = A_{Y_j,6,+} \qquad B_j = B_{Y_j,6,+} \tag{140}$$

We then define 2-variable under- and over-approximations:

$$\underline{\psi}(t,u) = \sum_{i=0}^{9} A_i(t)(4u)^i, \qquad \overline{\psi}(t,u) = \sum_{i=0}^{9} B_i(t)(4u)^i. \tag{141}$$

We use $4u$ in these sums because we want our domains to be the unit square. We have

$$\underline{\psi}(t,u) \leq \psi_{6+t}(4u) \leq \overline{\psi}(t,u), \qquad \forall (t,u) \in [0,1]^2. \tag{142}$$

Now we do the same thing for ψ_s'. We have rational vectors $Y_0', ..., Y_8' \in \boldsymbol{Q}^8$ which work for ψ' in place of ψ, and this gives under- and over-approximations $\underline{\psi}'$ and $\overline{\psi}'$ which satisfy the same kind of equation as Equation 142.

We run the parallel positive dominance algorithm on the set of functions $\{\underline{\phi}, -\overline{\phi}, \underline{\phi}', -\overline{\phi}'\}$ and the algorithm halts. This constitutes a proof that at least one of these functions is positive at each point. But then at least one of $\phi_s(r)$ or $\phi_s'(r)$ is nonzero for each $s \in [6,13]$ and each $r \in [0,4]$. Hence ψ_s only has simple roots in $(0,4)$. This completes our proof in Case 3.

Our proof of Lemma 1.2 is done.

9 Computational Details

9.1 Getting the Program

Our computer program is written in Java. At least in 2016, one can get it from my Brown University website:

http://www.math.brown.edu/∼res/Java/Riesz.tar

The directory has 3 relevant subdirectories:

- **Approximations:** This has the computer code for with Lemma 1.2.
- **Hessian:** This has the Mathematica code used in the the local analysis of the Hessian.
- **Riesz:** This has the main program, which runs the divide and conquer algorithm from §5.

This main directory has a README file which contains more information about these directories. Each subdirectory has a README file as well, which gives information about running the program. The **Approximations** and **Riesz** directories each contain the code for java programs. The programs each have a documentation window which explains how to operate the program. Each of the programs also has a debugging mode, where the main operations are checked.

9.2 Debugging

One serious concern about any computer-assisted proof is that some of the main steps of the proof reside in computer programs which are not printed, so to speak, along with the paper. It is difficult for one to directly inspect the code without a serious time investment, and indeed the interested reader would do much better simply to reproduce the code and see that it yields the same results.

The worst thing that could happen is if the code had a serious bug which caused it to suggest results which are not actually true. Let me explain the extent to which I have debugged the code. Each of the java programs has a debugging mode, in which the user can test that various aspects of the program are running correctly. While the debugger does not check every method, it does check that the main ones behave exactly as expected.

Here is what the user can check in the **Approximations** program:

- Our under-approximation (respectively over-approximation) of $\log(m)$ is less than (respectively greater than) the numerically computed value of $\log(m)$. Here $m = 2, 3, 4$.
- The series under-approximation (respectively over-approximation) to a random power combo evaluated at a random point in a random unit integer interval is less than (respectively greater than) and very close to the power combo when it is computed numerically using Java's power function.
- The series under-approximations to the functions ψ and ψ' from §8.6 behave graphically as they should. For instance, when $\underline{\psi}$ is plotted alongside the graph of the function $H = (1 - \Gamma/F)$, the zeros of $\underline{\psi}$ visually match the locations of the extrema of H. Likewise the zeros of $\underline{\psi}'$ visually match the extrema of $\underline{\psi}$.
- The polynomial subdivision from Equation 124 checks out correctly on random inputs – both in the 1-variable case and in the 2-variable case.

Here are the things one can check for the main program, the one in the **Riesz** directory.

- You can check on random inputs that the interval arithmetic operations are working properly.
- You can check on random inputs that the vector operations - dot product, addition, etc. - are working properly.
- You can check for random dyadic squares that the floating point and interval arithmetic measurements match in the appropriate sense.
- You can select a block of your choice and compare the estimate from the Energy Theorem with the minimum energy taken over a million random configurations in the block.
- You can open up an auxiliary window and see the grading step of the algorithm performed and displayed for a block of your choosing.

I have not included a debugger for the code in the **Hessian** directory because this Mathematica code is quite straightforward. The code performs the calculations described in detail in §6.

In any case, I view Lemmas 5.1 and 5.2 as the main mathematical contributions to the paper, and these are covered by the code in the **Riesz** directory.

9.3 Integer Calculations

Let me discuss the implementation of the divide and conquer algorithm. We manipulate blocks and dyadic squares using **longs**. These are 64 bit integers. Given a dyadic square Q with center (x, y) and side length 2^{-k}, we store the triple

$$(Sx, Sy, k). \tag{143}$$

Here $S = 2^{25}$ when we do the calculations for G_3, G_4, G_5, G_6 and $S = 2^{30}$ when we do the calculation for $\widehat{G}_{10}$. The reader can modify the program so that it uses any power of 2 up to 2^{40}. Similarly, we store a dyadic segment with center x and side length 2^{-k} as (Sx, k).

The subdivision is then obtained by manipulating these triples. For instance, the top right square in the subdivision of (Sx, Sy, k) is

$$(Sx - 2^{-k+1}S, Sy - 2^{-k+1}S, k + 1).$$

The scale 2^N allows for N such subdivisions before we lose the property that the squares are represented by integer triples. The biggest dyadic square is stored as $(0, 0, -2)$, and each subdivision increases the value of k by 1. We terminate the algorithm if we ever arrive at a dyadic square whose center is not an even pair of integers. We never reach this eventuality when we run the program on the functions from the Main Theorem, but it does occur if we try functions like G_7.

We make exact comparisons for Steps 1-4 in the grading part of the algorithm described in §5.1. For instance, the point of scaling our square centers by S is that the inequalities which go into the calculations in §2.7 are all integer inequalities. We are simply clearing denominators. It is only Step 5 which requires floating point (or interval) calculations.

9.4 Interval Arithmetic

We implement interval arithmetic the same way that we did in [**S1**]. Here we repeat some of the discussion, but abbreviate it. We make some changes to the way we do things in [**S1**], and also things are simpler here because we never use the square root function.

Java represents real numbers by **doubles**, essentially according to the scheme discussed in [**I**, §3.2.2]. A double is a 64 bit string where 11 of the bits control the exponent, 52 of the bits control the binary expansion, and one bit controls the sign. The non-negative doubles have a lexicographic ordering, and this ordering coincides with the usual ordering of the real numbers they represent. The lexicographic ordering for the non-positive doubles is the reverse of the usual ordering of the real numbers they represent. To *increment* x_+ of a positive double x is the very next double in the ordering. This amounts to treating the last 63 bits of the string as an integer (written in binary) and adding 1 to it. With this interpretation, we have $x_+ = x + 1$. We also have the decrement $x_- = x - 1$. Similar operations are defined on the non-positive doubles. These operations are not defined on the largest and smallest doubles, but our program never encounters (or comes anywhere near) these.

Let $\boldsymbol{D}$ be the set of all doubles. Let

$$\boldsymbol{R}_0 = \{x \in \boldsymbol{R} |\ |x| \leq 2^{50}\} \tag{144}$$

Our choice of 2^{50} is an arbitrary but convenient cutoff. Let $\boldsymbol{D}_0$ denote the set of doubles representing reals in $\boldsymbol{R}_0$.

According to the discussion in [**I**, 3.2.2, 4.1, 5.6], there is a map $\boldsymbol{R}_0 \to \boldsymbol{D}_0$ which maps each $x \in \boldsymbol{R}_0$ to some $[x] \in \boldsymbol{D}_0$ which is closest to x. In case there are several equally close choices, the computer chooses one according to the method in [**I**, §4.1]. This "nearest point projection" exists on a subset of $\boldsymbol{R}$ that is much larger than $\boldsymbol{R}_0$, but we only need to consider $\boldsymbol{R}_0$. We also have the inclusion $r : \boldsymbol{D}_0 \to \boldsymbol{R}_0$, which maps a double to the real that it represents.

Our calculations just use the arithmetic operations (plus, minus, times, divide) These operations act on $\boldsymbol{R}_0$ in the usual way. Operations with the same name act on $\boldsymbol{D}_0$. Regarding these operations, [**I**, §5] states that *each of the operations shall be performed as if it first produced an intermediate result correct to infinite precision and with unbounded range, and then coerced this intermediate result to fit into the destination's format*. Thus, for doubles

$x, y \in \boldsymbol{D}_0$ such that $x * y \in \boldsymbol{D}_0$ as well, we have

$$x * y = [r(x) * r(y)]; \qquad * \in \{+, -, \times, \div\}. \tag{145}$$

The operations on the left hand side represent operations on doubles and the operations on the right hand side represent operations on reals.

It might happen that $x, y \in \boldsymbol{D}_0$ but $x * y$ is not. To avoid this problem, we make the following checks before performing any arithmetic operation.

- For addition and subtraction, $\max(|x|, |y|) \leq 2^{40}$.
- For multiplication, either $|x| \leq 2^{40}$ and $|y| \leq 2^{10}$ or $|x| \leq 2^{10}$ and $|y| \leq 2^{40}$.
- For division, $|x| \leq 2^{40}$ and $|y| \leq 2^{10}$ and $|y| \geq 2^{-10}$.

We set the calculation to abort if any of these conditions fails.

For us, an *interval* is a pair $I = (x, y)$ of doubles with $x \leq y$ and $x, y \in \boldsymbol{D}_0$. Say that I *bounds* $z \in \boldsymbol{R}_0$ if $x \leq [z] \leq y$. This is true if and only if $x \leq z \leq y$. Define

$$[x, y]_o = [x_-, y_+]. \tag{146}$$

This operation is well defined for doubles in $\boldsymbol{D}_0$. We are essentially *rounding out* the endpoints of the interval. Let I_0 and I_1 denote the left and right endpoints of I. Letting I and J be intervals, we define

$$I * J = (\min_{ij} I_i * I_j, \max_{ij} I_i * I_j)_o. \tag{147}$$

That is, we perform the operations on all the endpoints, order the results, and then round outward. Given Equation 145, we the interval $I * J$ bounded $x * y$ provided that I bounds x and J bounds y. Except for the rounding out, this is the same as what we discussed for the rational interals in §7.1.

We also define an interval version of a vector in $\boldsymbol{R}^3$. Such a vector consists of 3 intervals. The only operations we perform on such objects are addition, subtraction, scaling, and taking the dot product. These operations are all built out of the arithmetic operations.

All of our calculations come down to proving inequalities of the form $x < y$. We imagine that x and y are the outputs of some finite sequence of arithmetic operations and along the way we have intervals I_x and I_y which respectively bound x and y. If we know that the right endpoint of I_x is less than the left endpoint of I_y, then this constitutes a proof that $x < y$. The point is that the whole interval I_x lies to the left of I_y on the number line.

10 References

[**A**] A. N. Andreev, *An extremal property of the icosahedron* East J Approx **2** (1996) no. 4 pp. 459-462

[**BBCGKS**] Brandon Ballinger, Grigoriy Blekherman, Henry Cohn, Noah Giansiracusa, Elizabeth Kelly, Achill Schurmann,
Experimental Study of Energy-Minimizing Point Configurations on Spheres, arXiv: math/0611451v3, 7 Oct 2008

[**BDHSS**] P. G. Boyvalenkov, P. D. Dragnev, D. P. Hardin, E. B. Saff, M. M. Stoyanova, *Universal Lower Bounds and Potential Energy of Spherical Codes*, Constructive Approximation 2016 (to appear)

[**C**] Harvey Cohn, *Stability Configurations of Electrons on a Sphere*, Mathematical Tables and Other Aids to Computation, Vol 10, No 55, July 1956, pp 117-120.

[**CK**] Henry Cohn and Abhinav Kumar, *Universally Optimal Distributions of Points on Spheres*, J.A.M.S. **20** (2007) 99-147

[**CCD**] online website:
http://www-wales.ch.cam.ac.uk/∼ wales/CCD/Thomson/table.html

[**DLT**] P. D. Dragnev, D. A. Legg, and D. W. Townsend, *Discrete Logarithmic Energy on the Sphere*, Pacific Journal of Mathematics, Volume 207, Number 2 (2002) pp 345–357

[**HS**], Xiaorong Hou and Junwei Shao, *Spherical Distribution of 5 Points with Maximal Distance Sum*, arXiv:0906.0937v1 [cs.DM] 4 Jun 2009

[**I**] IEEE Standard for Binary Floating-Point Arithmetic (IEEE Std 754-1985) Institute of Electrical and Electronics Engineers, July 26, 1985

[**RSZ**] E. A. Rakhmanoff, E. B. Saff, and Y. M. Zhou, *Electrons on the Sphere*,
Computational Methods and Function Theory, R. M. Ali, St. Ruscheweyh, and E. B. Saff, Eds. (1995) pp 111-127

[**S1**] R. E. Schwartz, *The* 5 *Electron Case of Thomson's Problem*, Journal of Experimental Math, 2013.

[**S2**] R. E. Schwartz, *The Projective Heat Map*, A.M.S. Research Monograph, 2017 (to appear).

[**SK**] E. B. Saff and A. B. J. Kuijlaars, *Distributing many points on a Sphere*, Math. Intelligencer, Volume 19, Number 1, December 1997 pp 5-11

[**Th**] J. J. Thomson, *On the Structure of the Atom: an Investigation of the Stability of the Periods of Oscillation of a number of Corpuscles arranged at equal intervals around the Circumference of a Circle with Application of the results to the Theory of Atomic Structure*. Philosophical magazine, Series 6, Volume 7, Number 39, pp 237-265, March 1904.

[**T**] A. Tumanov, *Minimal Bi-Quadratic energy of* 5 *particles on* 2*-sphere*, Indiana Univ. Math Journal, **62** (2013) pp 1717-1731.

[**W**] S. Wolfram, *The Mathematica Book*, 4th ed. Wolfram Media/Cambridge University Press, Champaign/Cambridge (1999)

[**Y**], V. A. Yudin, *Minimum potential energy of a point system of charges* (Russian) Diskret. Mat. **4** (1992), 115-121, translation in Discrete Math Appl. **3** (1993) 75-81