Divide and Conquer: A Distributed Approach to Five Point Energy Minimization

Richard Evan Schwartz

January 10, 2025

Abstract

This paper proves the 1977 Melnyk-Knopf-Smith phase transition conjecture for 5-point energy minimization. This result contains, as a special case, the solution of Thomson's 5 electron problem from 1904.

1 Introduction

1.1 History and Context

Let S^2 be the unit sphere in $\boldsymbol{R}^3$. Given a configuration $\{p_i\} \subset S^2$ of N distinct points and a function $F : (0, 2] \to \boldsymbol{R}$, define

$$F(P) = \sum_{1 \le i < j \le N} F(\|p_i - p_j\|). \tag{1}$$

This quantity is commonly called the *F-potential* or the *F-energy* of P. A configuration P is a *minimizer* for F if $F(P) \le F(P')$ for all other N-point configurations P'. The question of finding energy minimizers has a long literature; the classic case goes back to Thomsom [**Th**] in 1904.

The classic choice for this question is $F = R_s$, the *Riesz potential*, given by $R_s(d) = d^{-s}$. The Riesz potential is defined when $s > 0$. When $s < 0$ the corresponding function $R_s(d) = -d^{-s}$ is called the *Fejes-Toth potential*. The case $s = 1$ is specially called the *Coulomb potential* or the *electrostatic potential*. This case of the energy minimization problem is known as *Thomson's problem*. See [**Th**].

There is a large literature on the energy minimization problem. See [**Fö**] and [**C**] for some early local results. See [**MKS**] for a definitive numerical study on the minimizers of the Riesz potential for n relatively small. The website [**CCD**] has a compilation of experimental results which stretches all the way up to about $n = 1000$. The paper [**SK**] gives a nice survey of results, 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 $n = 2, 3$ the problem is fairly trivial. See [**KY**], [**A**], [**Y**] for the result that the three Platonic solids with triangular faces minimize all powe-law potentials. This result is subsumed by [**CK**, Theorem 1.2], a powerful result about the so-called sharp configurations.

The case $n = 5$ has been notoriously intractable. First let me introduce the two main players. The *Triangular Bi-Pyramid* (TBP) is the 5 point configuration having one point at the north pole, one point at the south pole, and 3 points arranged in an equilateral triangle on the equator. A *Four Pyramid* (FP) is a 5-point configuration having one point at the north pole and 4 points arranged in a square equidistant from the north pole. Here is a run-down on what is known so far:

- The paper [**HZ**] has a rigorous computer-assisted proof that the TBP is the unique minimizer for the potential $F(r) = -r$. (Polya's problem).

- 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. Again $R_s(d) = d^{-s}$.

- The paper [**DLT**] gives a traditional proof that the TBP is the unique minimizer for the logarithmic potential.

- In [**BHS**, Theorem 7] it is shown that, as $s \to \infty$, any sequence of 5-point minimizers w.r.t. R_s must converge (up to rotations) to the FP having one point at the north pole and the other 4 points on the equator. In particular, the TBP is not a minimizer w.r.t R_s when s is sufficiently large.

- Define $G_k(r) = (4 - r^2)^k$. In [**T**], A. Tumanov proves that the TBP is the unique minimizer for G_2. The minimizers for G_1 are those configurations (including the TBP) whose center of mass is the origin.

1.2 The Main Result

Our main result verifies the phase-transition for 5 point energy minimization first observed in [**MKS**], in 1977, by T. W. Melnyk, O, Knop, and W. R. Smith. Define

$$15_+ = 15 + \frac{25}{512}. \tag{2}$$

Theorem 1.1 (Phase Transition) *There exists* $\mathfrak{w} \in (15, 15_+)$ *such that:*

1. *For* $s \in (0, \mathfrak{w})$ *the TBP is the unique minimizer for* R_s*.*
2. *For* $s = \mathfrak{w}$ *the TBP and some FP are the two minimizers for* R_s*.*
3. *For each* $s \in (\mathfrak{w}, 15_+)$ *some FP is the unique minimizer for* R_s*.*

Remark: I can also prove that the TBP minimizes all Fejes-Toth potentials for $s \in (-2, 0)$. I am leaving out the proof of this result so as to have a shorter exposition. See the end of §4.3 for further discussion.

1.3 Verification

To make the proof easier to verify, I have divided it up into 7 self-contained units. Each of 7 readers only needs to read between 8 and 16 pages of the document and then communicate to a central "team-leader" (say Reader 0) that the portion they have read is correct. Here is the breakdown.

Part 0, Assembly: This part of the proof deduces the Phase Transition Theorem from smaller components. Reader 0 need only read §2 and §3.

Part 1, Interpolation: We introduce potentials which we call *hybrid triples*:

$$a_0 G_{b_0}(r) + a_1 G_{b_1}(r) + a_2 G_{b_2}(r), \qquad a_k \in \boldsymbol{R}, \qquad G_b(r) = (4 - r^2)^b. \tag{3}$$

See §2.2 for the precise list and §4.3 for motivation. This part of the proof establishes Lemma 4.1 in §4.1, which says that if the TBP minimizes various collections of hybrid triples, then it also minimizes the Riesz potentials within certain ranges. Reader 1 need only read §2 and §4. This part of the proof involves a moderate amount of Java code which a competent programmer could reproduce in under a week. The results here are obvious from computer plots.

Part 2, Local Analysis: In this part of the proof, we show that there is an explicitly defined neighborhood Ω_0 of the TBP in which the TBP minimizes certain hybrid triples. See §2.4. Reader 2 need only read §2 and §5. This part of the proof involves a moderate amount of Mathematica code, which a competent programmer could reproduce in less than a day. Parts 1 and 2 combine to prove the Phase Transition Theorem for all (configuration, exponent) pairs in $\Omega_0 \times (0, 15_+]$.

Part 3, Symmetrization: Let $\boldsymbol{K}_4$ denote the set of 5-point configurations which have 4-fold dihedral symmetry. The dihedral symmetry group fixes one of the points. This part of the proof deals with a small open subset Υ of configurations near $\boldsymbol{K}_4$, and power law exponents $s \in [12, 16]$. See §2.5. Here we produce a retraction $\Upsilon \to \boldsymbol{K}_4$ and show that it is energy-decreasing on $\Upsilon - \boldsymbol{K}_4$. Reader 3 need only read §2 and §6. This part of the proof has a moderate amount of Mathematica code that a competent programmer could reproduce in a few days.

Part 4, Symmetric Configurations: This part of the proof treats configurations in $\Upsilon \cap \boldsymbol{K}_4$. Our work here combines with Part 3 to prove the Phase Transition Theorem for all (configuration, exponent) pairs in $\Upsilon \times [13, 15_+]$. This is the region where the phase transition actuallly occurs. Reader 4 need only read §2 and §7. This part of the proof involves a moderate amount of Mathematica and Java code that a competent programmer could reproduce in under a week.

Part 5, Energy Estimate: This part of the proof establishes an estimate which allows us to prove, just using finitely many calculations, that an entire open subset of the configuration space consists of configurations having larger F-energy than the TBP. Here F is one of the hybrid triple potentials of interest to us. This part of the proof is completely theoretical. There are no computer calculations involved. Reader 5 need only read §2, §8, and §9.

Part 6, The Big Calculation: Parts 1,2,3,4 of the proof wipe out all the pairs (configuration, exponent) in the sets $\Omega_0 \times (0, 15_+]$ and $\Upsilon \times [13, 15_+]$. The remaining pairs do not pose a serious threat to the TBP. This part of the proof uses the energy estimate from Part 5 to deal with all the remaining configurations. Reader 6 need only read §2, §8 and §10. This part of the

proof is the hardest to verify because it relies on a massive computer calculation. On the other hand, the computer calculation is just doing the same thing over and over again. I think that a good programmer could reproduce the entire program in two weeks.

Verification Summary: Here is what each of the 7 readers needs to read:

- Reader 0 (assembly): §2, §3. (10 pages total.)
- Reader 1 (interpolation): §2, §4. (12 pages total.)
- Reader 2 (local analysis): §2, §5. (8 pages total.)
- Reader 3 (symmetrization): §2, §6. (13 pages total.)
- Reader 4 (symmetric configs.) §2 , §7. (13 pages total.)
- Reader 5 (energy estimate): §2, §8, §9 (16 pages total.)
- Reader 6 (big calculation): §2, §8, §10. (10 pages total.)

Actually, not all readers have to read all of §2. At the beginning of §2 there is a finer breakdown of the topics.

Computer Code: The computer code is all written in Java and Mathematica. The Java code runs on Java 8 Update 201. I ran everything on a 2017 iMac Pro with a 3.2 GHz Intel Zeon W processor, running the Mojava operating system. The Mathematica code seems to run on all modern versions of Mathematica. One can download the computer code from

http://www.math.brown.edu/∼res/Papers/TBP.tar

The code is divided up to match the 7-part division discussed above. So, e.g., Reader 4 only needs to run Part 4 of the code.

1.4 Acknowledgements

I would like to thank Doug Hardin, Ed Saff, Javi Gomez-Serrano, and Stephen D. Miller for their helpful comments and encouragement.

2 Preliminaries

Reading Guide:

- Reader 0 (assembly) should read everything except §2.6.
- Reader 1 (interpolation) should read §2.2.
- Reader 2 (local analysis) should read §2.1, §2.2, §2.4
- Reader 3 (symmetrization) should read §2.1, §2.5, §2.6.
- Reader 4 (symmetric configs.) should read §2.1, §2.5, §2.6.
- Reader 5 (energy estimate) should read §2.1, §2.2, §2.3.
- Reader 6 (big calculation) should read everything except §2.6

2.1 Avatars

Let $S^2 \subset \boldsymbol{R}^3$ be the unit 2-sphere. *Stereographic projection* is the map $\Sigma : S^2 \to \boldsymbol{R}^2 \cup \infty$ given by the following formula.

$$\Sigma(x,y,z) = \Big(\frac{x}{1-z}, \frac{y}{1-z}\Big). \tag{4}$$

Here is the inverse map:

$$\Sigma^{-1}(x,y) = \Big(\frac{2x}{1+x^2+y^2}, \frac{2y}{1+x^2+y^2}, 1 - \frac{2}{1+x^2+y^2}\Big). \tag{5}$$

Σ^{-1} maps circles in $\boldsymbol{R}^2$ to circles in S^2 and $\Sigma^{-1}(\infty) = (0,0,1)$.

Stereographic projection gives us a correspondence between 5-point configurations on S^2 having $(0,0,1)$ as the last point and planar configurations:

$$\widehat{p}_0, \widehat{p}_1, \widehat{p}_2, \widehat{p}_3, (0,0,1) \in S^2 \iff p_0, p_1, p_2, p_3 \in \boldsymbol{R}^2, \qquad \widehat{p}_k = \Sigma^{-1}(p_k). \tag{6}$$

We call the planar configuration the *avatar* of the corresponding configuration in S^2. We call 2 avatars *isomorphic* if the corresponding 5-point configurations on S^2 are isometric.

We write $F(p_1, p_2, p_3, p_4)$ when we mean the F-potential of the corresponding 5-point configuration. If $\xi = (p_0, p_1, p_2, p_3)$ then we will write $F(\xi) = F(p_0, p_1, p_2, p_3)$.

We call a pair of points $\widehat{p}, \widehat{q} \in S^2$ *far* if $\|\widehat{p} - \widehat{q}\| \geq 4/\sqrt{5}$. Note that $(\widehat{p}, \widehat{q})$ is a far pair if and only if $(\widehat{q}, \widehat{p})$ is a far pair. Our rather strange definition has a more natural interpretation in terms of the avatars. If we rotate S^2 so that $\widehat{p} = (0, 0, 1)$ then $q = \Sigma(\widehat{q})$ lies in the disk of radius $1/2$ centered at the origin if and only if $(\widehat{p}, \widehat{q})$ is a far pair.

We say that a point in a 5-point configuration is *odd* or *even* according to the parity of the number of far pairs it makes with the other points in the configuration. Correspondingly, define the parity of the avatar to be the parity of the number of points which are contained in the closed disk of radius $1/2$ about the origin.

Lemma 2.1 *Every avatar is isomorphic to an even avatar.*

Proof: We form a graph by joining two points in a 5-point configuration by an edge if and only if they make a far pair. As for any graph, the sum of the degrees is even. Hence there is some vertex having even degree. When we rotate so that this vertex is $(0, 0, 1)$, the corresponding avatar is even. ♠

Figure 2.1 shows the two possible avatars (up to rotations) of the triangular bi-pyramid, first separately and then superimposed. We call the one on the left the *even avatar*, and the one in the middle the *odd avatar*. Let ξ_0 denote the even avatar. The points of ξ_0 are $(\pm 1, 0)$ and $(0, \pm\sqrt{3}/3)$.

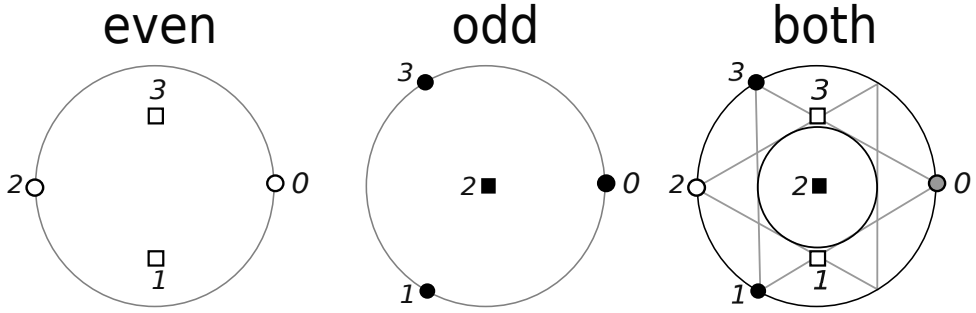

Figure 2.1: Even and odd avatars of the TBP.

When we superimpose the two avatars we see some extra geometric structure that is not relevant for our proof but worth mentioning. The two circles respectively have radii $1/2$ and 1 and the 6 segments shown are tangent to the inner one.

2.2 The Hybrid Triples

Now we introduce the potentials which we use in order to understand the Riesz potentials. Define

$$G_k(r) = (4 - r^2)^k. \tag{7}$$

Also define

$$\begin{aligned} G_5^{\flat} &= G_5 - 25G_1, \\ G_{10}^{\sharp\sharp} &= G_{10} + 28G_5 + 102G_2, \\ G_{10}^{\sharp} &= G_{10} + 13G_5 + 68G_2 \end{aligned} \tag{8}$$

I found these hybrid triples experimentally. They look rather arbitrary, but in fact they are close to the unique choices which function the right way in our proof.

2.3 The Big Domain

Given an avatar $\xi = (p_0, p_1, p_2, p_3)$, we write $p_k = (p_{k1}, p_{k2})$. We define a domain $\Omega \subset \boldsymbol{R}^7$ to be the set of avatars ξ satisfying the following conditions.

1. ξ is even.
2. $\|p_0\| \geq \max(\|p_1\|, \|p_2\|, \|p_3\|)$.
3. $p_{12} \leq p_{22} \leq p_{32}$ and $p_{22} \geq 0$.
4. $p_{01} \in (0, 2]$ and $p_{02} = 0$.
5. $p_j \in [-3/2, 3/2]^2$ for $j = 1, 2, 3$.

The Containment Theorem, stated in §3.2 and proved in §3.3, says that only configurations having avatars isomorphic to ones in Ω could be minimizers for the potentials we consider. So, Ω is our universe.

2.4 A Neighborhood of the TBP

Let ξ_0 denote the even avatar for the TBP. When we string out the points of ξ_0, we get $(1, \quad 0, -u, -1, 0, 0, u)$ where $u = \sqrt{3}/3$. The space indicates that we do not record $p_{02} = 0$. We let Ω_0 denote the cube of side-length 2^{-17} centered at ξ_0.

2.5 The Special Domain

We let $\Upsilon \subset (\boldsymbol{R}^2)^4$ denote those avatars p_0, p_1, p_2, p_3 such that

1. $\|p_0\| \geq \|p_k\|$ for $k = 1, 2, 3$.
2. $512p_0 \in [433, 498] \times [0, 0]$. (That is, $p_0 \in [433/512, 498/512] \times \{0\}$.)
3. $512p_1 \in [-16, 16] \times [-464, -349]$.
4. $512p_2 \in [-498, -400] \times [0, 24]$.
5. $512p_3 \in [-16, 16] \times [349, 464]$.

As we discussed above, Υ contains the avatars that compete with the TBP near the exponent $\mathfrak{w}$.

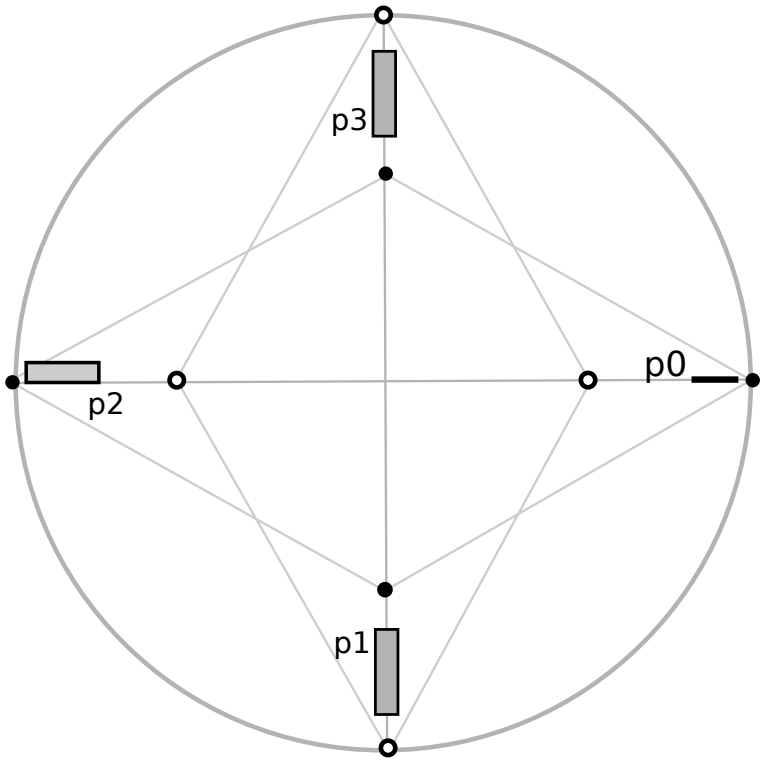

Figure 2.2: The sets defining Υ compared with two TBP avatars.

2.6 Polynomials and Exponential Sums

2.6.1 Positive Dominance

The works [**S2**] and [**S3**] give more details about positive dominance. Here I explain the basics. Let $P \in \boldsymbol{R}[x_1, ..., x_n]$ be a multivariable polynomial:

$$P = \sum_I c_I X^I, \qquad X^I = \prod_{i=1}^n x_i^{I_i}. \tag{9}$$

Given two multi-indices I and J, we write $I \preceq J$ if $I_i \leq J_i$ for all i. Define

$$P_J = \sum_{I \preceq J} c_I, \qquad P_\infty = \sum_I c_I. \tag{10}$$

We say that P is *weak positive dominant* (WPD) if $P_J \geq 0$ for all J and $P_\infty > 0$. We call P *positive dominant* if $P_J > 0$ for all J.

Lemma 2.2 (Weak Positive Dominance) *If P is weak positive dominant then $P > 0$ on $(0,1]^n$. If P is positive dominant then $P > 0$ on $[0,1]^n$.*

Proof: We prove the first statement. The second one has almost the same proof. Suppose $n = 1$. Let $P(x) = a_0 + a_1 x + ...$ Let $A_i = a_0 + ... + a_i$. The proof goes by induction on the degree of P. The case $\deg(P) = 0$ is obvious. Let $x \in (0,1]$. We have

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

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

Here $Q(x)$ is WPD and has degree $n-1$.

Now we consider the general case. We write

$$P = f_0 + f_1 x_k + ... + f_m x_k^m, \qquad f_j \in \boldsymbol{R}[x_1, ..., x_{n-1}]. \tag{11}$$

Since P is WBP so are the functions $P_j = f_0 + ... + f_j$. By induction on the number of variables, $P_j > 0$ on $(0,1]^{n-1}$. But then, when we arbitrarily set the first $n-1$ variables to values in $(0,1)$, the resulting polynomial in x_n is WPD. By the $n = 1$ case, this polynomial is positive for all $x_n \in (0,1]$. ♠

2.6.2 Polynomial Subdivision

2. Subdivision: Let $P \in \boldsymbol{R}[x_1, ..., x_n]$. For any x_j and $k \in \{0, 1\}$ we define

$$S_{x_j,k}(P)(x_1, ..., x_n) = P(x_1, ..., x_{j-1}, x_j^*, x_{j+1}, ..., x_n), \quad x_j^* = \frac{k}{2} + \frac{x_j}{2}. \quad (12)$$

If $S_{x_j,k}(P) > 0$ on $(0, 1]^n$ for $k = 0, 1$ then we also have $P > 0$ on $(0, 1]^n$.

2.6.3 Numerator Selection

If $f = f_1/f_2$ is a bounded rational function on $[0, 1]^n$, written in so that f_1, f_2 have no common factors, we always choose f_2 so that $f_2(1, ..., 1) > 0$. If we then show, one way or another, that $f_1 > 0$ on $(0, 1]^n$ we can conclude that $f_2 > 0$ on $(0, 1]^n$ as well. The point is that f_2 cannot change sign because then f blows up. But then we can conclude that $f > 0$ on $(0, 1]^n$. We write $\text{num}_+(f) = f_1$.

2.6.4 Exponential Sums

Lemma 2.3 (Convexity) *Suppose that $\alpha, \beta, \gamma \geq 0$ have the property that $\alpha + \beta \geq 2\gamma$. Then $\alpha^s + \beta^s \geq 2\gamma^s$ for all $s > 1$, with equality iff $\alpha = \beta = \gamma$.*

Proof: This is an exercise with Lagrange multipliers. ♠

Lemma 2.4 (Descartes) *Let $0 < r_1 \leq ... \leq r_n < 1$ be a sequence of positive numbers. Let $c_1, ..., c_n$ be a sequence of nonzero numbers and let $\sigma_1, ..., \sigma_n$ be the corresponding sequence of signs of these numbers. Define*

$$E(s) = \sum_{i=1}^{n} c_i \; r_i^s. \quad (13)$$

Let K denote the number of sign changes in the sign sequence. Then E changes sign at most K times on $\boldsymbol{R}$.

Proof: Suppose we have a counterexample. By continuity, perturbation, and taking mth roots, it suffices to consider a counterexample of the form $P(t) = \sum c_i t^{e_i}$ where $t = r^s$ and $r \in (0, 1)$ and $e_1 > ... > e_n \in \boldsymbol{N}$. As s ranges in r, the variable t ranges in $(0, \infty)$. But $P(t)$ changes sign at most K times on $(0, \infty)$ by Descartes' Rule of Signs. This gives us a contradiction. ♠

3 Proof Assembly

Reading Guide: This chapter is for Reader 0.

3.1 Interpolation

We use the notation from §2. Here is our main result about interpolation.

Theorem 3.1 (Interpolation) *Let T_0 be the TBP. Then*

1. *Suppose $s \in (0, 13]$ and T is any 5-point configuration. If we have $F(T_0) < F(T)$ for all $F = G_4, G_5, G_6, G_{10}^{\sharp\sharp}$ then $R_s(T_0) < R_s(T)$.*
2. *Suppose $s \in [13, 15_+]$ and T is any 5-point configuration. If we have $F(T_0) < F(T)$ for all $F = G_5^{\flat}, G_{10}^{\sharp}$ then $R_s(T_0) < R_s(T)$.*

3.2 The Containment Theorem

We prove the following theorem in the next section.

Theorem 3.2 (Containment) *Let $F = G_4, G_5^{\flat}, G_6, G_{10}^{\sharp}$. If ξ is not isomorphic to any avatar in Ω then $F(\xi_0) < F(\xi)$.*

Corollary 3.3 *If ξ is not isomorphic to any avatar in Ω and $F = G_5$ or $F = G_{10}^{\sharp\sharp}$, then $F(\xi_0) < F(\xi)$.*

Proof: Since ξ_0 (or indeed any configuration whose center of mass is the origin) is a global mininizer for G_1, we have $G_1(\xi_0) \le G_1(\xi)$. But then

$$G_5(\xi_0) = G_5^{\flat}(\xi_0) + 25G_1(\xi_0) < G_5^{\flat}(\xi) + 25G_1(\xi) = G_5(\xi).$$

The second inequality comes from the Containment Lemma.

We now know that $G_5(\xi_0) < G_5(\xi)$. But then

$$G_{10}^{\sharp\sharp\sharp}(\xi_0) = G_{10}^{\sharp}(\xi_0) + 15G_5(\xi_0) + 34G_2(\xi_0) <$$

$$G_{10}^{\sharp}(\xi) + 15G_5(\xi) + 34G_2(\xi) = G_{10}^{\sharp\sharp\sharp}(\xi).$$

The inequality follows from the Containment Theorem, the previous corollary, and Tumanov's result [**T**] that ξ_0 is a global minimizer for G_2. ♠

Corollary 3.4 *Let T_0 be the TBP and let T be a configuration that has no avatar isomorphic to one in Ω. Then $R_s(T_0) < R_s(T)$.*

Proof: This is an immediate corollary of the Interpolation Theorem, Containment Theorem, and Corollary 3.3. ♠

Corollary 3.4 tells us that we do not have to worry about configurations which do not have avatars isomorphic to ones in Ω. This means that, from the point of view of our proof, Ω is our universe. For the rest of the chapter, we will speak of the Phase Transition making a statement about $\Omega \times (0, 15_+]$. Each pair (ξ, s) is an avatar ξ at an exponent s. We will evaluate all our potentials directly on the avatars, with the understanding that in every case we are first applying inverse stereographic projection.

3.3 Proof of the Containment Theorem

Let ξ_0 the even avatar of the TBP. Let $[F] = F(\xi_0)$ for any F-potential. Since the TBP has 6 bonds of length $\sqrt{2}$, and 3 of length $\sqrt{3}$, and 1 of length $\sqrt{4}$, we have

$$[G_k] = 6 \times 2^k + 3. \tag{14}$$

Using this result, and the formulas for our energy functions, we compute

$$[G_4] = 99, \qquad [G_6] = 387, \qquad [G_5^\flat] = -180, \qquad [G_{10}^\sharp] = 10518. \tag{15}$$

Let $\xi = p_0, p_1, p_2, p_3$ some other avatar.

Lemma 3.5 *Let $F = G_6, G_5^\flat, G_{10}^\sharp$. If $\|p_0\| > 3/2$ then $[F] < F(\xi)$.*

Proof: Let τ_0 be the term in F corresponding to the pair (p_0, p_4). That is

$$\tau_0 = F(\|\Sigma^{-1}(p_0) - (0,0,1)\|). \tag{16}$$

When $\|p_j\| = 3/2$ we check using Equation 5 that $\tau_k = F(d)$. Here we have $d = 4/\sqrt{13}$. Also, each of our choices of F is monotone decreasing on $(0, d]$. So, if $\|p_0\| > 3/2$ then $\tau_0 > F(d)$.

Rather than work with $G_5^\flat$ we work with $G_5^* = G_5^\flat + 30$ so that all our functions are non-negative on $(0, 2]$. We have $[G_5^*] = 120$. Referring to the sequence $G_6, G_5^*, G_{10}^\sharp$, we have $\tau_0 > 450, 123, 26909$ if $\|p\| > 3/2$. These bounds respectively exceed. $[G_6], [G_5^*], [G_{10}^\sharp]$. ♠

Lemma 3.6 *If $F = G_4$ then $[F] < F(\xi)$ provided that either $\|p_0\| > 2$ or $\|p_0\|, \|p_j\| > 3/2$ for some $j = 1, 2, 3$.*

Proof: We keep the same notation from the previous result, and define τ_j just as we defined τ_0. When $\|p_0\| > 2$ we have $\tau_0 > 104 > [G_4]$. When $\|p_0\|, \|p_i\| > 3/2$ we have $\tau_0 + \tau_j > 58 + 58 > [G_4]$. ♠

Assume first that $F \neq G_4$. Assume ξ is a minimizer for F. As we have already discussed in the definition of even and odd avatars, we normalize so that ξ is even. Reordering p_0, p_1, p_2, p_3 and rotating, about the origin, we make $\|p_0\| \geq \|p_i\|$ for $i = 1, 2, 3$ and we move p_0 into the positive x-axis. Reflecting in the x-axis if necessary and reordering the points p_1, p_2, p_3 if necessary, we arrange that $p_{12} \leq p_{22} \leq p_{32}$ and $p_{22} \geq 0$. Lemma 3.5 tells us that $\|p_0\| \leq 3/2$, and this gives us $\|p_i\| \leq 3/2$ for $i = 1, 2, 3$. In particular $p_j \in [-3/2, 3/2]^2$ for $j = 0, 1, 2, 3$. We have also arranged that $p_{02} = 0$.

The case of $F = G_4$ follows from Lemma 3.6 just as the other cases follow from Lemma 3.5. This completes the proof of the Containment Theorem.

3.4 Local Analysis

Recall that Σ^{-1} is inverse stereographic projection. Here is the main local result we prove.

Theorem 3.7 (Local Convexity) *For $F = G_4, G_6, G_5^\flat, G_{10}^\sharp$, the Hessian of $F \circ \Sigma^{-1}$ is positive definite at every point of Ω_0.*

Corollary 3.8 *Let F be any of $G_4, G_5^\flat, G_5, G_6, G_{10}^\sharp, G_{10}^{\sharp\sharp}$. Then ξ_0, the TBP avatar, is the unique F-energy minimizer inside Ω.*

Proof: Let F be any of the functions from the Local Convexity Theorem. Let $\xi \in \Omega_0$ be other than ξ_0. The Local Convexity Theorem combines with the vanishing gradient to show that the restriction of $F \circ \Sigma_{-1}$ to the line segment γ joining ξ_0 to ξ is convex and has 0 derivative at ξ_0. Hence $F(\xi) > F(\xi_0)$. It remains to deal with $F = G_5$ and $F = G_{10}^{\sharp\sharp}$. The same argument as in Corollary 3.3 deals with G_5 and $G_{10}^{\sharp\sharp}$. ♠

Combining Corollary 3.8 with the Interpolation Theorem, we get:

Corollary 3.9 *The Phase Transition Theorem is true for for $\Omega_0 \times (0, 15_+]$. In this region, there is no phase transition: The TBP is always best.*

3.5 Symmetrization

In this section we deal directly with the Riesz potentials. We deal with the configurations in the region Υ from §2.5. Let $\boldsymbol{K}_4$ denote the set of avatars which are invariant under reflections in the coordinate axes. Recall that Υ is our special domain from §2.5. We describe a symmetrization operation which maps Υ into $\boldsymbol{K}_4$. Let (p_0, p_1, p_2, p_3) be an avatar with $p_0 \neq p_2$. Define

$$-p_2^* = p_0^* = (x, 0), \quad -p_1^* = p_3^* = (0, y), \quad x = \frac{\|p_0 - p_2\|}{2}, \quad y = \frac{\|\pi_{02}(p_1 - p_3)\|}{2}. \tag{17}$$

Here π_{02} is the projection onto the subspace perpendicular to $p_0 - p_2$. The avatar $(p_1^*, p_2^*, p_3^*, p_4^*)$ lies in $\boldsymbol{K}_4$. Note that our operation fixes avatars in $\boldsymbol{K}_4$.

Theorem 3.10 (Symmetrization) *Let $(p_0, p_1, p_2, p_3) \in \Upsilon - \boldsymbol{K}_4$. Then $R_s(p_0^*, p_1^*, p_2^*, p_3^*) < R_s(p_0, p_1, p_2, p_3)$ when $s \geq 12$.*

3.6 Symmetric Configurations

Let Ψ_4 denote the set of avatars of the form

$$(x, 0), \qquad (0, -y), \qquad (-x, 0), \qquad (0, y), \qquad 64(x, y) \in [43, 64]^2. \tag{18}$$

We have $\Upsilon \cap \boldsymbol{K}_4 \subset \Psi_4$. We identify Ψ_4 (and its special subsets below) as a subset of $\boldsymbol{R}^2$. Thus (x, y) names the configuration in Equation 18.

Let $\Psi_4^\sharp \subset \Psi_4$ denote the subset with

$$64(x, y) \in [55, 56]^2. \tag{19}$$

Let $\Psi_8 \subset \Psi_4$ and $\Psi_8^\sharp \subset \Psi_4^\sharp$ denote the diagonals, where $x = y$.

Define

$$\sigma(x, y) = (z, z), \qquad z = \frac{x + y + (x - y)^2}{2}. \tag{20}$$

This maps $\Psi_4^\sharp$ into Ψ_8.

Note that σ is the identity on Ψ_8. Here are the three results we prove in this section. All these results are about low dimensional subspaces.

Theorem 3.11 (Critical I) *$R_s(\sigma(p)) < R_s(p)$ for*

$$(p, s) \in (\Psi_4^\sharp - \Psi_8^\sharp) \times [14, 16].$$

Theorem 3.12 (Critical II) *$R_s(\xi_0) < R_s(\xi)$ for*

$$(\xi, s) \in (\Psi_4 \times [13, 15]) \quad \cup \quad ((\Psi_4 - \Psi_4^\sharp) \times [15, 15_+]).$$

Theorem 3.13 (Critical III) *There exist $\mathfrak{w} \in (15, 15_+)$ such that*

1. *$R_s(\xi_0) < R_s(\xi)$ for all $(\xi, s) \in \Psi_8^\sharp \times [15, \mathfrak{w})$.*
2. *$R_s(\xi_0) > R_s(\xi)$ for some (fixed) $\xi \in \Psi_8^\sharp$ and all $s \in (\mathfrak{w}, 15_+)$*
3. *R_s is uniquely minimized on $\Upsilon_8^\sharp$ for all $s \in (\mathfrak{w}, 15_+]$.*

Corollary 3.14 *The Phase Transition Theorem is true for $\Upsilon \times [13, 15_+]$.*

Proof: We first show that if $\xi \in \Upsilon$ and $s \in [13, \mathfrak{w})$ then $R_s(\xi_0) < R_s(\xi)$. We argue by contradiction. Suppose $R_s(\xi) < R_s(\xi_0)$. By the Symmetrization Theorem, it suffices to consider the case when $\xi \in \Upsilon \cap \boldsymbol{K}_4 \subset \Upsilon_4$. The Critical Theorem II tells us that $s \in [15, 15_+]$ and $\xi \in \Upsilon_4^\sharp$. By the Critical Theorem I, we can find some new $\xi' \in \Upsilon_8^\sharp$ such that $R_s(\xi') < R_s(\xi) < R_s(\xi_0)$. This contradicts Statement 1 of the Critical Theorem III.

Now suppose that that $s > \mathfrak{w}$. Statement 2 of the Critical Theorem III tells us that some FP minimizes the R_s potential.

Now we consider the case when $s = \mathfrak{w}$. By continuity both the TBP and some FP minimize R_s. Suppose ξ_1 and ξ_2 are both satisfy

$$R_s(\xi_1) = R_s(\xi_2) = R_s(\xi_0).$$

Note that both ξ_1 and ξ_2 are FPs which minimizer R_s. Statement 3 of the Critical Theorem III now says that at most one of ξ_1, ξ_2 can belong to $\Upsilon_8^\sharp$. Suppose $\xi_2 \notin \Upsilon_8^\sharp$.

The Critical Theorem II says that $\xi_2 \in \Upsilon_4^\sharp - \Upsilon_8^\sharp$. But then the Critical Theorem I says that there is some $\xi_2' \in \Upsilon_8^\sharp$ with $R_s(\xi_2') < R_s(\xi_2)$. This contradicts the fact that ξ_2 is an FP which minimizes R_s. Hence, when $s = \mathfrak{w}$, there is a unique FP in Υ that ties with the TBP. ♠

Corollaries 3.9 and 3.14 reduce us to showing that the Phase Transition Theorem is true on

$$(\Omega - \Omega_0) \times (0, 13] \quad \cup \quad (\Omega - \Omega_0 - \Upsilon) \times [13, 15_+]. \tag{21}$$

We handle this with a divide-and-conquer calculation.

3.7 Big Calculation

Theorem 3.15 (Calculation) *The following is true.*

1. *The TBP is the unique minimizer for $G_4, G_5^\flat, G_6$ amongst 5-point configurations which have avatars in $\Omega - \Omega_0$.*
2. *The TBP is the unique minimizer for $G_{10}^\sharp$ among 5-point configurations which have avatars in $\Omega - \Omega_0 - \Upsilon$.*
3. *The TBP is the unique minimizer for $G_{10}^{\sharp\sharp}$ among 5-point configurations which have avatars in Υ.*

The Calculation Theorem does not quite line up with our Interpolation Theorem. Let us now get the two results in line exactly.

Corollary 3.16 *The following is true.*

1. *The TBP is the unique minimizer for $G_4, G_5^\flat, G_6, G_{10}^{\sharp\sharp}$ among configurations having avatars in $\Omega - \Omega_0$.*
2. *The TBP is the unique minimizer for $G_{10}^\sharp$ among 5-point configurations having avatars in $\Omega - \Omega_0 - \Upsilon$.*

Proof: The only point that is not obvious from the Calculation Theorem is the statement about $G_{10}^{\sharp\sharp}$. Since the TBP is a global minimizer for G_1 and (uniquely so) for $G_5^\flat$ on $\Omega - \Omega_0$, we see that the TBP is the unique minimizer for G_5 on $\Omega - \Omega_0$. Since the TBP is the unique minimizer for $G_{10}^\sharp$ and G_5 and (by Tumanov's result [**T**]) G_2 on $\Omega - \Omega_0 - \Upsilon$ we see that the TBP is the unique minimizer for $G_{10}^{\sharp\sharp}$ on $\Omega - \Omega_0 - \Upsilon$. This combines with Statement 3 of the Calculation Theorem to show that the TBP is the unique minimizer for $G_{10}^{\sharp\sharp}$ on $\Omega - \Omega_0$. ♠

Combining Corollary 3.16 with the Interpolation Theorem, we see that the Phase Transition Theorem is true on the domain in Equation 21. This completes the proof of the Phase Transition Theorem.

4 The Interpolation Theorem

Reading Guide: This chapter is for Reader 1.

4.1 Main Result

Recall that $15_+ = 15 + \frac{25}{512}$. We let $R_s(T)$ be the Riesz s-potential of a configuration T. Referring to Equations 7 and 8, we define

$$P_1 = (G_4, G_6), \qquad P_2 = (G_5, G_{10}^{\sharp\sharp}), \qquad P_3 = (G_5^{\flat}, G_{10}^{\sharp}), \tag{22}$$

$$I_1 = (0, 6], \qquad I_2 = [6, 13], \qquad I_3 = [13, 15_+]. \tag{23}$$

We say that a pair (Γ_3, Γ_4) of functions *forces* the interval I if the following is true: If T is another 5-point configuration such that $\Gamma_k(T_0) < \Gamma_k(T)$ for $k = 3, 4$ then $R_s(T_0) < R_s(T)$ for all $s \in I$.

In this chapter we prove the following result, which immediately the Interpolation Theorem from §3.1

Lemma 4.1 (A) *The following is true.*

1. *The pair* (G_4, G_6) *forces* $(0, 6]$.
2. *The pair* $(G_5, G_{10}^{\sharp\sharp})$ *forces* $[6, 13]$.
3. *The pair* $(G_5^{\flat}, G_{10}^{\sharp})$ *forces* $[13, 15_+]$.

4.2 Reduction to Smaller Results

We say that a pair of functions (Γ_3, Γ_4) *specially forces* $s > 0$ if there are constants $a_0, ..., a_4$ (depending on s) such that

$$\Lambda_s = a_0 + a_1 G_1 + a_2 G_2 + a_3 \Gamma_3 + a_4 \Gamma_4, \tag{24}$$

1. $\Lambda_s(x) = R_s(x)$ for $x = \sqrt{2}, \sqrt{3}, \sqrt{4}$.
2. $a_1, a_2, a_3, a_4 > 0$.
3. $\Lambda_s(x) \le R_s(x)$ for all $x \in (0, 2]$.

We say that (Γ_3, Γ_4) *specially forces* the interval I if this pair specially forces all $s \in I$.

Lemma 4.2 (A1) *If* (Γ_3, Γ_4) *specially forces* I *then* Γ *forces* I*.*

Proof: Let T_0 be the TBP and let T be some other 5-point configuration. We simplify the notation and write $F(T) = \mathcal{E}_F(T)$. We assume

$$\Gamma_j(T_0) < \Gamma_j(T)$$

for $j = 3, 4$ and we want to show that that $R_s(T_0) < R_s(T)$ for all $s \in I$. It is well known that $\Gamma_1(T_0) \le \Gamma_1(T)$ and, by Tumanov's result [**T**], $\Gamma_2(T_0) \le \Gamma_2(T)$. Let $a_j = a_j(s)$ for $s \in I$. The quantities $\sqrt{2}, \sqrt{3}, \sqrt{4}$ are the distances which appear between pairs of points in T_0. Therefore $\Lambda_s(T_0) = R_s(T_0)$. But then

$$R_s(T) \ge \Lambda_s(T) = a_0 + \sum_{j=1}^{4} a_j \Gamma_j(T) > a_0 + \sum_{j=1}^{4} a_j \Gamma_j(T_0) = \Lambda_s(T_0) = R_s(T_0).$$

This completes the proof. ♠

Lemma 4.3 (A2) *For each* $i = 1, 2, 3$ *the pair* P_i *specially forces* I_i*.*

Lemma A is an immediate consequence of Lemma A1 and Lemma A2. It remains to prove Lemma A2.

4.3 Discussion

Before launching into the proof, let me explain what made me search for these results and how I found them. Tumanov remarks in [**T**] (using somewhat different language) that the pair (G_3, G_5) forces the parameter interval $(0, 2]$. He did not offer a proof but eventually I found one on my own. By finding the explicit equations for the coefficients, I saw that (G_3, G_5) specially forces $(0, 2]$. Finding the coefficients is just a linear algebra problem for each s.

Wanting to prove that the TBP is the minimizer for a larger range of exponents, I eventually saw that (G_4, G_6) specially forces $(0, 6]$. This is the case $i = 1$ of Lemma A2.

Our luck somewhat runs out for G_k when $k \ge 7$. The TBP is not the global minimizer for $G_7, G_8, ...$ If we want to use expressions like G_7, etc, we need to average it with G_k for smaller values of k to give the TBP a chance

of being the minimizer. I experimented with expressions $a_1 G_{b_1} + a_2 G_{b_2}$ and wasn't having much luck. So, I then broadened the search to the hybrid triples. This worked quite well.

My computer program allows the reader to specify a hybrid triple, solve for the coefficients needed for Property 1, and then check visually whether Properties 2 and 3 hold. After fooling around for a while I hit on the specific expressions that appear in Lemma A2. Once I got the expressions, it was a matter of computer algebra to prove that the plots on my computer program are indeed an accurate reflection of mathematical reality.

The proof of Lemma A2 relies on interval arithmetic calculations in Java. The reader can download the code and see that it works. I think that it would take a competent programmer less than a week to reproduce the code. Also, I give explicit expressions for everything (with computer plots), so a really energetic reader could find their own ways to verify that the plots are accurate reflections of mathematical reality.

As an aside, Tumanov also observes that the pair (Γ_3, Γ_5) forces the interval $(-2, 0)$ if we use the Fejes-Toth potentials. (A proof similar to the ones given in this chapter would establish this fact.) This is how I prove that the TBP minimizes the Fejes-Toth potentials for all $s \in (-2, 0)$.

4.4 Techniques of Proof

In our proofs below, we will need to deal with expressions of the form

$$F(s) = \sum c_i s^{t_i} b_i^{s/2}, \tag{25}$$

where $b_i, c_i \in \boldsymbol{Q}$ and $t_i \in \boldsymbol{Z}$ and $b_i > 0$. For each summand we compute a floating point value, x_i. We then consider the floor and ceiling of $2^{32} x_i$ and divide by 2^{32}. This gives us rational numbers x_{i0} and x_{i1} such that $x_{i0} \le x_i \le x_{i1}$. Since we don't want to trust floating point operations without proof, we formally check these inequalities with what we call the *expanding out method*.

Expanding Out Method: Suppose we want to establish an inequality like $(\frac{a}{b})^{\frac{p}{q}} < \frac{c}{d}$, where every number involved is a positive integer. This inequality is true iff $b^p c^q - a^p d^q > 0$. We check this using exact integer arithmetic. The same idea works with $(>)$ in place of $(<)$.

To check the positivity of F on some interval $[s_0, s_1]$ we produce, for each term, the 4 rationals $x_{i00}, x_{i10}, x_{i01}, x_{i01}$. Where x_{ijk} is the approximation computed with respect to s_k. We then let y_i be the minimum of these expressions. The sum $\sum y_i$ is a lower bound for Equation 25 for all $s \in [s_0, s_1]$. On any interval exponent I where we want to show that Equation 25 is positive, we pick the smallest dyadic interval $[0, 2^k]$ that contains I and then run the following subdivision algorithm.

1. Start with a list L of intervals. Initially $L = \{[0, 2^k]\}$.
2. If L is empty, then **HALT**. Otherwise let Q be the last member of L.
3. If either $Q \cap I = \emptyset$ or the method above shows that Equation 25 is positive on Q we delete Q from L and go to Step 2.
4. Otherwise we delete Q from L and append to L the 2 intervals obtained by cutting Q in half. Then we go to to Step 2.

If this algorithm halts then it constitutes a proof that $F(s) > 0$ for all $s \in I$.

Here is another tool we will use below in the proof. This kind of result is discussed in much more generality in §2.6.1. All we need here is the single variable case and so we give a short and self-contained account.

Lemma 4.4 (Positive Dominance) *A real polynomial $a_0 + a_1 t + ... a_n t^n$ is positive on $[0, 1]$ provided that the sums $a_0, a_0 + a_1, a_0 + a_1 + a_2, ..., a_0 + ... + a_n$ are all positive.*

Proof: Call the polynomial P. We do induction on the degree of P. For $x \in [0, 1]$ we have

$$P(x) \geq xQ(x), \qquad Q(x) = (a_0 + a_1) + a_2 x + a_3 x^2 ...$$

The polynomial $Q(x)$ satisfies the same hypotheses as $P(x)$ concerning the coefficients, so $Q(x) > 0$. Hence $P(x) > 0$ for $x \in (0, 1]$. Finally $P(0) = a_0 > 0$. ♠

4.5 Reduction of Lemma A2

Referring to Equation 24 we solve the equations

$$\Lambda_s(\sqrt{m}) = R_s(\sqrt{m}), \quad m = 2,3,4, \qquad \Lambda_s'(\sqrt{m}) = R_s'(\sqrt{m}), \quad m = 2,3. \tag{26}$$

Here f' denotes the derivative of f, a function defined on $(0,2]$. We don't need to constrain $f'(2)$. For each s this gives us a linear system with 5 variables and 5 equations. In all cases, our solutions have the following structure

$$(a_0, a_1, a_2, a_3, a_4) = M(2^{-s/2}, 3^{-s/2}, 4^{-s/2}, s2^{-s/2}, s3^{-s/2}) \tag{27}$$

We will list M below for each of the 3 cases.

Lemma 4.5 (A21) *For each $i = 1,2,3$ the following is true. When M is defined relative to the pair P_i then the coefficients a_1, a_2, a_3, a_4 are positive functions on the interval I_i.*

We want to see that the function

$$H_s = 1 - \frac{\Lambda_s}{R_s}. \tag{28}$$

takes its minima at $r = \sqrt{2}, \sqrt{3}$ on $(0,2]$. Differentiating with respect to $r \in (0,2]$ we have

$$H_s'(r) = r^{s-1}(s\Lambda_s(r) + r\Lambda_s'(r)). \tag{29}$$

Using the general equation $rG_k'(r) = 2kG_k(r) - 8kG_{k-1}(r)$, we see that

$$\psi_s = s\Lambda_s(r) + r\Lambda_s'(r) \tag{30}$$

is a polynomial in $t = 4 - r^2$.

Lemma 4.6 (A22) *For each choice P_j and each $s \in I_j$ the following is true. The function ψ_s has 4 simple roots in $[0,4]$. Two of the roots are 1 and 2 and the other two respectively lie in $(0,1)$ and $(1,2)$.*

Let us deduce Lemma A2. Our construction and Lemma A21 immediately take care of Conditions 1 and 2 of special forcing. Condition 3: The roots of ψ_s in $[0,4)$ are in bijection with the roots of H_s' in $(0,2]$ and their nature (min, max, simple) is preserved under the bijection. We check for one parameter in each of the three cases that the roots 1 and 2 correspond to local minima and the other two roots correspond to local maxima. Since these roots remain simple for all s in the relevant interval, the nature of the roots cannot change as s varies. Hence H_s has exactly 2 local minima in $(0,2]$, at $r = \sqrt{2}, \sqrt{3}$. But then $H_s \geq 0$ on $(0,2]$. This completes the proof.

4.6 Data and Plots

Referring to Equation 27, we list out the matrix M in each of the three cases and also show computer plots. The reader can interact with these plots and see others like (and unlike) them using our computer software.

Here is Case 1.

$$M = \frac{1}{792}\begin{bmatrix} 0 & 0 & 792 & 0 & 0 \\ 792 & 1152 & -1944 & -54 & -288 \\ -1254 & -96 & 1350 & 87 & 376 \\ 528 & -312 & -216 & -39 & -98 \\ -66 & 48 & 18 & 6 & 10 \end{bmatrix} \tag{31}$$

The left side of Figure 4.1 shows a graph of

$$80a_1, \qquad 200a_2, \qquad 2000a_3, \qquad 10000a_4,$$

considered as functions of the exponent s. Here a_1, a_2, a_3, a_4 are colored darkest to lightest. The completely unimportant positive multipliers are present so that we get a nice picture. On the left side of Figure 4.1, the thick vertical segments are $s = 0, 1, 2, 3, 4, 5, 6$.

It turns out that a_3 goes negative between 6 and 6.1, so the interval $(0, 6]$ is fairly near to the maximal interval of positivity.

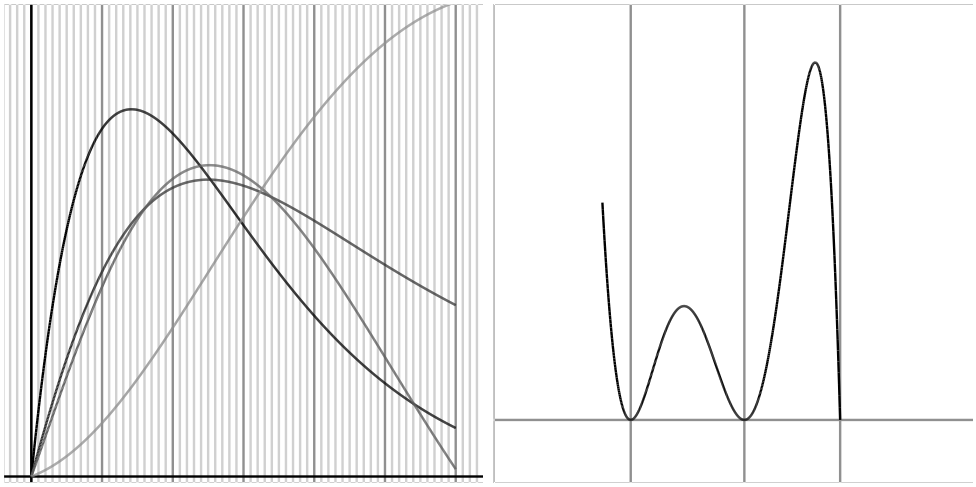

Figure 4.1: Plots for Case 1.

We cannot directly apply our positivity algorithm to Case 1 because this algorithm only works for functions which have uniform positive lower bounds. We will deal with Case 1 below.

Here is Case 2.

$$\frac{1}{368536}\begin{bmatrix} 0 & 0 & 268536 & 0 & 0 \\ 88440 & 503040 & -591480 & -4254 & -65728 \\ -77586 & -249648 & 327234 & 2361 & 65896 \\ 41808 & -19440 & -22368 & -2430 & -9076 \\ -402 & 264 & 138 & 33 & 68 \end{bmatrix} \tag{32}$$

Figure 4.2 does for Case 2 what Figure 4.1 does for Case 1. This time the left hand side plots

$$500a_1 \qquad 500a_2, \qquad 5000a_3, \qquad 500000a_4.$$

for $s \in [6, 13]$. The think vertical segments are $s = 6, 7, 8, 9, 10, 11, 12, 13$.

The coefficients a_1, a_2, a_3 go negative for s just a tiny bit larger than 13. I worked hard to find the function $\Gamma_4 = G_{10} + 28G_5 + 102G_2$ so that we could get all the way up to $s = 13$. The right hand side shows a plot of H_{10} from $r = 5/4$ to $r = 2$.

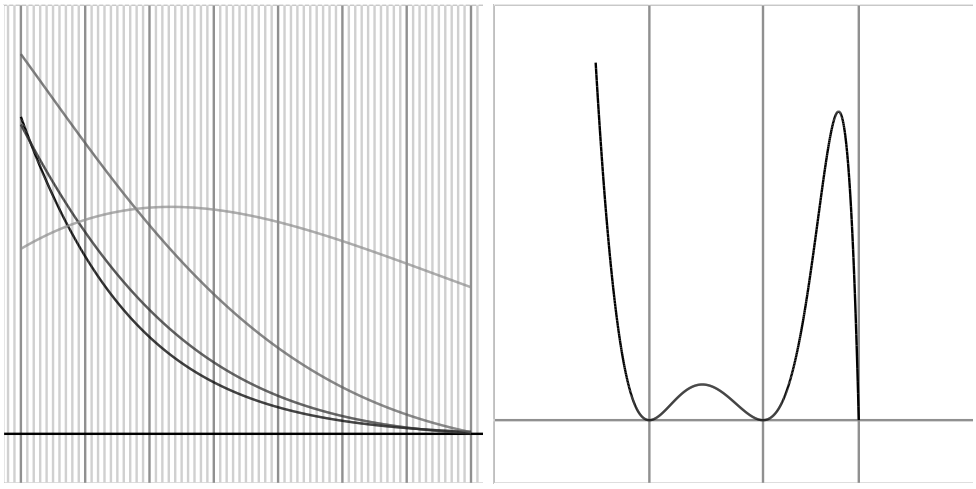

Figure 4.2: Plot for Case 2.

For Case 2 we run the positivity algorithm and show that for $k = 1, 2, 3, 4$ the function $a_k(s)$ is positive on $[6, 13]$, as the plot indicates.

Here is Case 3.

$$\frac{1}{368536}\begin{bmatrix} 0 & 0 & 268536 & 0 & 0 & 0 \\ 982890 & 116040 & -1098930 & -52629 & -267128 & 0 \\ -91254 & -240672 & 331926 & 3483 & 68208 & 0 \\ 35778 & -15480 & -20298 & -1935 & -8056 & 0 \\ -402 & 264 & 138 & 33 & 68 & 0 \end{bmatrix} \tag{33}$$

This matrix is quite similar to the one in the previous case, because we are essentially still taking combinations of $G_0, G_1, G_2, G_5, G_{10}$. We are just grouping the functions differently. Figure 3.3 does for Case 3 what Figure 3.2 does for Case 2. This time we plot

$$500a_1 \qquad 15000a_2, \qquad 20000a_3, \qquad 500000a_4,$$

for $s \in [13, 16]$. The thick vertical segments are $s = 13, 14, 15$.

The coefficients a_1, a_2, a_3 go negative for s just a tiny bit larger than 15.05. In particular, everything up to and including our cutoff of $5 + 25/512$ is covered. The right hand side shows a plot of H_{14} from $r = 5/4$ to $r = 2$.

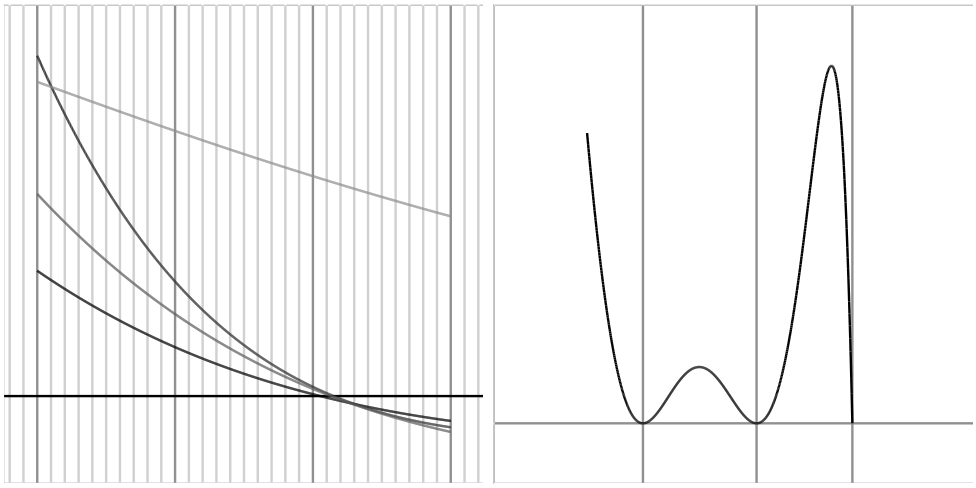

Figure 4.3: Plot for Case 3.

For Case 3 we run the positivity algorithm and show that for $k = 1, 2, 3, 4$ the function $a_k(s)$ is positive on $[13, 15_+]$, as the plot indicates.

4.7 Case 1 of Lemma A21

Before we launch into Case 1, we add two quantities we test, namely $\psi_s(0)$ and $\psi_s(4)$. We have

$$11\psi_s(0) = \begin{bmatrix} -88 \\ -128 \\ +216 \\ +6 \\ +32 \\ +11 \end{bmatrix} \cdot \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2} \\ s4^{-s/2} \end{bmatrix}, \qquad \frac{11}{s}\psi_s(4) = \begin{bmatrix} -2112 \\ +1664 \\ +459 \\ +219 \\ 288 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 2^{-s/2} \\ 3^{-s/2} \\ 4^{-s/2} \\ s2^{-s/2} \\ s3^{-s/2} \\ s4^{-s/2} \end{bmatrix}$$

In other words, these quantities have the same form as the functions $a_j(s)$ for $j = 1, 2, 3, 4$. We run the positivity algorithm and show that all 6 quantities are positive on $[1/4, 6]$.

Now we deal with the interval $(0, 1/4]$. Note that

$$\sup_{m=2,3,4} \sup_{s\in[0,1]} \left| \frac{\partial^6}{\partial s^6} m^{-s/2} \right| < \frac{1}{8}. \tag{34}$$

All our (scaled) expressions have the form $Y \cdot V(s)$,

$$V(s) = (2^{-s/2}, 3^{-s/2}, 4^{-s/2}, s2^{-s/2}, s3^{-s/2}, s4^{-s/2}).$$

For an integer vector Y. Moreover the sum of the absolute values of the coefficients in each of the Y vectors is at most 5000. This means that, when we take the 5th order Taylor series expansion for $Y \cdot V(s)$, the error term is at most

$$5000 \times \frac{1}{8} \times \frac{1}{6!} < 1.$$

We compute each Taylor series, set all non-leading positive terms to 0, and crudely round down the other terms:

$$\begin{aligned}
792a_1(s) &: \quad 98s - 69s^2 + 0s^3 - 6s^4 + 0s^5 - 1s^6 \\
792a_2(s) &: \quad 14s - 3s^2 - 2s^3 + 0s^4 - 1s^5 - 1s^6. \\
792a_3(s) &: \quad 1s + 0s^2 - 1s^3 + 0s^4 + 0s^5 - 1s^6. \\
792a_4(s) &: \quad .03s + 0s^2 + 0s^3 - .01s^4 + 0s^5 - 1s^6. \\
11\psi_s(0) &: \quad .08s + 0s^2 - .02s^3 + 0s^4 - .01s^5 - 1s^6. \\
(11/s)\psi_s(4) &: \quad 11 + 0s + 0s^2 - 1s^3 - 1s^4 + 0s^5 - 1s^6.
\end{aligned}$$

These under-approximations are all easily seen to be positive on $(0, 1/4]$. My computer code does these calculations rigorously with interval arithmetic, but it hardly seems necessary.

4.8 Proof of Lemma A22

Case 1: In Case 1 we compute that

$$\psi_s(t) = t^6 - \frac{48}{12+s}t^5 + ... \tag{35}$$

We don't care about the other terms. Since ψ_s has degree 6 we conclude that ψ_s has at most $N = 6$ roots, counting multiplicity. By construction $H_s(\sqrt{m}) = H_s'(\sqrt{m}) = 0$ for $m = 2, 3$ and $H_s(\sqrt{4}) = 0$. This means that H_s has extrema at $r_2 = \sqrt{2}$ and $r_3 = \sqrt{3}$ and at points $r_{23} \in (\sqrt{2}, \sqrt{3})$ and $r_{34} \in (\sqrt{3}, \sqrt{4})$. Correspondingly ψ_s has roots $t_1 = 1$ and $t_2 = 2$ and $t_{01} \in (0,1)$ and $t_{12} \in (1,2)$. The sum of all the roots of ψ_s is $48/(12+s) < 4$. Since $t_1 + t_2 + t_{01} + t_{12} > 4$ we see that not all roots can be positive. Hence $N < 6$. Since ψ_s is positive at $t = 0, 4$ we see that N is even. Hence $N = 4$. This means that the only roots of ψ_s in $(0,4)$ are the 4 roots we already know about. Since these roots are distinct, they are simple roots.

Cases 2 and 3: First of all, the functions H_s are the same in Cases 2 and 3. This is not just a computational accident. In both cases we are building H_s from the functions G_1, G_2, G_5, G_{10}. So, we combine Cases 2 and 3 by proving that the common polynomial ψ_s just has 4 roots for each $s \in [6, 16]$. I will describe a proof which took me quite a lot of experimentation to find.

The same analysis as in Case 1 shows that ψ_s has roots at $1, 2$, and in $(0,1)$ and in $(1,2)$. We just want to see that there are no other roots.

We can factor ψ_s as $(t-1)(t-2)\beta_s$ where β_s is a degree 8 polynomial. Taking derivatives with respect to t, we notice that

1. $\gamma_s = 268536 \times 12^{s/2} \times (\beta_s'' - \beta_s')$ is positive for $s \times t \in [6,16] \times [0,4]$.
2. $-\beta_s'(0) > 0$ for all $s \in [6,16]$.
3. $\beta_s'(4) > 0$ for all $s \in [6,16]$.

Statement 1 shows in particular that β_s' never has a double root. This combines with Statements 2 and 3 to show that the number of roots of β_s' in $[0,4]$ is independent of $s \in [6,16]$. We check explicitly that β_6' has only one root in $[0,4]$. Hence β_s' always has just one root. But this means that β_s has at most 2 roots in $[0,4]$. This, in turn, means that ψ_s has at most 4 roots in $[0,4]$. This completes the proof modulo the 3 statements.

Now we establish the 3 statements. We first give a formula for γ_s. Define matrices M_3, M_4, M_6 respectively as:

$$\begin{bmatrix} -546840 & -1800480 & 99720 & -397440 & -234600 & -33120 & 173880 & -22080 \\ 18366 & 17112 & 80766 & 24288 & 18630 & 11592 & 4830 & -1104 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -345600 & -1576320 & -509760 & -760320 & -448800 & -63360 & 332640 & -42240 \\ -199296 & -698784 & 75216 & -149376 & -79960 & 5856 & 94920 & -12992 \\ 7104 & 8432 & 33960 & 11968 & 9180 & 5712 & 2380 & -544 \end{bmatrix}$$

$$\begin{bmatrix} 892440 & 3376800 & 410040 & 1157760 & 683400 & 96480 & -506520 & 64320 \\ -73350 & -246888 & -228942 & -165792 & -110370 & -41688 & 27510 & -2064 \\ 1473 & 4092 & 10557 & 5808 & 4455 & 2772 & 1155 & -264 \end{bmatrix}$$

Define 3 polynomials P_3, P_4, P_6 by the formula:

$$P_k(s,t) = (1, s, s^2) \cdot M_k \cdot (1, ..., t^7) = \sum_{i=0}^{2} \sum_{j=0}^{7} (M_k)_{ij} s^i t^j, \qquad k = 3, 4, 6. \quad (36)$$

We have

$$\gamma = P_3 3^{s/2} + P_4 4^{s/2} + P_6 6^{s/2}. \quad (37)$$

To check the positivity of γ_s we check that each of the 16 functions

$$\gamma_s(v/4 + 1/4) = a_{v,0} + a_{v,1} t + ... a_{v,7} t^7 \quad (38)$$

satisfies the following condition: $A_{v,k} = a_{v,0} + ... + a_{v,k}$ is positive for all $k = 0, ..., 7$ and all $s \in [6, 16]$. The Positive Dominance Lemma now implies that the corresponding polynomial is positive on $[0, 1]$.

For each $v = 0, ..., 15$ and each $k = 0, ..., 7$ we have a 3×3 integer matrix $\mu_{v,k}$ such that

$$A_{v,k} = (1, s, s^2) \cdot \mu_{v,t} \cdot (3^{s/2}, 4^{s/2}, 6^{s,2}). \quad (39)$$

This gives 128 matrices to check. We get two more such matrices from the conditions $-\beta_s'(0) > 0$ and $\beta_s'(4) > 0$. All in all, we have to check that 130 expressions of the form in Equation 39 are positive for $s \in [6, 16]$. These expressions are all special cases of Equation 25, and we use the method discussed above to show positivity in all 130 cases. The program runs in several hours.

5 The Local Convexity Theorem

Reading Guide: This chapter is for Reader 2.

We use the notation from §2. Recall that Σ^{-1} is inverse stereographic projection. The small domain Ω_0 is defined in §2.4. Here is the result we prove in this chapter.

Theorem 5.1 (Local Convexity) *For $F = G_4, G_6, G_5^\flat, G_{10}^\sharp$, the Hessian of $\mathcal{E}_F \circ \Sigma^{-1}$ is positive definite at every point of Ω_0.*

5.1 Discussion

Before we launch into the proof, we discuss why the proof has the structure that it does. We are interested in showing that certain functions, essentially the eigenvalues of the Hessian matrix of various energy potentials, are positive in a certain definite neighborhood.

Consider a toy version of this problem where we want to show that a real valued smooth function f is positive on an interval $[0, a]$. We only want to evaluate f and its derivatives at 0. In general, this is a hopeless task, but let us discuss it anyhow.

We evaluate $f(0)$ and we notice that it is positive. if we knew that $|f'|$ was small enough on $[0, a]$ then we could use Taylor's Theorem with Remainder to show that $f > 0$ on $[0, a]$. We could evaluate $|f'(0)|$ and observe that is is quite small. If we knew that $|f''|$ was small on $[0, a]$ then we could again use Taylor's Theorem to show that $|f'|$ is small enough on $[0, a]$.

In general, we have a recursive problem with no end in sight. However, in our specific situation we have some good algebraic luck that saves us. It turns out that certain *a priori* algebraic bounds on the nth derivatives grow slowly in comparison to the large constant $n!$ we divide by when we apply Taylor's Theorem with Remainder.

We will use an algebraic trick to get reasonable bounds on high derivatives of the functions of interest to us, and then we will use Taylor's Theorem with remainder to promote these decent bounds on high derivatives to excellent bounds on the lower derivatives.

5.2 Reduction to Simpler Statements

We consider F to be any of the 4 functions

$$G_4, \quad G_6, \quad G_5^\flat = G_5 - 25G_1, \quad 2^{-5}G_{10}^\sharp = 2^{-5}(G_{10} + 13G_5 + 68G_2).$$

Scaling the last function by 2^{-5} makes our estimates more uniform.

Recall that Ω_0 is the cube of side length 2^{-17} centered at the point

$$\xi_0 = \Big(1, 0, \frac{-1}{\sqrt{3}}, -1, 0, 0, \frac{1}{\sqrt{3}}\Big) \in \boldsymbol{R}^7 \tag{40}$$

In general, the point $(x_1, ..., x_7)$ represents the avatar

$$p_0 = (x_1, 0),\ p_1 = (x_2, x_3),\ p_2 = (x_4, x_5),\ p_3 = (x_6, x_7). \tag{41}$$

The quantity $F(x_1, ..., x_7)$ is the F-potential of the 5-point configuration associated to the avatar under inverse stereographic projection Σ^{-1}.

$$F(x_1, ..., x_7) = \sum_{i<j} F(\|\widehat{p}_i - \widehat{p}_j\|), \qquad \widehat{p} = \Sigma^{-1}(p). \tag{42}$$

Equation 5 gives the formula for Σ^{-1}.

Let HF be the Hessian of F. The Local Convexity Theorem says HF is positive definite in Ω_0. Let $\partial_J F$ be the (iterated) partial derivative of F with respect to a multi-index $J = (j_1, ..., j_7)$. Let $|J| = j_1 + ... + j_7$. Let

$$M_N = \sup_{|J|=N} M_J, \qquad M_J = \sup_{\xi \in \Omega_0} |\partial_J F(\xi)|, \tag{43}$$

Let $\lambda(M)$ be the smallest eigenvalue of a real symmetric matrix M. The Local Convexity Theorem is an immediate consequence of the following two lemmas.

Lemma 5.2 (L1) *If $M_3(F) < 2^{12}\lambda(HF(\xi_0))$ then $\lambda(HF(\xi)) > 0$ for all $\xi \in \Omega_0$.*

Lemma 5.3 (L2) *$M_3(F) < 2^{12}\lambda(H\mathcal{E}_F(\xi_0)))$ in all cases.*

5.3 Proof of Lemma L1

Let

$$H_0 = HF(\xi_0), \qquad H = HF(\xi), \qquad \Delta = H - H_0. \tag{44}$$

For any real symmetric matrix X define the L_2 matrix norm:

$$\|X\|_2 = \sqrt{\sum_{ij} X_{ij}^2} = \sup_{\|v\|=1} \|Xv\|. \tag{45}$$

Given a unit vector $v \in \boldsymbol{R}^7$ we have $H_0 v \cdot v \geq \lambda$. Hence

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

So, to prove Lemma L1 we just need to establish the implication

$$M_3 < 2^{12}\lambda(H_0) \quad \Longrightarrow \quad \|\Delta\|_2 < \lambda(H_0).$$

Let $t \to \gamma(t)$ be the *unit speed parametrized* line segment connecting p_0 to p in Ω_0. Note that γ has length $L \leq \sqrt{7} \times 2^{-18}$. We write $\gamma = (\gamma_1, ..., \gamma_7)$. Let H_t denote the Hessian of F evaluated at $\gamma(t)$. Let D_t denote the directional derivative along γ.

Now $\|D_t(H_t)\|_2$ is the speed of the path $t \to H_t$ in $\boldsymbol{R}^{49}$, and $\|\Delta\|_2$ is the Euclidean distance between the endpoints of this path. Therefore

$$\|\Delta\|_2 \leq \int_0^L \|D_t(H_t)\|_2 \; dt. \tag{46}$$

Let $(H_t)_{ij}$ denote the ijth entry of H_t. From the definition of directional derivatives, and from the Cauchy-Schwarz inequality, we have

$$(D_t H_t)_{ij}^2 = \Big(\sum_{k=1}^{7} \frac{d\gamma_k}{dt} \frac{\partial H_{ij}}{\partial k} \Big)^2 \leq 7 M_3^2. \qquad \|D_t(H_t)\|_2 \leq 7^{3/2} M_3. \tag{47}$$

The second inequality follows from summing the first one over all 7^2 pairs (i, j) and taking the square root. Equation 46 now gives

$$\|\Delta\|_2 \leq L \times 7^{3/2} M_3 = 49 \times 2^{-18} M_3 < 2^{-12} M_3 < \lambda(H_0). \tag{48}$$

This completes the proof.

5.4 Proof of Lemma L2

Let F be any of our functions. Let $H_0 = HF(\xi_0)$.

Lemma 5.4 (L21) $\lambda(H_0) > 39$.

Proof: Let χ be the characteristic polynomial of H_0. This turns out to be a rational polynomial. We check in Mathematica that the signs of the coefficients of $\chi(t+39)$ alternate. Hence $\chi(t+39)$ has no negative roots. The file we use is `L21.m`. ♠

Recalling that $\xi_0 \in \boldsymbol{R}^7$ is the point representing the TBP, we define

$$\mu_N(F) = \sup_{|I|=N} |\partial_I F(\xi_0)|. \tag{49}$$

Lemma 5.5 (L22) *For any of our functions we have the bound*

$$\mu_3 < 45893, \qquad \frac{(7 \times 2^{-18})^j}{j!}\mu_{j+3} < 38, \quad j = 1,2,3. \tag{50}$$

Proof: We compute this in Mathematica. The file we use is `L22.m`. ♠

Lemma 5.6 (L23) *For any of our functions we have the bound*

$$\frac{(7 \times 2^{-18})^4}{4!} M_7 < 2354.$$

Proof: We give this proof in the next section. ♠

Lemma 5.7 (L24) *We have*

$$M_3 \le \mu_3 + \sum_{j=1}^{3} \frac{(7 \times 2^{-18})^j}{j!}\mu_{j+3} + \frac{(7 \times 2^{-18})^4}{4!} M_7 \tag{51}$$

Proof: Choose any multi-index J with $|J| = 3$. Let γ be the line segment connecting ξ_0 to any $\xi \in \Omega$. We parametrize γ by unit speed and furthermore set $\gamma(0) = \xi_0$. Let

$$f(t) = \partial_J F \circ \gamma(t).$$

The bound for $|M_J|$ follows from Taylor's Theorem with remainder once we notice that

$$0 \le t \le \sqrt{7} \times 2^{-18}, \qquad \Big|\frac{\partial^n f(0)}{\partial t^n}\Big| \le (\sqrt{7})^n \mu_n \qquad \Big|\frac{\partial^n f}{\partial t^n}\Big| \le (\sqrt{7})^n M_n.$$

Since this works for all J with $|J| = 3$ we get the same bound for M_3. ♠

The lemmas above and Equation 50 imply

$$M_3 < 45893 + 3 \times 38 + 2354 \le 65536 = 2^{16} \le 2^{12}\lambda(H_0).$$

This completes the proof of Lemma L2.

5.5 Proof of Lemma L23

Now we come to the interesting part of the proof, the one place where we need to go beyond specific evaluations of our functions. When $r, s \ge 0$ and $r + s \le 2d$ we have

$$\sup_{(x,y)\in \boldsymbol{R}^2} \frac{x^r y^s}{(1 + x^2 + y^2)^d} \le (1/2)^{\min(r,s)}. \tag{52}$$

One can prove Equation 52 by factoring the expression into pieces with quadratic denominators. Here is a more general version. Say that a function $\phi : \boldsymbol{R}^4 \to \boldsymbol{R}$ is *nice* if it has the form

$$\sum_i \frac{C_i a^{\alpha_i} b^{\beta_i} c^{\gamma_i} d^{\delta_i}}{(1 + a^2 + b^2)^{u_i}(1 + c^2 + d^2)^{v_i}}, \;\; \alpha_i, \beta_i, \gamma_i, \delta_i \ge 0, \quad \alpha_i + \beta_i \le 2u_i, \quad \gamma_i + \delta_i \le 2v_i.$$

It follows from Equation 52 that

$$\sup_{\boldsymbol{R}^4} |\phi| \le \langle \phi \rangle, \qquad \langle \phi \rangle = \sum_i |C_i| (1/2)^{\min(\alpha_i,\beta_i)+\min(\gamma_i,\delta_i)}. \tag{53}$$

Equation 53 is useful to us because it allows us to bound certain kinds of functions without having to evaluate then anywhere. We also note that if

ϕ is nice, then so is any iterated partial derivative of ϕ. Indeed, the nice functions form a ring that is invariant under partial differentiation. This fact makes it easy to identify nice functions.

For any $\phi : \boldsymbol{R}^n \to \boldsymbol{R}$ we define

$$\overline{M}_7(\psi) = \sup_{|J|=7} \overline{M}_J(\psi), \qquad \overline{M}_J(\psi) = \sup_{\xi \in \boldsymbol{R}^n} |\partial_J(\phi)|. \tag{54}$$

We obviously have

$$M_7(F) \le \overline{M}_7(F). \tag{55}$$

Recall that $\widehat{p} = \Sigma^{-1}(p)$, the inverse stereographic image of p. Define

$$f(a,b) = 4 - \|\widehat{(a,b)} - (0,0,1)\|^2 = \frac{4(a^2+b^2)}{1+a^2+b^2}. \tag{56}$$

$$g(a,b,c,d) = 4 - \|\widehat{(a,b)} - \widehat{(c,d)}\|^2 = \frac{4(1+2ac+2bd+(a^2+b^2)(c^2+d^2))}{(1+a^2+b^2)(1+c^2+d^2)}. \tag{57}$$

Notice that g is nice. Hence g^k is nice and $\partial_I g^k$ is nice for any multi-index. That means we can apply Equation 53 to $\partial_I g^k$.

G_k is a 10-term expression involving 4 instances of f^k and 6 of g^k. However, each variable appears in at most 4 terms. So, as soon as we take a partial derivative, at least 6 of the terms vanish. Moreover, $\partial_I f$ is a limiting case of $\partial_I g$ for any multi-index I. From these considerations, we see that

$$\overline{M}_7(G_k) \le 4 \times \overline{M}_7(g^k). \tag{58}$$

The function $\partial_I(g^k)$ is nice in the sense of Equation 53. Therefore

$$4 \times \overline{M}_7(g^k) \le 4 \times \max_{|I|=7} \langle \partial_I g^k \rangle. \tag{59}$$

Using this estimate, and the Mathematica file `L23.m`, we get

$$\max_{k \in \{1,2,3,4,5,6\}} \frac{(7 \times 2^{-18})^4}{4!} \times 4 \times \overline{M}_7(g^k) \le \frac{1}{1000}.$$

$$2^{-5} \times \frac{(7 \times 2^{-18})^4}{4!} \times 4 \times \overline{M}_7(g^{10}) \le 2353. \tag{60}$$

The bounds in Lemma 5.6 follow directly from Equations 58 - 60 and from the definitions of our functions.

6 The Symmetrization Theorem

Reading Guide: This is for Reader 3. We prove the Symmetrization Theorem from §3.5.

6.1 Reduction to Four Lemmas

What makes our proof possible is that we can break the 10-term sum into smaller sums, each involving just a few terms, which are separately decreased by the symmetrization operation.

The domain Υ is defined in §3.7. Let $X = (p_0, p_1, p_2, p_3)$ be an avatar in Υ. We let X' be the planar configuration which is obtained by rotating X about the origin so that p_0' and p_2' lie on the same horizontal line, with p_0' lying on the right. This operation does not change the R_s-energy. Let Υ' denote the domain of avatars X' such that (comparing with Υ)

1. $\|p_0'\| \geq \|p_k'\|$ for $k = 1, 2, 3$.
2. $512p_0' \in [432, 498] \times I_{16}$. (Compare $[433, 498] \times I_0$.)
3. $512p_1' \in I_{32} \times [-465, -348]$. (Compare $I_{16} \times [-464, -349]$.)
4. $512p_2' \in [-498, -400] \times I_{16}$. (Compare $[-498, -400] \times [0, 24]$.)
5. $512p_3' \in I_{32} \times [348, 465]$. (Compare $I_{16} \times [349, 464]$.)
6. $p_{02}' = p_{22}'$. (Compare $p_{02} = 0$.)

Lemma 6.1 (B1) *If $X \in \Upsilon$ then $X' \in \Upsilon'$.*

Proof: This is the most tedious proof in the whole paper! Condition 6 holds by construction. Rotation about the origin does not change the norms, so X' satisfies Condition 1. Now we check the other conditions.

Let ρ_θ denote the counterclockwise rotation through the angle θ. Since p_0 lies on the x axis and p_2 lies on or above it, we have to rotate by a small amount counterclockwise to get p_0' and p_2' on the same horizontal line. Hence $\theta \geq 0$. This angle is maximized when p_0 is an endpoint of its segment of constraint and p_2 is one of the two upper vertices of rectangle of constaint.

We check for all 4 pairs (p_0, p_2) that the second coordinate of $\rho_{1/34}(p_0)$ is larger than the second coordinate of $\rho_{1/34}(p_2)$. Hence $\theta < 1/34$. This yields

$$512\cos(\theta) \in [0,1], \qquad 512\sin(\theta) \in [0,16]. \tag{61}$$

From Equation 61, the map $512p_0 \to 512p_0'$ changes the first coordinate by $512\delta_{01} \in [0,16]$ and $512\delta_{02} \in [-1,0]$. Condition 2 follows. Next, we have $512\delta_{21} \in [0,1]$. This gives Condition 4 for Υ' because $|p_{21}'| \le |p_{01}'|$.

Condition 3 follows from $512\delta_{11} \in [0,16]$ and $512\delta_{12} \in [-1,1]$. The first bound comes from $512\sin(\theta) < 16$. For the second bound we note that the angle that p_1 makes with the y-axis is maximized when p_1 is at the corners of its constraints in Υ. That is, $512p_1 = (16, 349)$. Since $\tan(1/21) > 16/349$ we conclude that this angle is at most $1/21$. Hence

$$|512\delta_{12}| \le \max_{|x| \le 1/21} \left| \cos\left(x + \frac{1}{34}\right) - \cos(x) \right| < 1.$$

This gives Condition 3. The same argument gives Condition 5. ♠

Given an avatar $X' \in \Upsilon'$, there is a unique configuration X'', invariant under under reflection in the y-axis, such that p_j' and p_j'' lie on the same horizontal line for $j = 0,1,2,3$ and $\|p_0'' - p_2''\| = \|p_0' - p_2'\|$. We call this *horizontal symmetrization*. In a straightforward way we see that horizontal symmetrization maps Υ' into Υ'', the set of avatars $p_0'', p_1'', p_2'', p_3''$ such that

1. $-512p_2'', 512p_0'' \in [416, 498] \times I_{16}$
2. $-512p_1'', 512p_3'' \in I_0 \times [348, 465]$.
3. $p_{02}'' = p_{22}''$.

Given a configuration $X'' \in \Upsilon''$ there is a unique configuration $X''' \in \mathbf{K4}$ such that p_j'' and p_j''' lie on the same vertical line for $j = 0,1,2,3$. We call this operation *vertical symmetrization*. Here $X''' = X^*$ from Lemma B.

Given an avatar $X = (p_0, p_1, p_2, p_3)$ define

$$r_{ij} = \frac{1}{\|\Sigma^{-1}(p_i) - \Sigma^{-1}(p_j)\|}. \tag{62}$$

Given a list L of pairs of points we define $R_s(X, L)$ to be the sum of the R_s-potentials just over the pairs in L. E.g. $R_s(X, \{(0,2),(0,4)\}) = r_{02}^s + r_{04}^s$.

We call L *good* for s, and with respect to one of the operations, if the operation does not increase the value of $R_s(X, L)$. We call L *great* if the operation strictly lowers $R_s(X, L)$ unless the operation fixes P. We mean to take the appropriate domains in all cases. The Symmetrization Theorem follows immediately from Lemma B1 and from the 3 lemmas below.

Lemma 6.2 (B2) *On* Υ, $\{(0,2),(0,4),(2,4)\}$ *and* $\{(1,3),(1,4),(3,4)\}$ *are both great for all* $s \geq 2$ *and with respect to symmetrization.*

Lemma 6.3 (B3) *On* Υ', *the lists* $\{(0,1),(1,2)\}$ *and* $\{(0,3),(3,2)\}$ *are both good for all* $s \geq 2$ *and with respect to horizontal symmetrization.*

Lemma 6.4 (B4) *on* Υ'', *the lists* $\{(0,1),(0,3)\}$ *and* $\{(2,1),(2,3)\}$ *are both good for all* $s \geq 12$ *and with respect to vertical symmetrization.*

6.2 Proof of Lemma B2

Let $s_3 = \sqrt{3}/3$. Inverse stereographic projection maps the triangle with vertices $(\pm s_3, 0)$ and ∞ to an equilateral triangle on S^2. Avatars in Υ satisfy

$$\|p_0\|, \|p_1\|, \|p_2\|, \|p_3\|, \frac{\|p_0 - p_2\|}{2}, \frac{\|p_1 - p_2\|}{2} \in (s_3, 1).$$

Let (u, v) stand for either $(0,2)$ or $(1,3)$.

1. Let $a > 0$ be such that $\|p_u - p_v\|/2 = s_3 + a$. Let $-q_u = q_v = (s_3 + a, 0)$. The points q_u, q_v are symmetric w.r.t the y-axis. Also set $a_u = a_v = a$.

2. Choose b_u, b_v with $0 < b_u \leq a_u$ and $0 < b_v \leq a_v$. Let $r_u = (-s_3 - b_u, 0)$ and $r_v = (s_3 + b_v, 0)$. Note that $\|r_u - r_v\| \leq \|q_u - q_v\|$.

up to rotation about the origin, our symmetrization operation does the map $(p_u, p_v) \to (r_u, r_v)$ for suitable a_u, a_v, b_u, b_v. For our symmetrization operation we have the additional properties $b_0 = b_2 = a$ and $b_1 = b_3$, but we want to consider the more general case as part of our proof strategy.

Recall that $\widehat{p}$ is the image of p under inverse stereographic projection. Lemma B2 is implied by:

$$\|\widehat{r}_u - \widehat{r}_v\|^{-s} + \|\widehat{r}_u - (0,0,1)\|^{-s} + \|\widehat{r}_v - (0,0,1)\|^{-s} \leq$$
$$\|\widehat{p}_u - \widehat{p}_v\|^{-s} + \|\widehat{p}_u - (0,0,1)\|^{-s} + \|\widehat{p}_v - (0,0,1)\|^{-s} \tag{63}$$

for all $s \geq 2$, with equality iff $(r_u, r_v) = (p_u, p_v)$ up to rotation about the origin. We will establish this in two steps.

Lemma 6.5 (B21) *Let $s \geq 2$ and*

$$A_s = \|\widehat{p}_u - \widehat{p}_v\|^{-s} - \|\widehat{q}_u - \widehat{q}_v\|^{-s},$$

$$B_s = \|\widehat{p}_u - (0,0,1)\|^{-s} + \|\widehat{p}_v - (0,0,1)\|^{-s} - \|\widehat{q}_u - (0,0,1)\|^{-s} - \|\widehat{q}_v - (0,0,1)\|^{-s}.$$

Then $A_s, B_s \geq 0$, with equality iff $p_u = q_u$ and $p_v = q_v$ up to a rotation.

Proof: Note that if $A_2 > 0$ then $A_s > 0$ for all $s > 0$. If $B_2 > 0$ then the Convexity Lemma implies that $B_s > 0$ for all $s > 2$. So, it suffices to prove that $A_2, B_2 > 0$. We rotate so that

$$p_u = (-x + h, y), \quad p_v = (x + h, y), \quad q_u = (-x, 0), \quad q_v = (x, 0). \tag{64}$$

We compute

$$A_2 = \frac{h^4 + y^2(2 + 2x^2 + y^2) + 2h^2(1 - x^2 + y^2)}{16x^2}, \qquad B_2 = \frac{y^2 + h^2}{2}. \tag{65}$$

Since $x \in (0,1)$ we have $A_2, B_2 > 0$ unless $h = y = 0$. ♠

Define

$$F_s(a_u, a_v) = \|\widehat{q}_u - \widehat{q}_v\|^{-s} + \|\widehat{q}_u - (0,0,1)\|^{-s} + \|\widehat{q}_v - (0,0,1)\|^{-s}, \tag{66}$$

Likewise define $F_s(b_u, b_v)$. This is the same expression with respect to $\widehat{r}_u$ and $\widehat{r}_v$. Finally, define

$$E(s) = F_s(a_u, a_v) - F_s(b_u, b_v). \tag{67}$$

Lemma 6.6 (B22) *$E(s) \geq 0$ with equality iff $b_u = a_u$ and $b_v = a_v$.*

Proof: It suffices to prove this result in the intermediate case when $a_u = b_u$ or $a_v = b_v$ because then we can apply the intermediate result twice to get the general case. Without loss of generality we consider the case when $a_v = b_v$ and $b_u < a_u$. With the file `LemmaB22.m` we compute that $\partial F_2/\partial a_u$ and $-\partial F_{-2}/\partial a_u$ are both rational functions of a_u, a_v with all positive coefficients. Hence $E(2) > 0$ and $E(-2) < 0$.

Referring to §2.6.4, consider the sign sequence for $E(s)$. When $a_u = b_u$, the expression $E(s)$ is an exponential sum with 4 terms. When $a_u = a_v = 0$ the points $\widehat{\zeta}_u, \widehat{\zeta}_v$ and $(0,0,1)$ make an equilateral triangle on a great circle.

Hence, when $a_u, a_v, b_u, b_v > 0$ the point $\widehat{\zeta}_u$ is closer to $(0,0,1)$ than it is to $\widehat{\zeta}_v$ both in its old location and in its new location. The inward motion of the point ζ_u increases the shorter (corresponding spherical) distance and decreases the longer (corresponding spherical) distance. More to the point, our move decreases the longer inverse-distance and increases the shorter inverse-distance. Thus the sign sequence for $E(s)$ is $+, -.-, +$.

By Descartes' Lemma, $E(s)$ changes sign at most twice and also $E(s) > 0$ when $|s|$ is sufficiently large. Since $E(-2) < 0$ as see that E changes sign on $(-\infty, -2)$. If E has a root in $(2, \infty)$ then in fact E has at least 2 roots (counted with multiplicity) because it starts and ends positive on this interval. But then E has at least 3 roots, counting multiplicity. This is contradiction. Hence $E(s) > 0$ for $s \geq 2$. ♠

6.3 Proof of Lemma B3

The domain Υ' is symmetric with respect to reflection in the X-axis. Thanks to this symmetry, it suffices to prove Lemma B3 for the list $\{(0,1),(1,2)\}$. We set $q_j = p'_j$ and $q'_j = p''_j$.

We introduce the notation $q_1 = (q_{10}, q_{11})$, etc. The horizontal symmetrization operation is given by

$$(q_0, q_1, q_2) \to (q'_0, q'_1, q'_2),$$

where

$$q'_0 = \Big(\frac{q_{01} - q_{21}}{2}, q_{02}\Big), \qquad q'_1 = (0, q_{21}), \qquad q'_2 = \Big(\frac{q_{21} - q_{01}}{2}, q_{22}\Big), \tag{68}$$

Note that $\|q'_0 - q'_1\| = \|q'_2 - q'_1\|$. This means that the kind of inequality we are trying to establish has the form $2A^s \leq B^s + C^s$ for choices of A, B, C which depend on the points involved. Therefore, by the Convexity Lemma, it suffices to prove that $\{(0,1),(1,2)\}$ is good for the parameter $s = 2$.

Let D denote the set of triples of points $(q_0, q_1, q_2) \in (\boldsymbol{R}^2)^3$ such that there is some q_3 such that $q_0, q_1, q_2, q_3 \in \Upsilon'$. Most of our proof involves finding a concrete parametrization of a subset of $\boldsymbol{R}^6$ that contains D. Note that D is really a 5 dimensional set, because $q_{22} = q_{02}$. We will use parameters a, b, c, d, e to parametrize a subset of $\boldsymbol{R}^6$ that contains D.

We define

$$[a, b, t] = \frac{a(1-t)}{512} + \frac{bt}{512}. \tag{69}$$

Here $F_{512}(a,b,\cdot)$ maps the interval $[0,1]$ onto the interval $[a,b]/512$. Given $(a,b,c,d,e) \in [0,1]^5$ and $\sigma_1, \sigma_2 \in \{-,+\}$ we define

$$\begin{aligned} p0 &= ([+416,+498,a]+[0,49,e],[0,16\sigma_1,b]);\\ p1 &= ([0,32\sigma_2,d],[348,465,c]);\\ p2 &= ([-416,-498,a]+[0,49,e],[0,16\sigma_1,b]); \end{aligned} \tag{70}$$

We call this map ϕ_{σ_1,σ_2}. In these coordinates, horizontal symmetrization is the map

$$(a,b,c,d,e) \to (a,b,c,0,0). \tag{71}$$

We have two steps we need to take. First we really need to show that we have parametrized a superset of D. Second, we need to calculate the energy change as a function of a,b,c,d,e and check at it decreases.

Lemma 6.7 (B31) *We have*

$$D \subset \phi_{+,+}([0,1]^5) \cup \phi_{+,-}([0,1]^5) \cup \phi_{-,+}([0,1]^5) \cup \phi_{-.-}([0,1]^5).$$

Proof: Recall that $q_i = (q_{i1}, q_{i2})$. Let D_{ij} denote the set of possible coordinates q_{ij} that can arise for points in D. Thus, for instance

$$D_{01} = [-16,16]/512.$$

Let D^*_{ij} denote the set of possible coordinates q_{ij} that can arise from the union of our parametrizations. By construction $D_{i2} \subset D^*_{i2}$ for $i = 0,1,2$ and $D_{11} \subset D^*_{11}$.

Remembering that we have $q_{01} \geq |q_{21}|$, we see that the set of pairs $512(q_{01}, q_{21})$ satisfying all the conditions for inclusion in D lies in the triangle Δ with vertices

$$(498,-498), \quad (498,-400), \quad (432,-400).$$

At the same time, the set of pairs $(512)(p^*_{01}, p^*_{21})$ that we can reach with our parametrization is the rectangle Δ^* with vertices

$$(498,-498), \quad (416,-416), \quad (498,-498)+(49,49), \quad (416,-416)+(49,49).$$

One checks easily that hence $\Delta \subset \Delta^*$. Indeed, Δ is inscribed in Δ^*. ♠

Using our coordinates above, we define

$$F_{\pm,\pm}(a,b,c,d,e) = \|\widehat{q}_0 - \widehat{q}_1\|^{-2} + \|\widehat{q}_2 - \widehat{q}_1\|^{-2},$$

$$\Phi_{\pm,\pm}(a,b,c,d,e) = \text{num}_+(F_{\pm,\pm}(a,b,c,d,e) - F_{\pm,\pm}(a,b,c,0,0)). \tag{72}$$

Here q_0, q_1, q_2 are the points which correspond to (a,b,c,d,e) under our map $\phi_{\pm,\pm}$ and $\widehat{q}_0, \widehat{q}_1, \widehat{q}_2$ are their images under inverse stereographic projection. To finish our proof, we just have to show that $\Phi_{\pm,\pm}(a,b,c,d,e) \geq 0$ on $[0,1]^5$. The following lemma, and continuity, gives us this result.

Lemma 6.8 (B32) *For any sign choice,* $\Phi_{\pm,\pm} \geq 0$ *on* $[0,1]^5$.

Proof: We let $\Phi_a = \partial\Phi/\partial a$, and likewise for the other variables. Iterating this notation, we let Φ_{aa}, etc., denote the second partials.

Let Φ be any of the 4 polynomials. The file `LemmaB32.m` opencomputes that

1. Φ and Φ_d and Φ_e are zero when $d = e = 0$.
2. Φ_{dd} and Φ_{ee} are weak positive dominant, hence nonnegative on $[0,1]^5$.
3. $\Phi_d + 2\Phi_e$ is weak positive dominant, hence nonnegative on $[0,1]^5$.

Let $Q_d \subset [0,1]^5$ be the sub-cube where $d = 0$. We fix (a,b,c) and consider the single variable function $\phi(d) = \Phi(a,b,c,d,0)$. From Items 1 and 2 above, $\phi(0) = \phi'(0) = 0$ and $\phi''(d) \geq 0$. Hence $\phi(d) \geq 0$ for $d \geq 0$. Hence $\Phi \geq 0$ on Q_d. A similar argument shows that likewise $\Phi \geq 0$ on Q_e.

Any point in $(0,1)^5$ can be joined to a point in $Q_d \cup Q_e$ by a line segment L which is parallel to the vector $(0,0,0,1,2)$. From Item 3 above, Φ increases along such a line segment as we move out of $Q_d \cup Q_e$. Hence $\Phi \geq 0$ on $[0,1]^5$. ♠

6.4 Proof of Lemma B4

The set Υ'' is symmetric with respect to reflections in both coordinate axes. Thanks to these symmeties, it suffices to prove that $\{(0,1),(0,3)\}$ is good for all $s \geq 12$, and it suffices to consider the case when $p''_{02} \geq 0$. That is, the

point p_0 lies on or above the X-axis. For ease of notation set $q_k = p_k''$ and $q_k' = p_k'''$. We are considering the case when $q_{02} \geq 0$.

Let D be the set of configurations (q_0, q_1, q_3) such that $q_{02} \geq 0$ and $(q_0, q_1, q_2, q_3) \in \Upsilon''$ when q_2 is the reflection of q_0 in the Y-axis. Let $D_\pm \subset D$ denote those configurations with $\pm(q_{12} + q_{32}) \geq 0$. Obviously $D = D_+ \cup D_-$.

The sets $D_\pm$ are 4-dimensional subsets of $(\boldsymbol{R}^2)^3$. We parametrize a superset of $D_\pm$ much as we did in the proof of Lemma B3. As in Equation 69 we define

$$[a, b, t] = \frac{(1-t)a}{512} + \frac{bt}{512}.$$

Given $(a, b, c, d) \in [0, 1]^4$ and $\sigma \in \{+, -\}$ we define

$$\begin{aligned} p_0 &= ([416, 498, b], [0, 16, d]); \\ p_1 &= (0, -[348, 465, a] + [0, 59\sigma, c]); \\ p_3 &= (0, +[348, 465, a] + [0, 59\sigma, c]); \end{aligned} \tag{73}$$

We call this map ϕ_σ. In these coordinates, the symmetrization operation is $(a, b, c, d) \to (a, b, 0, 0)$.

Lemma 6.9 (B41) *$D_\pm \subset \phi_\pm([0, 1]^4)$.*

Proof: This is just like the proof of Lemma B31. The only non-obvious point is why every pair (p_{12}, p_{32}) is reached by the map $\phi_\pm$. The essential point is that for configurations in $D_\pm$ we have $512|p_{12} + p_{32}| \leq 2 \times 59$. ♠

Following the same idea as in the proof of Lemma B3, we define

$$F_{s,\pm}(a, b, c, d) = \|\Sigma^{-1}(q_0) - \Sigma^{-1}(q_1)\|^{-s} + \|\Sigma^{-1}(q_0) - \Sigma^{-1}(q_3)\|^{-s}, \tag{74}$$

$$\Phi_{s,\pm}(a, b, c, d) = \text{num}_+(F_{s,\pm}(a, b, c, d) - F_{s,\pm}(a, b, 0, 0)). \tag{75}$$

The points on the right side of Equation 74 are coordinatized by the map $\phi_\pm$. We can finish the proof by showing that $\phi_{2,+} \geq 0$ and $\phi_{12,-} \geq 0$ on $[0, 1]^4$. The Convexity Lemma then takes care of all exponents greater than 2 on D_+ and all exponents greater than 12 on D_-. Notice the asymmetry in the calculation. The (+) side is much less delicate.

Lemma 6.10 (B42) *$\Phi_{2,+} \geq 0$ on $[0, 1]^4$.*

Proof: Let $\Phi = \Phi_{2,+}$. Let $\Phi|_{c=0}$ denote the polynomial we get by setting $c = 0$. Etc. Let $\Phi_c = \partial\Phi/\partial c$, etc. The Mathematica file `LemmaB42.m` computes that $\Phi|_{c=0}$ and $\Phi|_{d=0}$ and $\Phi_c + \Phi_d$ are weak positive dominant. Hence $\Phi \geq 0$ when $c = 0$ or $d = 0$ and the directional derivative of Φ in the direction $(0, 0, 1, 1)$ is non-negative. This suffices to show that $\Phi \geq 0$ on $[0, 1]^4$. ♠

Lemma 6.11 (B43) $\Phi_{12,-} \geq 0$ *on* $[0, 1]^4$.

Proof: The file `LemmaB43.m` has the calculations. Let $\Phi = \Phi_{12,-}$. This monster has 102218 terms.

Step 1: Let M denote the maximum coefficient of Φ. We let Φ^* be the polynomial we get by taking each coefficient of c of Φ and replacing it with $\text{floor}(10^{10}c/M)$. Note that if Φ^* is nonnegative on $[0, 1]^4$ then so is Φ.

Step 2: Now Φ^* has 37760 monomials in which the coefficient is -1. We check that each such monomial is divisible by one of c^2 or d^2 or cd. Let $\Psi = \Phi^{**} - 37760(c^2 + d^2 + cd)$, where Φ^{**} is obtained from Φ^* by setting all the (-1) monomials to 0. We have $\Psi \leq \Phi^*$ on $[0, 1]^4$. Hence, if Ψ is non-negative on $[0, 1]^4$ then so is Φ^*. The polynomial Ψ has 5743 terms.

Step 3: We check that Ψ_{aaa} is WPD and hence non-negative on $[0, 1]^4$. This massive calculation reduces us to showing that the restrictions $\Psi|_{a=0}$ and $\Psi_a|_{a=0}$ and $\Psi_{aa}|_{a=0}$ are all non-negative on $[0, 1]^3$. Consider

$$f|_{c=0}, \qquad f|_{d=0} \qquad 4f_c + f_d, \tag{76}$$

We show that all three functions are WPD when either $f = \Psi_a|_{a=0}$ or $f = \Psi_{aa}|_{a=0}$. This shows that $\Psi_a|_{a=0}$ and $\Psi_{aa}|_{a=0}$ are non-negative on $[0, 1]^3$. Also, we show that the first two functions are WPD when $f = \Psi|_{a=0}$.

Step 4: Let $g = 4f_c + f_d \geq 0$ on $[0, 1]^3$ when $f = \Psi|_{a=0}$. We check that g_d is WPD and hence non-negative on $[0, 1]^3$. This reduces us to showing that $h = g|_{d=0}$ is non-negative on $[0, 1]^2$. here h is a 2-variable polynomial in b, c. Referring to the operation in §2.6, we check that the two subdivisions $S_{b,0}(h)$ and $S_{b,1}(h)$ are WPD. This proves $h \geq 0$ on $[0, 1]^2$. ♠

7 Symmetric Configurations

Reading Guide: This chapter is for Reader 4. We prove the Critical Theorems from §3.6. We use the notation from §3.6.

7.1 Critical Theorem I

As in Equation 20, we write $(z,z) = \sigma(x,y)$. Let $\phi : [0,1]^2 \to \Psi_4^\sharp$ be map which scales the coordinates by a factor of $1/64$. We use coordinates a, b on $[0,1]^2$ so that $(x,y) = \phi(a,b)$.

For any rational function $F : \Psi_4^\sharp \to \boldsymbol{R}$ we define

$$N_F(a,b) = \frac{\text{num}_+((F - F \circ \sigma) \circ \phi)}{(a-b)^2}. \tag{77}$$

See §2.6.3. For all the choices of F we make, N_F will be a polynomial.

Recall that $\Sigma^{-1}(p_4) = (0,0,1)$, and define

$$r_{ij} = \frac{1}{\|\Sigma^{-1}(p_i) - \Sigma^{-1}(p_j)\|}. \tag{78}$$

Note that r_{ij}^s is a rational function when s is an even integer.

Let $R_s(x,y)$ be the R_s-energy of the avatar represented by (x,y). We write $R_s(x,y) = G_s(x,y) + H_s(x,y)$, where

$$G_s = r_{02}^s + r_{13}^s, \qquad H_s = 2r_{04}^s + 2r_{14}^s + 4r_{01}^s. \tag{79}$$

Lemma 7.1 (C1) *$G_s - G_s \circ \sigma > 0$ on $\Psi_4^\sharp \times (2,\infty)$.*

Proof: The file `LemmaC1.m` computes that N_{G_2} is a WPD polynomial. This combines with the Convexity Lemma of §2.6.4 to show $G_s - G_s \circ \sigma > 0$ on $\Psi_4^\sharp \times (2,\infty)$. ♠

To finish the proof, we need to show

Lemma 7.2 (C2) *$H_s - H_s \circ \sigma \geq 0$ on $\Psi_4^\sharp \times [14,16]$.*

We first prove two smaller lemmas and then deduce Lemma C2.

We will suppose, for the sake of contradiction, that there is some $(x,y) \in \Psi_4^\sharp$ and some $s \in [14,16]$ such that

$$h(s) = H_s(x,y) - H_s(z,z) < 0. \tag{80}$$

We study the single-variable function h. The idea is to use Descartes' Lemma from §2.6.4 to get a contradiction. We first need some preliminary results.

Lemma 7.3 (C21) *h has at least 3 roots in $[2, 16]$.*

Proof: The file `LemmaC21.m` computes that $-N_{H_2}$ and $N_{H_{14}}$ and $N_{H_{16}}$ are all WPD polynomials. Hence $h(2) < 0$ and $h(14) > 0$ and $h(16) > 0$. ♠

Let (p_0, p_1, p_2, p_3) and (p_0', p_1', p_2', p_3') respectively be the configurations corresponding to (x, y) and $(z, z) = \sigma(x, y)$. Without claiming to have the terms in order, we have

$$h(s) = +2r_{04}^s - 4(r_{04}')^s + 2r_{14}^s + 4r_{01}^s - 4(r_{01}')^s. \tag{81}$$

The next result gives us control on the ordering of these terms.

Lemma 7.4 (C22) *$r_0, r_1, r_0' < 1/\sqrt{2} < r_{01}, r_{01}'$ and $r_{01} < r_{01}'$.*

Proof: We have $x, y, z \in (0, 1)$. We compute

$$(1/2) - r_0^2 = \frac{1 - x^2}{4} > 0, \quad (1/2) - r_1^2 = \frac{1 - y^2}{4} > 0, \quad (1/2) - (r_0')^2 = \frac{1 - z^2}{4} > 0,$$

$$(r_{01})^2 - (1/2) = \frac{(1 - x^2)(1 - y^2)}{4(x^2 + y^2)} > 0, \qquad (r_{01}')^2 - (1/2) = \frac{(1 - z^2)^2}{8z^2} > 0.$$

This proves the first statement.

For the second statement, the file `LemmaC22.m` computes that $-N_{r_{01}^2}$ is a WPD polynomial. Hence $r_{01}' \geq r_{01}$. If we really had $r_{01}' = r_{01}$. Then Equation 81 would only have 3 terms. There would then be at most 2 sign changes and Lemma 7.3 would contradict Descartes' Lemma. We conclude that $r_{01}' > r_{01}$. ♠

Since the largest term in Equation 81 is $-4(r_{01}')^s$ we see that h vanishes for some $s > 16$. Combining this with Lemma 7.3 and the fact that $h(16) > 0$, we see that h changes sign 4 times on $[2, \infty)$. This is only possible if the sign sequence is $- + - + -$. But this is impossible because there are two pluses and three minuses. We have a contradiction. The only way out is that h does not vanish on $[14, 16]$. This proves Lemma C2 and thereby finishes the proof of the Critical Theorem I.

7.2 Critical Theorem II modulo Lemma C3

Derivative Bounds: As above, we identify Ψ_4 with a square in $\boldsymbol{R}^2$. The point $(1, \sqrt{3}/3)$, which is outside Ψ_4, names the TBP. We define

$$\Theta(x, y, s) = R_s(x, y) - R_s(1, \sqrt{3}/3). \tag{82}$$

Here we are comparing the R_s-energy of an avatar in Ψ_4 to the R_s energy of the TBP. Let Θ_x be the partial derivative of Θ with respect to x, etc. In §7.3 we establish the following bound.

Lemma 7.5 (C3) $|\Theta_{xx}|, |\Theta_{yy}| \le 4$ *and* $|\Theta_{ss}| \le 1/64$ *on* $\Psi_4 \times [13, 16]$.

Blocks: We say that a *block* is a rectangular solid of the form

$$X = Q \times J \subset [0,1]^2 \times [0, 16], \tag{83}$$

where Q is a square and J is an interval. We define $|X|_1$ to be the length of J and $|X|_2$ to be the side length of Q. Let $v(X)$ denote vertex set of X.

Lemma 7.6 *For any block* $X \subset \Psi_4 \times \subset [13, 16]$ *we have*

$$\min_X \Theta \ge \min_{v(X)} \Theta - \Big(\frac{|X|_1^2}{512} + |X|_2^2 \Big).$$

Proof: Write $I = [s_0, s_1]$ and $Q = [x_0, x_1] \times [y_0, y_1]$. Choose $(x, y, s) \in X = I \times Q$. Taylor's Theorem with remainder (applied at the point of $[a, b]$ where f is minimized) implies that for any function $f : [a, b] \to \boldsymbol{R}$ and any $x \in [a, b]$ we have

$$f(x) \ge \min(f(a), f(b)) - \frac{1}{8} \max_{[a,b]} |f''| \times |a - b|^2.$$

Applying this result 3 times and using Lemma C3 we have

$$\Theta(x, y, s) \ge \min_i \Theta(x, y, s_i) - \frac{|I|^2}{512} \ge \min_{i,j} \Theta(x_j, y, s_i) - \frac{|I|^2}{512} - \frac{|X|_2^2}{2} \ge$$

$$\min_{i,j,k} \Theta(x_j, y_j, s_i) - \frac{I}{512} - \frac{|X|_2^2}{2} - \frac{|X|_2^2}{2} = \min_{v(X)} \Theta - \frac{|X|_1}{512} - |X|_2^2.$$

This completes the proof. ♠

Verifying Inequalities: Suppose we want to establish an inequality like

$$(\frac{a}{b})^{\frac{p}{q}} < \frac{c}{d},$$

where every number involved is a positive integer. This inequality is true iff

$$b^p c^q - a^p d^q > 0.$$

We check this using exact integer arithmetic. The same idea works with $(>)$ in place of $(<)$. We call this the *expanding out method*.

More generally, we will want to verify inequalities like

$$\sum_{i=1}^{10} b_i^{-s} - \sum_{i=1}^{10} a_i^{-s/2} > C. \tag{84}$$

where all a_i belong to the set $\{2,3,4\}$, and b_i, c, s are all rational. more specifically $s \in [13, 15_+]$ will be a dyadic rational and c will be positive. The expression on the left will be $\mathcal{E}_s(p) - \mathcal{E}_s(p_0)$ for various choices of p, and the constant C is related to the error term we define below.

Here is how we handle expressions like this. For each index $i \in \{1, ..., 10\}$ we produce rational numbers A_i and B_i such that

$$A_i^{s/2} > a_i \qquad B_i^s < b_i. \tag{85}$$

We use the expanding out method to check these inequalities. We then check that

$$\sum_{i=1}^{10} B_i - \sum_{i=1}^{10} A_i > C. \tag{86}$$

This last calculation is again done with integer arithmetic. Equations 85 and 86 together imply Equation 84. Logically speaking, the way that we produce the rational A_i and B_i does not matter, but let us explain how we find them in practice. For A_i we compute $2^{32} a_i^{-s/2}$ and round the result up to the nearest integer N_i. We then set $A_i = N_i/2^{32}$. We produce B_i in a similar way. When we have verified Equation 84 in this manner we say that we have used the *rational approximation method* to verify Equation 84. We will only need to make verifications like this on the order of 20000 times.

The Grading Step: We say that a rational number p/q is *dyadic* if q is a power of 2. We say that a block (defined in the previous chapter) is *dyadic* if all coordinates of all the block vertices are dyadic rationals.

We perform the following pass/fail evaluation of X.

1. If $I \subset [0,13]$ or $I \subset [15_+,16]$ or $Q \cap \Psi_4 = \emptyset$, we pass X because X is irrelevant to the calculation.

2. If $s_0 \geq 15$ and $Q \subset \widehat{\Psi}_4$ we pass X.

3. $s_0 < 13$ and $s_1 > 13$ we fail X because we don't want to make any computations which involve exponents less than 13.

4. If X has not been passed or failed, we try to use the rational approximation method to verify that $\Theta(v) > |X|_1^2/512 - |X|_2^2$ for each vertex v of X. If we succeed at this, then we pass X. Otherwise we fail X.

To prove the Critical Theorem II it suffices to find a partition of

$$[0,16] \times [0,1]^2$$

into blocks which all pass the evaluation.

Subdivision: Let $X = I \times Q$. Here is the rule we use to subdivide X: If $16|X|_2 > |X|_1$ we subdivide X along Q dyadically, into 4 pieces. Otherwise we subdivide X along I, into two pieces. This method takes advantage of the lopsided form of Lemma C22 and produces a small partition.

Running the Algorithm: We perform the following algorithm.

1. We start with a list L of blocks. Initially L has the single member $\{0,16\} \times \{0,1\}^2$.

2. We let B be the last block on L. We grade B. If B passes, we delete B from L. If $L = \emptyset$ then **HALT**. If B fails, we delete B from L and append to L the subdivision of B. Then we go back to Step 1.

For the calculation, I used the computer discussed at the end of the introduction. When I run the algorithm, it halts with success after 21655 steps and in about 1 minute. The partition it produces has 14502 blocks.

This establishes the Critical Theorem II modulo the proof of Lemma C3. In the next section we prove Lemma C3 and also some derivative bounds needed for the Critical Theorem III.

7.3 Critical Theorem III

Let Θ be the function from the previous section.

Lemma 7.7 (C31) *On* $\Psi_4 \times [13,16]$ *we have* $\Theta_{xx}, \Theta_{yy}, \Theta_{xy} > 0$.

Proof: We prove this for Θ_{xx} and Θ_{xy}. The case of Θ_{yy} follows from this and symmetry. Setting $u = s/2$ we compute

$$\mathcal{E}_s(x,y) = A(x,s) + A(y,s) + 2B(x,s) + 2B(y,s) + 4C(x,y,s), \tag{87}$$

$$A(x) = a(x)^u, \qquad B(x) = b(x)^u, \qquad C(x) = c(x)^u,$$

$$a(x) = \frac{(1+x^2)^2}{16x^2} \quad b(x) = \frac{1+x^2}{4} \quad c(x,y) = \frac{(1+x^2)(1+y^2)}{4(x^2+y^2)}$$

Hence

$$\Theta_{xx} = A_{xx} + 2B_{xx} + 4C_{xx}, \qquad \Theta_{xy} = C_{xy}. \tag{88}$$

For each choice of $F = A, B, C$ we have

$$F_{xx} = u(u-1)f^{u-2}f_x^2 + uf^{u-1}f_{xx}, \quad C_{xy} = u(u-1)c^{u-2}c_xc_y + uc^{u-1}c_{xy}. \tag{89}$$

Our notation is such that $f = a$ when $F = A$, etc.

We compute

$$a_{xx} = \frac{3+x^4}{8x^4} > 0, \qquad b_{xx} = \frac{1}{2}, \qquad c_{xx} = \frac{(1-y^4)(3x^2-y^2)}{2(x^2+y^2)^3} \geq 0.$$

$$c_x = \frac{x(y^4-1)}{2(x^2+y^2)^2} < 0, \quad c_y = \frac{y(x^4-1)}{2(x^2+y^2)^2} < 0, \quad c_{xy} = \frac{2xy(1+x^2y^2)}{(x^2+y^2)^3} > 0.$$

Equation 89 combines with all this to prove that $\Theta_{xx} > 0$ and $\Theta_{xy} > 0$ on $\Psi_4 \times [13,16]$. ♠

Proof of Statement 3 of the Critical Theorem III: We actually prove a broader result. Let $s \in [13,16]$. We know by Lemma C31 that all the second partials of Θ are positive at each point of $\Psi_8^\sharp$ for this value of s. But then the restriction of Θ to $\Psi_8^\sharp$, at this parameter s, is a convex function. This shows that for each $s \in [13,16]$ the restriction of Θ to $\Psi_8^\sharp$ has a unique minimizer. In particular, this is true on the smaller interval $(\mathfrak{w}, 15_+]$. ♠

Proof of Lemma C3: We keep the notation from the proof of Lemma C31. We first consider Θ_{xx}. We already know $\Theta_{xx} > 0$ on our domain. An easy exercise in calculus shows that $f \in (0, 3/5)$ on Ψ_4 for each $f = a, b, c$. From this bound, we see that the expression in Equation 89 is decreasing as a function of u for $u \geq 6$. (Recall that $u = s/2$.) Hence it suffices to prove that $4 - \Theta_{xx} \geq 0$ on $\{12\} \times [43/64, 1]^2$.

We define $\phi(t) = (43/64)(1-t) + t$. The file `LemmaC3.m` computes that for $s = 12$ the polynomial $\Phi = \text{num}_+(4 - \Theta_{xx} \circ \phi)$ is WPD and hence non-negative on $[0,1]^2$. Hence $4 - \Theta_{xx} \geq 0$ when $s = 12$ and $(x, y) \in \Psi_4$. The same bound for Θ_{yy} follows from symmetry.

Now we consider Θ_{ss}. Let $\psi(s) = b^{-s}$. Let $\beta = (1.3, \sqrt{2}, \sqrt{3})$ and also let $\gamma = (440, 753, 4184)$. We first establish the following bound:

$$0 < \min_{b \geq \beta_j} \psi_{ss}(s, b) \leq 1/\gamma_j, \qquad j = 1, 2, 3, \qquad \forall s \geq 13. \tag{90}$$

As a function of s, and for $b > 1$ fixed, $\psi_{ss}(s, b) = b^{-s} \log(b)^2$ is decreasing. Hence, it suffices to prove Equation 90 when $s = 13$. Choose $b \geq 1.3$. The equation $\psi_{ssb}(13, b) = 0$ has its unique solution in $[1, \infty)$ at the value $b = \exp(2/13) < 1.3$. Moreover, the function $\psi_{ss}(13, b)$ tends to 0 as $b \to \infty$. Hence the restriction of the function $b \to \psi_{ss}(13, b)$ to $[b, \infty)$ takes its maximum value at b. Evaluating at $b = 1.3, \sqrt{2}, \sqrt{3}$ we get Equation 90.

For $x, y \in [43/64, 1]$ we easily check the inequalities

$$A(-1, x) \geq 3, \quad B(-1, x) \geq 2, \quad C(-1, x, y) \geq (1.3)^2.$$

The quantities on the left are the square distances of the various pairs of points in the corresponding configuration on S^2. From this analysis we conclude that the 10 distances associated to a 5-point configuration parametrized by a point in Ψ_4 exceed 1.3, and at least 6 of them exceed $\sqrt{2}$, and at least 2 of them exceed $\sqrt{3}$. The same obviously holds for the TBP.

Now, 10 of the 20 terms comprising $\Theta_{ss}(x, y, s)$ are positive and 10 are negative. Also, for the terms of the same sign, all 10 of them are less than $1/440$, and at least 6 of them are less than $1/753$, and at least 2 of them are less than $1/4184$. Hence, by Equation 90, we have the final bound $|\Theta_{ss}| \leq (4/440) + (4/753) + (2/4184) < 1/64$. ♠

With the proof of Lemma C3, we have finished the proof of the Critical Theorem II.

All that remains is to prove Statements 1 and 2 of the Critical Theorem III. We first prove the derivative bounds we need for this and then we give the final argument. Let

$$I = \Big[\frac{55}{64}, \frac{56}{64}\Big]. \tag{91}$$

Lemma 7.8 (C4) $\Theta_{tts}(t,t,15) < 0, \qquad \forall t \in I.$

The file `LemmaC4.m` does the calculations for this proof. Because the s-energy of the TBP does not depend on the t-variable, we have

$$\Theta_{stt}(t,t,15) = 2A_{stt}|_{s=15} + 4B_{stt}|_{s=15} + 4C_{stt}|_{s=15} := \alpha(t) + \beta(t) + \gamma(t). \tag{92}$$

We write $f \sim f^*$ if

$$\frac{f}{f^*} = 2^u t^v (1+t^2)^w (2+t^2+t^{-2})^x$$

for exponents $u, v, w, x \in \boldsymbol{R}$. In this case, f and f^* have the same sign.

Step 1: Taking $(u,v,w,x) = (-14, 0, 11/2, 0)$ we have $\beta \sim -\beta^*$,

$$\beta^*(t) = (-2 + 30\log(2)) + t^2(-58 + 420\log(2)) - 15(1+14t^2)\log(1+t^2).$$

Noting that $\log(2) = 0.69...$ we eyeball β^* and see that it is positive for $t \in I$. The term $+420\log(2)t^2$ dominates. Hence $\beta < 0$ on I.

Step 2: Taking $(u,v,w,x) = (-41/2, -16, 12, 1/2)$ we have $\gamma \sim -\gamma^*$,

$$\gamma^*(t) = (-31 + 360\log(2)) + \underline{t^2(56 - 585\log(2))} + t^4(-29 + 315\log(2)) +$$
$$15(-8 + 13t^2 - 7t^4)\log(2 + t^2 + t^{-2}).$$

We have $\gamma^*(55/64) > 2^4$ and we estimate easily that $\gamma^*_t > -2^{10}$ on I. Only the underlined term has negative derivative in I. Noting that I has length 2^{-6}, we see that γ^* cannot decrease more than 2^4 as we move from x_0 to any other point of I. Hence $\gamma^* > 0$ on I. Hence $\gamma < 0$ on I.

Step 3: Taking $(u,v,w,x) = (-29, -14, 10, 3/2)$ we have $\alpha \sim -\alpha*$,

$$\alpha^*(t) = \gamma^*(t) + \delta^*(t), \qquad \delta^*(t) = 15\log 2 \times (8 - 13t^2 + 7t^4).$$

We see easily that $\delta^* > 0$ on I. So, from our result for γ^*, we have $\alpha^* > 0$ on I. Hence $\alpha < 0$ on I. ♠

Lemma 7.9 *For any* $\xi \in \widehat{\Psi}_8$ *let* $\Theta(s,\xi) = \mathcal{E}_s(\xi) - \mathcal{E}_s(\xi_0)$. *Then* $\Theta_s < 0$ *for* $s \in [15, 15_+]$.

Proof: Let $t_0 = 55/64$ be the left endpoint of the interval I. We compute that

$$\Theta_{st}(t_0, t_0, 15) < 0, \qquad \Theta_s(t_0, t_0, 15) < -2^{-7}. \tag{93}$$

The previous lemma now tells us that

$$\frac{d}{dt}\Theta_{st}(t,t,15) = \Theta_{tts} < 0, \qquad \forall t \in I. \tag{94}$$

The last two equations therefore combine to show that

$$\Theta_s(t,t,15 < -2^{-7}). \qquad \forall t \in I. \tag{95}$$

We also have the bound $|\Theta_{ss}| \le 2^{-6}$ on $[13,16] \times \Psi_4$. Hence

$$|\Theta_{ss}| \times |15_+ - 15| \le 2^{-6} \times \frac{25}{512} < 2^{-7}. \tag{96}$$

Hence $\Theta_s(s,t,t)$ varies by less than 2^{-7} as s ranges in $[15, 15_+]$. Hence $\Theta_s(s,t,t) < 0$ for all $s \in [15, 15_+]$ and all $t \in I$. ♠

Proof of Statements 1 and 2 of the Critical Theorem III: By the Critical Theorem II, we have $\Theta(15, *) > 0$ on $\Psi_8^\sharp$. We compute $\Theta(15_+, x, x) < 0$ for $x = 445/512 \in [55,56]/64$. Combining this with Lemma 7.9, we see that there exists a smallest parameter $\mathfrak{w} \in (15, 15_+)$ such that $\Theta(\mathfrak{w}, p^*) = 0$ for some $p^* \in \Psi_8^\sharp$. For $s > \mathfrak{w}$, Lemma 7.9 now says that $\Theta(s, p^*) < 0$. This establishes Statements 1 and 2 of the Critical Theorem III. ♠

8 The Energy Theorem

Reading Guide: This chapter is for Readers 5 and 6. For Reader 5, we prove the Energy Theorem in §9. For Reader 6, we use the Energy Theorem in our big computation in §10.

8.1 Background Definitions

We first give some background definitions and then we give our main result.

Energy Hybrids: We say that an *energy hybrid* is a potential of the form

$$F = \sum_{k=1}^{m} c_k G_k, \qquad G_k(r) = (4 - r^2)^k, \qquad c_1 \in \boldsymbol{Q}, \qquad c_2, ..., c_k \in \boldsymbol{Q}_+. \quad (97)$$

We normalize our avatars so that p_0 lies on the positive X-axis. In this way, and by stringing out the coordinates, we identify an avatar with a point in $\boldsymbol{R}^7 = \boldsymbol{R} \times (\boldsymbol{R}^2)^3$. Thus we think of the potential $\mathcal{E}_F$ as a function on $\boldsymbol{R}^7$. It will turn out that we only need to consider points in the cube $\square_{3/2}$ where

$$\square_r := [0, r] \times [-r, r]^r \times [-r, r]^r \times [-r, r]^2. \quad (98)$$

Dyadic Subdivision: The *dyadic subdivision* of a D-dimensional cube is the list of 2^D cubes obtained by cutting the cube in half in all directions. We sometimes blur this terminology and say that any one of these 2^D smaller cubes is a *dyadic subdivision* of the big cube.

Blocks: We define a *block* to be a product of the form

$$B = Q_0 \times Q_1 \times Q_2 \times Q_3 \subset \square_{3/2}, \quad (99)$$

where Q_0 is a segment and Q_1, Q_2, Q_3 are squares, each obtained by iterated dyadic subdivision respectively of $[0, 2]$ and $[-2, 2]^2$.

We call B *acceptable* if Q_0 has length at most 1 and Q_1, Q_2, Q_3 have sidelength at most 2. When B is acceptable, each Q_k is contained in a quadrant of $\boldsymbol{R}^2$.

8.2 The Main Result

We let $\mathcal{Q}$ denote the set of components of acceptable blocks. The elements of $\mathcal{Q}$ are either dyadic seqments in $[0, 3/2]$ or dyadic squares in $[-3/2, 3/2]^2$. Thanks to the subdivision process, each of these squares lies on one of the quadrants of the plane - it does not cross the coordinate axes. We also let $\{\infty\}$ be a member of $\mathcal{Q}$.

We first define 4 basic measurements we take of members in $\mathcal{Q}$.

0. The Flat Approximation: Given $Q \in \mathcal{Q}$ we define

$$Q^\bullet = \text{Convex Hull}(\Sigma^{-1}(v(Q)). \tag{100}$$

$Q^\bullet$ is either the point $(0, 0, 1)$, a chord of S^2 or else a convex planar quadrilateral with vertices in S^2 that is inscribed in a circle. We let $d_\bullet$ be the diameter of $Q_\bullet$. The quantity $d_\bullet^2$ is a rational function of the vertices of Q.

1. The Hull Approximation Constant: We think of $Q^\bullet$ as the linear approximation to

$$\widehat{Q} = \Sigma^{-1}(Q). \tag{101}$$

The constant we define here turns out to measure the distance between $\widehat{Q}$ and $Q^\bullet$. When $Q = \{\infty\}$ we define $\delta(Q) = 0$. Otherwise, let

$$\chi(D, d) = \frac{d^2}{4D} + \frac{(d^2)^2}{4D^3}. \tag{102}$$

This wierd function turns out to be an upper bound to a more geometrically meaningful non-rational function that computes the distance between an chord of length d of a circle of radius D and the arc of the circle it subtends.

When Q is a dyadic segment we define

$$\delta(Q) = \chi(2, \|\widehat{q}_1 - \widehat{q}_2\|). \tag{103}$$

Here q_1, q_2 are the endpoints of Q. When Q is a dyadic square we define

$$\delta(Q) = \max(s_0, s_2) + \max(s_1, s_3), \qquad s_j = \chi(1, \|q_j - q_{j+1}\|). \tag{104}$$

Here q_1, q_2, q_3, q_4 are the vertices of Q and the indices are taken cyclically. These are rational computations because $\chi(2, d)$ is a polynomial in d^2.

2. The Dot Product Estimator: By way of motivation, we point out that if $V_1, V_2 \in S^2$ then

$$G_k(\|V_1 - V_2\|) = (2 + 2V_1 \cdot V_2)^k.$$

Now suppose that Q_1 and Q_2 are two dyadic squares. We set $\delta_j = \delta(Q_j)$. Given any $p \in \boldsymbol{R}^2 \cup \infty$ let $\widehat{p} = \Sigma^{-1}(p)$. Define

$$Q_1 \cdot Q_2 = \max_{i,j}(\widehat{q}_{1i} \cdot \widehat{q}_{2j}) + (\tau) \times (\delta_1 + \delta_2 + \delta_1\delta_2). \tag{105}$$

Here $\{q_{1i}\}$ and $\{q_{2j}\}$ respectively are the vertices of Q_1 and Q_2. The constant τ is 0 if one of Q_1 or Q_2 is $\{\infty\}$ and otherwise $\tau = 1$. Finally, we define

$$T(Q_1, Q_2) = 2 + 2(Q_1 \cdot Q_2). \tag{106}$$

3. The Local Error Term: For $Q_1, Q_2 \in \mathcal{Q}$ and $k \geq 1$ we define

$$\epsilon_k(Q_1, Q_2) = \frac{1}{2}k(k-1)T^{k-2}d_1^2 + 2kT^{k-1}\delta_1, \tag{107}$$

$$d_1 = d_\bullet(Q_1), \quad \delta_1 = \delta(Q_1), \quad T = T(Q_1, Q_2).$$

The first term on the right in Equation 107 comes from the analysis of the flat approximation and the second term comes from the analysis of the difference between the flat approximation and the actual subset of the sphere. The quantity is not symmetric in the arguments, and $\epsilon_k(\{\infty\}, Q_2) = 0$.

4. The Global Error Estimate: Given $B = Q_0 \times Q_1 \times Q_2 \times Q_3$ let

$$\mathbf{ERR}_k(B) = \sum_{i=0}^{N} \mathbf{ERR}_k(B, i), \qquad \mathbf{ERR}_k(B, i) = \sum_{j \neq i} \epsilon(Q_i, Q_j). \tag{108}$$

More generally, when $F = \sum c_k G_k$ is as in Equation 97, we define

$$\mathbf{ERR}_F(B) = \sum_{k=0}^{N} \mathbf{ERR}_F(B, i), \qquad \mathbf{ERR}_F(B, i) = \sum |c_k| \ \mathbf{ERR}_k(B, i) \tag{109}$$

For the most part we only care about the $(+)$ case of the lemma. We only need the $(-)$ case when we deal with the potential $G_5 - 25G_1$.

Theorem 8.1 (Energy) *Let B be a acceptable block. Let $F = G_k$ for any $k \geq 1$ or $F = -G_1$. Then $\min_{p \in B} \mathcal{E}_F(v) \geq \min_{p \in v(B)} \mathcal{E}_k(v) - \mathbf{ERR}_k(B)$.*

9 Proof of the Energy Theorem

Reading Guide: This chapter is for Reader 5.

9.1 Guide to the Proof

Our proof of the Energy Theorem splits into two halves, an algebraic part and a geometric part. The algebraic part, which we do in this chapter, simply promotes a "local" result to a "global result". The geometric explains the meaning of the local error term $\epsilon_k(Q_1, Q_2)$ for $Q_1, Q_2 \in \mathcal{Q}$. Here $\mathcal{Q}$ is the space of components of good blocks, and also the point ∞.

The algebraic part involves what we call an *averaging system*. For the purpose of giving a uniform treatment, we treat every member of $\mathcal{Q}$ as a quadrilateral by the trick of repeating vertices. Thus, if we have a dyadic segment with vertices q_1, q_2 we will list them as q_1, q_1, q_2, q_2. For the point $\{\infty\}$ we will list the single vertex $q_1 = \infty$ as q_1, q_1, q_1, q_1. We say that an *averaging system* for a member of $\mathcal{Q}$ is a collection of maps $\lambda_1, \lambda_2, \lambda_3, \lambda_4 : Q \to [0,1]$ such that

$$\sum_{i=1}^{4} \lambda_i(z) = 1, \qquad \forall\ z \in Q.$$

The functions need not vary continuously. In case Q is a segment, we would have $\lambda_1 = \lambda_2$ and $\lambda_3 = \lambda_4$. In case $Q = \{\infty\}$ we would have $\lambda_j = 1/4$ for $j = 1, 2, 3, 4$.

We say that an *averaging system* for $\mathcal{Q}$ is a choice of averaging system for each member Q of $\mathcal{Q}$. The averaging systems for different members need not have anything to do with each other. In this chapter we will posit some additional properties of an averaging system and then prove the Energy Theorem under the assumption such such an averaging system exists. In the next chapter we will prove the existence of the desired averaging system.

Our naming system for the lemmas is designed to indicate the logic tree. Thus, the Energy Theorem follows from Lemma E1 and Lemma E2. Lemma E1 follows from Lemma E11 and Lemma E12. And so on.

9.2 Reduction to a Local Result

We fix the function $F = G_k$ for some $k \geq 1$ or else $F = -G_1$. We write $\mathcal{E} = \mathcal{E}_F$. We let $\epsilon = \epsilon_k$, as in Equation 107. Our algebraic argument would

work for any choice of F, but we need to use the choices above to actually get the averaging system we need. Let $q_{1,1}, q_{1,2}, q_{1,3}, q_{1,4}$ be the vertices of Q_1.

Lemma 9.1 (E1) *There exists an averaging system on $\mathcal{Q}$ with the following property: Let Q_1, Q_2 be distinct members of $\mathcal{Q}$. Given any $z_1 \in Q_1$ and $z_2 \in Q_2$ we have*

$$\sum_{i=1}^{4} \lambda_i(z_1) F(\|\widehat{q}_{1,i} - \widehat{z}_2\|) - F(\|\widehat{z}_1 - \widehat{z}_2\|) \le \epsilon(Q_1, Q_2). \tag{110}$$

We prove this result at the end of the chapter.

We are interested in 5-point configurations but we will work more generally so as to elucidate the general structure of the argument. We suppose that we have the good dyadic block $B = Q_0 \times ... \times Q_N$. The vertices of B 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 in each component member of the block. We define (as always, *via* inverse stereographic projection) the energy of the corresponding vertex configuration:

$$\mathcal{E}(I) = \mathcal{E}(q_{0,i_0}, ..., q_{N,i_N}) \tag{111}$$

Here is one more piece of notation. Given $z = (z_0, ..., z_n) \in B$ and a multi-index I we define

$$\lambda_I(z) = \prod_{i=0}^{N} \lambda_{i_j}(z_j). \tag{112}$$

Here λ_{i_j} is defined relative to the averaging system on Q_j.

Now we are ready to state our main global result. The global result uses the existence of an efficient averaging system. That is, it relies on the Energy Theorem1.

Lemma 9.2 (E2) *Let $z = (z_0, ..., z_N) \in B$. Then*

$$\sum_{I} \lambda_I(z) \mathcal{E}(I) - \mathcal{E}(z) \le \sum_{i=0}^{N} \sum_{j=0}^{N} \epsilon(Q_i, Q_j). \tag{113}$$

The lefthand sum is taken over all multi-indices. In the righthand sum, we set $\epsilon(Q_i, Q_i) = 0$ for all i.

Now let us deduce the Energy Theorem from Lemma E2. Notice that

$$\sum_I \lambda_I(z) = \prod_{j=0}^{N} \Big(\sum_{a=1}^{4} \lambda_a(z_j) \Big) = 1. \tag{114}$$

Choose some $(z_1, ..., z_N) \in B$ which minimizes $\mathcal{E}$. We have

$$0 \leq \min_{p \in v(B)} \mathcal{E}(v) - \min_{v \in B} \mathcal{E}(v) = \min_{p \in v(B)} \mathcal{E}(v) - \mathcal{E}(z) \leq^*$$

$$\sum_I \lambda_I(z)\mathcal{E}(I) - \mathcal{E}(z) \leq \sum_{i=0}^{N} \sum_{j=0}^{N} \epsilon(Q_i, Q_j). \tag{115}$$

The starred inequality comes from the fact that a minimum is less or equal to a convex average. The last expression is $\mathbf{ERR}(B)$ when $N = 4$ and $Q_4 = \infty$.

9.3 From Local to Global

Now we deduce the global Lemma E2 from the local Lemma E1.

Lemma 9.3 (E21) *Lemma E2 holds when $N = 1$.*

Proof: In this case, we have a block $B = Q_0 \times Q_1$. Setting $\epsilon_{ij} = \epsilon(Q_i, Q_j)$, Lemma E1 gives us

$$F(\|z_0 - z_1\|) \geq \sum_{\alpha=1}^{4} \lambda_\alpha(z_0) F(\|q_{0\alpha} - z_1\|) - \epsilon_{01}. \tag{116}$$

Applying Lemma E1 to the pair of points $(z_1, q_{0\alpha}) \in Q_1 \times Q_0$ we have

$$F(\|z_1 - q_{0\alpha}\|) \geq \sum_{\beta=1}^{4} \lambda_\beta(z_1) F(\|q_{1\beta} - q_{0\alpha}\|) - \epsilon_{10}. \tag{117}$$

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

$$F(\|z_0 - z_1\|) \geq \sum_{\alpha,\beta} \lambda_\alpha(z_0)[\lambda_\beta(z_1) F(\|q_{1\beta} - q_{0\alpha}\|) - \epsilon_{10}] - \epsilon_{01} =$$

$$\sum_{\alpha,\beta} \lambda_\alpha(z_0)\lambda_\beta(z_1) F(\|q_{1\beta} - q_{0\alpha}\|) - (\epsilon_{10} + \epsilon_{01}). \tag{118}$$

Equation 118 is equivalent to Equation 113 when $N = 1$. ♠

Now we do the general case.

Lemma 9.4 (E22) *Lemma E2 holds when $N \geq 2$.*

Proof: We rewrite Equation 118 as follows:

$$F(\|z_0 - z_1\|) \geq \sum_A \lambda_{A_0}(z_0)\lambda_{A_1}(z_1)\ F(\|q_{0A_0} - q_{1A_1}\|) - (\epsilon_{01} + \epsilon_{10}). \quad (119)$$

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

We also observe that

$$\sum_{I'} \lambda_{I'}(z') = 1, \qquad z' = (z_2, ..., z_N). \quad (120)$$

The sum is taken over all multi-indices $I' = (i_2, ..., i_N)$. Therefore, if we hold $A = (A_0, A_1)$ fixed, we have

$$\lambda_{A_0}(z_0)\lambda_{A_1}(z_1) = \sum_{I''} \lambda_{I''}(z). \quad (121)$$

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

$$F(\|z_0 - z_1\|) \geq \sum_I \lambda_I(z) F(\|q_{0I_0} - q_{1I_1}\|) - (\epsilon_{01} + \epsilon_{10}). \quad (122)$$

The same argument works for other pairs of indices, giving

$$F(\|z_i - z_j\|) \geq \sum_I \lambda_I(z) F(\|q_{iI_i} - q_{jI_j}\|) - (\epsilon_{ij} + \epsilon_{ji}). \quad (123)$$

Let us restate this as $X_{ij} - Y_{ij} \geq Z_{ij}$, where

$$X_{ij} = \sum_I \lambda_I(z) F(\|q_{iI_i} - q_{jI_j}\|), \quad Y_{ij} = F(\|z_i - z_j\|), \quad Z_{ij} = \epsilon_{ij} + \epsilon_{ji}.$$

When we sum Y_{ij} over all $i < j$ we get the second term in Equation 113. When we sum Z_{ij} over all $i < j$ we get the third term in Equation 113. When we sum X_{ij} over all $i < j$ we get

$$\sum_{i<j} \Big(\sum_I \Lambda_I(z) F(\|q_{iI_i} - q_{jI_j}\|) \Big) = \sum_I \sum_{i<j} \Lambda_I(z)\ F(\|q_{iI_i} - q_{jI_j}\|) =$$

$$\sum_I \Lambda_I(z) \Big(\sum_{i<j} F(\|q_{iI_i} - q_{jI_j}\|) \Big) = \sum_I \lambda_I(z) \mathcal{E}(I).$$

This is the first term in Equation 113. This proves Lemma E2. ♠

9.4 The Efficient Averaging System

The rest of the chapter is devoted to proving Lemma E1. Lemma E1 posits the existence of what we call an efficient averaging system. Here we define it. Recall that $Q^\bullet$ is the convex hull of the vertices $\widehat{q}_1, \widehat{q}_2, \widehat{q}_3, \widehat{q}_4$ of $\widehat{Q} = \Sigma^{-1}(Q)$. What we want from the system is that for any $z^\bullet \in Q^\bullet$

$$z^\bullet = \sum_{i=1}^{4} \lambda_i(z^\bullet)\widehat{q}_i. \tag{124}$$

If $z^\bullet$ lies in the convex hull of $\widehat{q}_1$, $\widehat{q}_2$, $\widehat{q}_3$, then we let $\lambda_1(z^\bullet)$, $\lambda_2(z^\bullet)$, $\lambda_3(z^\bullet)$ be barycentric coordinates on this triangle and we set $\lambda_4(z^\bullet) = 0$. If $z^\bullet$ lies in the convex hull of $\widehat{q}_1$, $\widehat{q}_2$, $\widehat{q}_4$, then we let $\lambda_1(z^\bullet)$, $\lambda_2(z^\bullet)$, $\lambda_4(z^\bullet)$ be barycentric coordinates on this triangle and we set $\lambda_3(z^\bullet) = 0$. This definition agrees on the overlap, which is the line segment joining $\widehat{q}_3$ to $\widehat{q}_4$.

To get our averaging system on $Q \in \mathcal{Q}$ we define

$$\lambda_j(z) = \lambda_j(z^\bullet), \tag{125}$$

where $z^\bullet$ is some choice of point in $Q^\bullet$ which is <u>closest</u> to $\widehat{z}$. If there are several closest points we pick the one (say) which has the smallest first coordinate. We prove Lemma E1 with respect to the averaging system above.

9.5 Reduction to Simpler Statements

Let F be either G_k for some $k \geq 1$ or else $F = -G_1$. For convenience we expand out the statement of Lemma E1.

Lemma 9.5 (E1) *The efficient averaging system on $\mathcal{Q}$ has the following property. Let Q_1, Q_2 be distinct members of $\mathcal{Q}$. Given any $z_1 \in Q_1$ and $z_2 \in Q_2$ we have*

$$\sum_{i=1}^{4} \lambda_i(z_1) F(\|\widehat{q}_{1,i} - \widehat{z}_2\|) - F(\|\widehat{z}_1 - \widehat{z}_2\|) \leq \frac{1}{2}k(k-1)T^{k-2}d_1^2 + 2kT^{k-1}\delta_1. \tag{126}$$

Here δ_1 and d_1 respectively are the Hull Approximation constant and diameter of Q_1, and

$$T = 2 + 2(Q_1 \cdot Q_1), \qquad Q_1 \cdot Q_2 = \max_{i,j}(\widehat{q}_{1,i} \cdot \widehat{q}_{2,j}) + (\tau) \times (\delta_1 + \delta_2 + \delta_1\delta_2). \tag{127}$$

$\tau = 0$ or $\tau = 1$ depending on whether one of Q_1, Q_2 is $\{\infty\}$. We are maximizing over the dot product of the vertices and then either adding an error term or not. Define

$$X_\bullet = F(z_1^\bullet - \widehat{z}_2) = (2 + 2z_1^\bullet \cdot \widehat{z}_2)^k \quad \text{or} \quad -2 - 2z_1^\bullet \cdot \widehat{z}_2. \tag{128}$$

Lemma E1 is an immediate consequence of the following two results.

Lemma 9.6 (E11) $\sum_{i=1}^4 \lambda_i(z_1) F(\|\widehat{q}_{1,i} - \widehat{z}_2\|) - X_\bullet \le \frac{1}{2}k(k-1)T_\bullet^{k-2} d_1^2$.

Lemma 9.7 (E12) $X_\bullet - F(\|\widehat{z}_1 - \widehat{z}_2\|) \le 2kT^{k-1}\delta$.

9.6 Proof of Lemma E11

Suppose first $F = -G_1$. We hold $\widehat{z}_2$ fixed and define

$$L(\widehat{q}) = F(\|\widehat{q} - \widehat{z}_2\|) = -2 - 2\widehat{q} \cdot \widehat{z}_2.$$

Lemma E2, in this special case, says that

$$\sum_{i=1}^4 \lambda_i(z_1) L(\widehat{q}_{1,i}) - L(z_1^\bullet) = 0.$$

But this follows from Equation 125 and the (bi) linearity of the dot product.

Now we deal with the case where $F = G_k$ for $k \ge 1$. We prove the following two lemmas at the end of the chapter.

Lemma 9.8 (E111) *For $j = 1, 2$ let γ_j be a point on a line segment connecting a point of $\widehat{Q}_j$ to a closest point on $Q_j^\bullet$. Then $\gamma_1 \cdot \gamma_2 \le Q_1 \cdot Q_2$.*

Lemma 9.9 (E112) *Let $M \ge 2$ and $k = 1, 2, 3....$ Suppose*

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

Then

$$0 \le \sum_{i=1}^M \lambda_i x_i^k - \Big(\sum_{I=1}^M \lambda_i x_i\Big)^k \le \frac{1}{8}k(k-1)x_M^{k-2}\ (x_M - x_1)^2. \tag{129}$$

Recall that $q_{1,1}, q_{1,2}, q_{1,3}, q_{1,4}$ are the vertices of Q_1. Let $\lambda_i = \lambda_i(z_1)$. We set

$$x_i = 4 - \|\widehat{q}_{1,i} - \widehat{z}_2\|^2 = 2 + 2\widehat{q}_{1,i} \cdot \widehat{z}_2, \qquad i = 1, 2, 3, 4. \tag{130}$$

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

$$\sum_{i=1}^{4} \lambda_i(z) F(\|q_{1,i} - z_2\|) = \sum_{i=1}^{4} \lambda_i x_i^k, \tag{131}$$

$$X_\bullet = (2 + 2z_1^\bullet \cdot \widehat{z}_2)^k = \Big(\sum_{i=1}^{4} \lambda_i \times (2 + \widehat{q}_i \cdot \widehat{z}_2)\Big)^k = \Big(\sum_{i=1}^{4} \lambda_i x_i\Big)^k. \tag{132}$$

By Equation 131, Equation 132, and the case $M = 4$ of Lemma E112, we have

$$\sum_{i=1}^{4} \lambda_i(z) F(\|q_{1,i} - z_2\|) - X_\bullet = \sum_{i=1}^{4} \lambda_i x_i^k - \Big(\sum_{i=1}^{4} \lambda_i x_i\Big)^k \leq \frac{1}{8} k(k-1) x_4^{k-2} (x_4 - x_1)^2. \tag{133}$$

By Lemma E111

$$x_4 = 2 + 2(\widehat{q}_4 \cdot \widehat{z}_2) \leq T. \tag{134}$$

Since d_1 is the diameter of $Q_1^\bullet$, and $\widehat{z}_2$ is a unit vector,

$$x_4 - x_1 = 2\widehat{z}_2 \cdot (\widehat{q}_4 - \widehat{q}_1) \leq 2\|\widehat{q}_4 - \widehat{q}_1\| \leq 2d_1 \tag{135}$$

Plugging Equations 134 and 135 into Equation 133, we get Lemma E12.

9.7 Proof of Lemma E12

Let $\delta(Q)$ be the hull approximation constant for $Q \in \mathcal{Q}$, as defined (depending on Q) in Equation 103 or Equation 104.

Lemma 9.10 (E121) *Let Q be any good dyadic square or segment. Then every point of $\widehat{Q}$ is within $\delta(Q)$ of the quadrilateral $Q^\bullet$.*

Lemma E121 implies that $\|\widehat{z}_1 - z_1^\bullet\| < \delta(Q)$. Let γ_1 denote the unit speed line segment connecting $z_1^\bullet$ to $\widehat{z}_1$. The length L of γ_1 is at most δ_1, by Lemma E11. So, $\gamma_1(0) = z_1^\bullet$ and $\gamma_1(L) = \widehat{z}_1$. Define

$$f(t) = \Big(2 + 2\widehat{z}_2 \cdot \gamma_1(t)\Big)^k \text{ or } -2 - 2\widehat{z}_2 \cdot \gamma_1(t), \tag{136}$$

depending on the case. The argument we give works equally well more generally when we use $F = \pm G_k$.

We have $f(0) = X_\bullet$ and $f(L) = F(\|\widehat{z}_1 - \widehat{z}_2\|)$. Hence

$$X_\bullet - F(\|\widehat{z}_1 - \widehat{z}_2\|) = f(0) - f(L), \qquad L \le \delta_1. \tag{137}$$

Combining the Chain Rule, the Cauchy-Schwarz inequality, and Lemma E111, we have

$$|f'(t)| = \left|(2\widehat{z}_2 \cdot \gamma_1'(t)) \times k\Big(2 + 2\widehat{z}_2 \cdot \gamma_1(t)\Big)^{k-1}\right| \le$$

$$2k\left|(2 + 2\widehat{z}_2 \cdot \gamma_1(t))\right|^{k-1} \le 2k(2 + 2(Q_1 \cdot Q_2))^{k-1} = 2kT^{k-1}.$$

In short

$$|f'(t)| \le 2kT^{k-1}. \tag{138}$$

Lemma E13 follows Equation 138, Equation 137, and integration.

9.8 Proof of Lemma E111

See Equation 127 for the definition of $Q_1 \cdot Q_2$. We first treat the case $\tau = 1$, meaning that neither Q_1 nor Q_1 is $\{\infty\}$. Since the dot product is bilinear,

$$q_1^\bullet \cdot q_2^\bullet \le \max_{i,j}(\widehat{q}_{1i} \cdot \widehat{q}_{2j}). \tag{139}$$

By Lemma E11, and by hypothesis, we can find points $z_1^\bullet$ and $z_2^\bullet$ such that

$$\gamma_j = z_1^\bullet + h_1, \qquad \gamma_2 = z_2^\bullet + h_2, \qquad \|h_j\| \le \delta_j.$$

But then by the triangle inequality and the Cauchy-Schwarz inequality

$$|(\gamma_1 \cdot \gamma_2) - (z_1^\bullet \cdot z_2^\bullet)| \le |z_1^\bullet \cdot h_2| + |z_2^\bullet \cdot h_1| + |h_1 \cdot h_2| \le \delta_1 + \delta_2 + \delta_1\delta_2.$$

This combines with Equation 139 to complete the proof when $\tau = 1$.

Suppose $\tau = 0$. Without loss of generality assume that $Q_2 = \{\infty\}$. The maximum of $\widehat{q}_1 \cdot (0, 0, 1)$, for $q_1 \in Q_1$, is achieved when q_1 is vertex of Q_1. At the same time, the maximum of $q_1^\bullet \cdot (0, 0, 1)$, for $q_1^\bullet \in Q_1^\bullet$ is achieved when $q_1^\bullet$ is a vertex of $Q_1^\bullet$. But then our lemma is true for the endpoints of the segment containing γ. Since the dot product with $(0, 0, 1)$ varies linearly along this line segment, the same result is true for all points on the line segment.

9.9 Proof of Lemma E112

Lemma 9.11 (E1121) *Suppose $a, x \in [0,1]$ and $k \geq 2$. Then $f(x) \leq g(x)$, where*

$$f(x) = (ax^k + 1 - a) - (ax + 1 - a)^k; \qquad g(x) = \frac{1}{8}k(k-1)(1-x)^2. \quad (140)$$

Proof: Since $f(1) = g(1) = f'(1) = g'(1) = 0$ the Cauchy Mean Value Theorem (applied twice) tells us that for any $x \in (0,1)$ there are values $y < z \in [x,1]$ such that

$$\frac{f(x)}{g(x)} = \frac{f'(y)}{g'(y)} = \frac{f''(z)}{g''(z)} = 4az^{k-2}\Big[1 - a\Big(a + \frac{1-a}{z}\Big)^{k-2}\Big] \leq 4a(1-a) \leq 1. \quad (141)$$

This completes the proof. ♠

Remark: The above proof, suggested by an anonymous referee of [**S4**], is better than my original proof.

Now we prove the main inequality The lower bound is a trivial consequence of convexity, and both bounds are trivial when $k = 1$. So, we take $k = 2, 3, 4, ...$ and prove the upper bound. Suppose first 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 $M - 1$. Hence, it suffices to prove the inequality when $M = 2$.

In the case $M = 2$, we set $a = \lambda_1$. Both sides of the inequality in Lemma E112 are homogeneous of degree k, so it suffices to consider the case when $x_2 = 1$. We set $x = x_1$. Our inequality then becomes exactly the one treated in Lemma E1121. This completes the proof.

9.10 Proof of Lemma E121

We remind the reader of the wierd function $\chi(D)$ and we introduce a more geometrically meaningfun function

$$\chi(D,d) = \frac{d^2}{4D} + \frac{d^4}{4D^3}, \qquad \chi^*(D,d) = \frac{1}{2}(D - \sqrt{D^2 - d^2}). \quad (142)$$

Lemma 9.12 (E1211) *$\chi^*(D,d) \le \chi(D,d)$ for all $d \in [0,D]$.*

Proof: By homogeneity, it suffices to prove the result when $D = 1$. To simpify the algebra we define $A = 2\chi(1,d) - 1$ and $A^* = 2\chi^*(1,d) - 1$. We compute $4A^2 - 4(A^*)^2 = d^4(d-1)(d+1)(d^2+3)$. Hence, the sign of $A - A^*$ does not change on $(0,1)$. We check that $A > A^*$ when $d = 1/2$. Hence $A > A^*$ on $(0,1)$. This implies the inequality. ♠

Segment Case: Let Q be dyadic segment. Here $\widehat{Q}$ is the arc of a great circle and $Q^\bullet$ is the chord of the arc joining the endpoints of this arc. Let d be the length of $Q^\bullet$. The point of $\widehat{Q}$ farthest from $Q^\bullet$ is the midpoint of this $\widehat{Q}$. Let x be the distance between the midpoint of $\widehat{Q}$ and the midpoint of $Q^\bullet$. From elementary geometry, $x(D-x) = (d/2)^2$. Solving for x we find that $x = \chi^*(2,d)$. Lemma E1211 finishes the proof.

Square Case: Let Q be a dyadic square and let $z \in Q$ be a point. Let L be the vertical line through x and let z_{01}, z_{23} be the endpoints of the segment $L \cap Q$. We label the vertices of Q (in cyclic order) so that z_{01} lies on the edge joining q_0 to q_1 and z_{23} lies on the edge joining q_2 to q_3.

If M is a horizontal line intersecting Q then the circle $\Sigma^{-1}(M \cup \infty)$ has diameter at least 1. The point is that this circle contains $(0,0,1)$ and also $\Sigma^{-1}(0,y)$ for some $|y| \le 3/2$. In fact the diameter is at least $4/\sqrt{13}$. The same goes for vertical lines intersecting Q.

Define $d_j = \|\widehat{p}_j - \widehat{p}_{j+1}\|$ with the indices taken cyclically. The length of the segment σ joining the endpoints of $\Sigma^{-1}(L \cap Q)$ varies monotonically with the position of L. Hence, σ has length at most $\max(d_1, d_3)$. At the same time, $\Sigma^{-1}(L \cap Q)$ is contained in a circle of diameter at least 1. The same argument as in the segment case now shows that there is a point $z^* \in \sigma$ which is within $t_{13} = \max(\chi(1,d_1), \chi(1,d_3))$ of $\widehat{z}$.

The endpoints of σ respectively are on the spherical arcs obtained by mapping the top and bottom edge of Q onto S^2 via Σ^{-1}. Hence, one endpoint of σ is within $\chi(1,d_0)$ of a point on the corresponding edge of $\partial Q^\bullet$ and the other endpoint of σ is within $\chi(1,d_2)$ of a point on the opposite edge of $\partial Q^\bullet$. But that means that either endpoint of σ is within $t_{02} = \max(\chi(1,d_0), \chi(1,d_2))$ of a point in $Q^\bullet$. But then every point of the segment σ is within t_{02} of some point of the line segment joining these two points of $Q^\bullet$. In particular, there is a point $z^\bullet \in Q^\bullet$ which is within t of z^*. The triangle inequality completes the proof of Lemma E121.

10 The Calculation Theorem

Reading Guide: This chapter is for Reader 6. We prove the Calculation Theorem from §3.7

10.1 A Preliminary Lemma

We first prove a result that cuts down on our calculation time. With the exception of the potential $G_5^\flat$ the remaining potentials are strictly monotone in the sense that the functions decrease as the distance increases.

Lemma 10.1 *Let F be a strictly monotone decreasing potential and suppose that $\xi = (p_0, p_1, p_2, p_3)$ is an avatar. If $\min(p_{1k}, p_{2k}, p_{3k}) > 0$ for one of $k = 1, 2$ then ξ does not minimize the F-potential.*

Proof: The corresponding 5-point configuration in S^2 is contained in a hemisphere H, and at least 3 of the points are in the interior of H. If we reflect one of the interior points across ∂H then we increase at least 2 of the distances in the configuration and keep the rest the same. ♠

10.2 The Four Calculation Ingredients

We say that a *rational block computation* is a finite calculation, only involving the arithmetic operations and min and max. The output of a rational block computation will be one of two things: **yes**, or an integer. A return of an integer is a statement that the computation does not definitively answer to the question asked of it. If the integer is -1 then there is no more information to be learned. If the integer lies in $\{0, 1, 2, 3\}$ we use this integer as a guide in our algorithm. Let Ω_0 and Υ be as in the Calculation Theorem.

Ingredient 1: We describe a rational block computation C_1 such that an output of **yes** for a block B implies that $B \subset \Omega_0$.

Define intervals $I_0, I_1, I_{\sqrt{3}/3}$ such that

$$I_0 = [-2^{-17}, 2^{-17}], \quad I_1 = [1 - 2^{-17}, 1 + 2^{-17}] \quad 2^{30} I_{\sqrt{3}/3} = [619916940, 619933323] \tag{143}$$

$I_{\sqrt{3}/3}$ is a rational interval that is just barely contained inside the interval of length 2^{-17} centered at $\sqrt{3}/3$. Define

$$\Omega_{00} = (I_1 \times \{0\}) \times (I_0 \times -I_{\sqrt{3}/3}) \times (-I_1 \times I_0) \times (I_0 \times I_{\sqrt{3}/3}). \tag{144}$$

We have $\Omega_{00} \subset \Omega_0$, though just barely. There are 128 vertices of B. We simply check whether each of these vertices is contained in Ω_{00}. If so then we return **yes**. In practice our program scales up all the coordinates by 2^{30} so that this test just involves integer comparisons.

Ingredient 2: We describe a rational block computation C_3 such that an output of **yes** for an acceptable block B implies that either B is disjoint from the interior of Ω or else all configurations in B are elliminated by Lemma 10.1.

Let $B = Q_0 \times Q_1 \times Q_2 \times Q_3$ be an acceptable block. These blocks are such that the squares Q_1, Q_2, Q_3 do not cross the coordinate axes. For such squares, the minimum and maximum norm of a point in the square is realized at a vertex. Thus, we check that a square lies inside (respectively outside) a disk of radius r centered at the origin by checking that the square norms of each vertex is at most (respectively at least) r^2.

We check whether there is an index $j \in \{1,2,3\}$ such that all vertices of Q_j have norm at least $\max Q_0$. We return **yes** if this happens, because then all avatars in the interior of B will have some p_j with $\|p_j\| > \|p_0\|$.

We check whether there is an index $j \in \{1,2,3\}$ such that all vertices of Q_j have norm at least $3/2$. If so, we return **yes**. If this happens then $\|p_0\|, \|p_j\| > 3/2$ for all avatars in the interior of B.

We count the number a of indices j such that the vertices of Q_j all have norm at most $1/2$. We then count the number b of indices j such that all vertices of Q_j have norm at least $1/2$. We return **yes** if a is odd and $a+b=4$. In this case, every avatar in the interior of B is odd.

We write $I \le J$ to indicate that all values in an interval I are less or equal to all values in an interval J. We also allow I and J to be single points in this notation. For each $j = 0,1,2,3$ we let Q_{jk} be the projection of Q_j onto the kth factor. Thus Q_{j1} and Q_{j2} are both line segments in $\boldsymbol{R}$.

We return **yes** for each of the following reasons:

- If $Q_{jk} \le -3/2$ or $Q_{jk} \ge 3/2$ for any $j = 1,2,3$ and $k = 1,2$.
- $Q_{12} \ge Q_{22}$ or $Q_{12} \ge Q_{32}$ or $Q_{22} \ge Q_{32}$ or $Q_{22} \le 0$.

- $Q_{j1} \geq 0$ for $j = 1, 2, 3$, unless we are working with $G_5^\flat$.
- $Q_{j2} \geq 0$ for $j = 1, 2, 3$, unless we are working with $G_5^\flat$.

Lemma 10.1 justifies the use of the last two criteria.

Ingredient 3: We describe a rational block computation $C_3^\sharp$ such that an output of **yes** for a block B implies that $B \subset \Upsilon$. Likewise, there exists a rational block computation $C_3^{\sharp\sharp}$ such that an output of **yes** for a block B implies that B is disjoint from Υ.

For $C_3^\sharp$ we return **yes** if all the vertices of B lie in Υ. For $C_3^{\sharp\sharp}$ we return **yes** if one of the factors of B is disjoint from the corresponding factor of Υ. This amounts to checking whether a pair of rational squares in the plane are disjoint. We do this using the projections defined for Ingredient 2.

Ingredient 4: For any function F given by Equation 97, we describe a rational block computation $C_{4,F}$ such that an output of **yes** for an acceptable block B implies that the minimum of $\mathcal{E}_F$ on B is at least $\mathcal{E}_F(\xi_0) + 2^{-50}$. Otherwise $C_{4,F}(B)$ is an integer in $\{0, 1, 2, 3\}$. Our calculation refers to the Energy Theorem from §8.

Let B be an acceptable block. Let F be an energy hybrid. Let $[F]$ denote the F-potential of the TBP. If

$$\min_{p \in v(B)} \mathcal{E}_F(v) - \mathbf{ERR}_k(B) \geq [F] + 2^{-50} \tag{145}$$

we return **yes**. Otherwise we return the index i such that $\mathbf{ERR}_F(B, i)$ is the largest. In case of a tie, which probably never happens, we pick the lowest such index. ♠

10.3 The Computational Algorithm

Here is the main calculation.

1. We start with the list $L = \{\square\}$.
2. If $L = \emptyset$ then **HALT**. Otherwise let $B = Q_0 \times Q_1 \times Q_2 \times Q_3$ be the last block of L.

3. If B is not acceptable we delete B from L and append to L the subdivision of B along the offending index. We then return to Step 2. Any blocks considered beyond this step are acceptable.

4. If $C_1(B) = \textbf{yes}$ or $C_2(B) = \textbf{yes}$ we remove B from L and go to Step 2. Here we are eliminating blocks disjoint from the interior of Ω or else contained in Ω_0.

5. If $F = G_{10}^{\sharp}$ and $C_3^{\sharp}(B) = \textbf{yes}$ we remove B from L and go to Step 2. If $F = G_{10}^{\sharp\sharp}$ and $C_3^{\sharp\sharp}(B) = \textbf{yes}$ we remove B from L and go to Step 2.

6. If $C_{4,F}(B) = \textbf{yes}$ then we remove B from L and go to Step 2. Here we have verified that the F-energy of any avatar in B exceeds $[F] + 2^{-50}$.

7. If $C_{4,F}(B) = k \in \{0, 1, 2, 3\}$ then we delete B from L and append to L the blocks of the subdivision $S_k(B)$ and return to step 2.

Remark: There is one fine point of our calculation. We eliminate blocks which are disjoint from the *interior* of Ω (or the interior of the set ruled out by Lemma 10.1). This is not a problem because any point in the boundary is also contained in a block that is not disjoint from the interior of our domain.

10.4 Discussion of the Implementation

Representing Blocks: We represent the coordinates of blocks by `longs`, which have 31 digits of accuracy. What we list are 2^{30} times the coordinates. Our algorithm never does so many subdivisions that it defeats this method of representation. In all but the main step (Lemma A134) in the algorithm below we compute with exact integers. When the calculation (such as squaring a `long`) could cause an overflow error, we first recast the `longs` as a `BigIntegers` in Java and then do the calculations.

Interval Arithmetic: For the main step of the algorithm we use interval arithmetic. We use the same implementation as we did in [**S1**], where we explain it in detail. Here is how it works in brief. If we have a calculation involving numbers $r_1, ..., r_n$, and we produce intervals $I_1, ..., I_n$ with dyadic rational numbers represented exactly by the computer such that $r_i \in I_i$ for $i = 1, ..., n$. We then perform the usual arithmetic operations on the intervals, rounding outward at each step. The final output of the calculation, an interval, contains the result of the actual calculation.

In our situation here, the numbers $r_1, ..., r_n$ are, with one exception, dyadic rationals. (The exception is that the coordinates of the point representing the TBP are quadratic irrationals.) In principle we could do the entire computation, save for this one small exception, with expicit integer arithmetic. However, the complexity of the rationals involved, meaning the sizes of their numerators and denominators, qets quite large this way and the calculation is too slow.

One way to think about the difference between our explicitly defined exact integer arithmetic and interval arithmetic is that the integer arithmetic interrupts the calculation at each step and rounds outward so as to keep the complexity of the rational numbers from growing too large.

Guess and Check: Here is how we speed up the calculation. When we do Steps 6-7, we first do the calculation $C_{4,F}$ using floating point operations. If the floating version returns an integer, we use this integer to subdivide the box and return to step 2. If $C_{4,F}$ says **yes** then we retest the box using the interval arithmetic. In this way, we only pass a box for which the interval version says **yes**. This way of doing things keeps the calculation rigorous but speeds it up by using the interval arithmetic as sparingly as possible.

Parallelization: We also make our calculation more flexible using some parallelization. We classify each block $B = Q_0 \times Q_1 \times Q_2 \times Q_3$ with a number in $\{0, ..., 7\}$ according to the formula

$$\text{type}(B) = \sigma(c_{01} - 1) + 2\sigma(c_{11}) + 4\sigma(c_{31}) \in \{0, ..., 7\}.$$

Here c_{j1} is the first coordinate of the center of B_j and $\sigma(x)$ is 0 if $x < 0$ and 1 if $x > 0$. Step 3 of our algorithm guarantees that $\sigma(\cdot)$ is always applied to nonzero numbers.

We wrote our program so that we can select any subset $S \subset \{0, ..., 7\}$ we like and then (after Step 3) automatically pass any block whose type is not in S. To be able to do the big calculations in pieces, we run the program for various subsets of $\{0, ..., j\}$, sometimes in parallel.

10.5 Record of the Calculation

If the algorithm reaches the **HALT** state for a given choice of F, this constitutes a proof that the corresponding statement of the Computation Theorem

is true. In fact this happens in all cases. Here I give an account of one time I ran the computations to completion during January 2023 using the computer discussed at the end of the introduction. In listing the calculations I will give the approximate time and the exact number of blocks passed. Since we use floating point calculations to guide the algorithm, the sizes of the partitions can vary slightly with each run.

For G_4 : 2 hrs 14 min, 10848537 blocks.
For G_6: 5 hr 11 min, 25159337 blocks.
For $G_5^\flat$ types 1&2: 2 hr 31 min, 6668864 blocks.
For $G_5^\flat$ types 3&4: 1 hr 55 min, 4787489 blocks.
For $G_5^\flat$ types 5&6: 5 hr 33 min, 14160332 blocks.
For $G_5^\flat$ types 7&8: 3 hr 49 min, 9219550 blocks.
For $G_{10}^\sharp$ type 1: 4 hr 23 min, 6885912 blocks.
For $G_{10}^\sharp$ type 2: 9 hr 47 min, 15982122 blocks.
For $G_{10}^\sharp$ type 3: 3 hr 47 min, 5872029 blocks.
For $G_{10}^\sharp$ type 4: 7 hr 59 min, 13475260 blocks.
For $G_{10}^\sharp$ type 5: 8 hr 30 min, 13313492 blocks.
For $G_{10}^\sharp$ type 6: 15 hr 16 min, 24110457 blocks.
For $G_{10}^\sharp$ type 7: 5 hr 19 min, 7862780 blocks.
For $G_{10}^\sharp$ type 8: 8 hr 33 min, 13478467 blocks.
For $G_{10}^{\sharp\sharp}$ (on the domain Υ): 28 minutes, 805242 blocks.

11 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)

[**BHS**], S. V. Bondarenko, D. P. Hardin, E.B. Saff, *Mesh Ratios for Best Packings and Limits of Minimal Energy Configurations*,

[**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

[**Fö**], Föppl *Stabile Anordnungen von Electron in Atom*, J. fur die Reine Agnew Math. **141**, 1912, pp 251-301.

[**HZ**], Xiaorong Hou and Junwei Zhao, *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

[**KY**], A. V. Kolushov and V. A. Yudin, *Extremal Dispositions of Points on the Sphere*, Anal. Math **23** (1997) 143-146

[**MKS**], T. W. Melnyk, O. Knop, W.R. Smith, *Extremal arrangements of point and and unit charges on the sphere: equilibrium configurations revisited*, Canadian Journal of Chemistry 55.10 (1977) pp 1745-1761

[**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

[**S0**] R. E. Schwartz, *Divide and Conquer: A Distributed Approach to* 5*-Point Energy Minimization*, Research Monograph (preprint, 2023)

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

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

[**S3**] R. E. Schwartz, *Lengthening a Tetrahedron*, Geometriae Dedicata, 2014.

[**S4**], R. E. Schwartz, *Five Point Energy Minimization: A Summary*, Journal of Constructive Approximation (2019)

[**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