# Five Point Energy Minimization 2: Big Calculation

Richard Evan Schwartz

April 12, 2024

**Abstract**

This is Paper 2 of series of 7 self-contained papers which together prove the Melnyk-Knopf-Smith phase transition conjecture for 5-point energy minimization. (Paper 0 has the main argument.) This paper deals with the big computer assisted part of the proof.

## 1 Introduction

### 1.1 Context

During the past decade I have written several versions of a proof that rigorously 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. See [**S0**] for the latest version. This work implies and extends my solution [**S1**] of Thomson's 1904 5-electron problem [**Th**]. Unfortunately, after a number of attempts I have not been able to publish my work on this. Even though I have taken great pains to make the proof modular and checkable, the monograph still gives the impression of being too difficult to referee.

Now I am taking a new approach. I have broken down the proof into a series of 7 independent papers, each of which may be checked without any reference to the others. The longest of the papers is 20 pages. The drawback of this approach is twofold. First, there will necessarily be some redundancy in these papers. Second, none of the papers has a blockbuster result in itself. To help offset the second drawback, I will state the main result in full in each paper, and I will try to explain how the small result proved in each paper relates to the overall goal.

## 1.2 The Phase Transition Result

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

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

This quantity is commonly called the *F-potential* or the *F-energy* of $P$. A configuration $P$ is a *minimizer* for $F$ if $\mathcal{E}_F(P) \le \mathcal{E}_F(P')$ for all other $N$-point configurations $P'$.

We are interested in the *Riesz potentials*:

$$R_s(d) = d^{-s}, \qquad s > 0. \tag{2}$$

$R_s$ is also called a *power law potential*, and $R_1$ is specially called the *Coulomb potential* or the *electrostatic potential*. The question of finding the $N$-point minimizers for $R_1$ is commonly called *Thomson's problem*.

We consider the case $N = 5$. 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.

Define

$$15_+ = 15 + \frac{25}{512}. \tag{3}$$

**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* $R_s$.

The proof has many moving parts. The largest part involves eliminating all the configurations and energy exponents outside a set of the form $\Upsilon \times [13, 15^+]$ using a computer-assisted divide-and-conquer algorithm. This paper details the divide-and-conquer calculation.

### 1.3 The Result in This Paper

In order to state the precise result proved here, I first need to introduce some background information.

**Stereographic Projection:** 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.

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

Here is the inverse map:

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

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

**Avatars:** 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 $S^2$. By a slight abuse of notation we write $\mathcal{E}_F(p_1,p_2,p_3,p_4)$ when we mean the $F$-potential of the corresponding 5-point configuration.

Figure 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*. The points for the even avatar are $(\pm 1, 0)$ and $(0, \pm\sqrt{3}/3)$. 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.

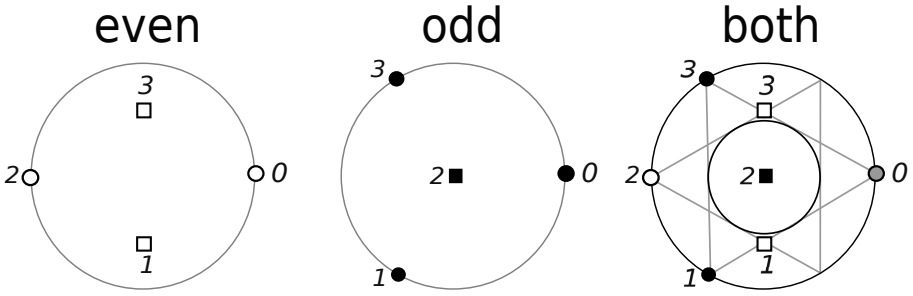


**Figure 1:** Even and odd avatars of the TBP.

**Even and Odd Avatars:** 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. This extends our definition for the TBP avatars.

We call 2 avatars *isomorphic* if the corresponding 5-point configurations on $S^2$ are isometric. Every avatar is isomorphic to an even avatar. To see this, 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. By focusing on the even avatars, and further using symmetry, we arrive at a configuration space where there is just one TBP avatar.

**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 $\xi$ satisfying the following conditions.

1. $\xi$ 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 $p_{01} = 0$.
5. $p_j \in [-3/2, 3/2]^2$ for $j = 1, 2, 3$.
6. $\min(p_{1k}, p_{2k}, p_{3k}) \leq 0$ for $k = 1, 2$.

We define $\Omega^\flat$ to be the same domain except that we leave off Condition 6.

**The Definite Neighborhood:** We specially treat avatars very near the TBP. When we string out the points of $\xi_0$, we get $(1,\ 0, -u, -1, 0, 0, u)$ where $u = \sqrt{3}/3$. The space indicates that we do not record $p_{02} = 0$. We let $\Omega_0$ denote the cube of side-length $2^{-17}$ centered at $\xi_0$.

**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, $\Upsilon$ contains the avatars that compete with the TBP near the exponent $\mathfrak{w}$.

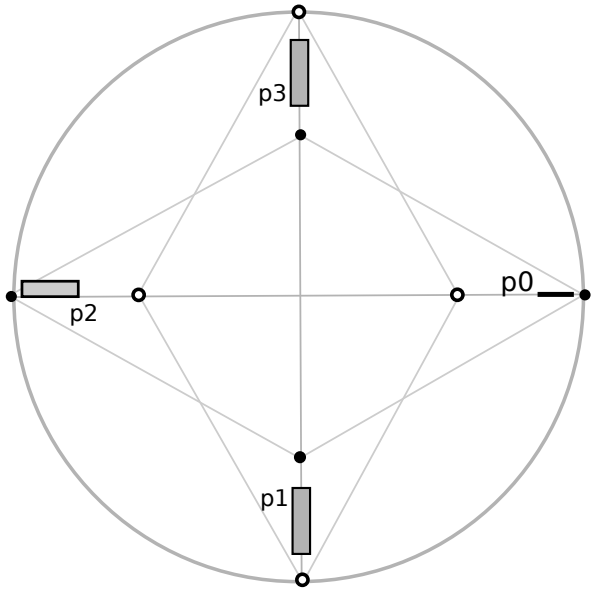


**Figure 2:** The sets defining $\Upsilon$ compared with two TBP avatars.

**The Special Potentials:** Rather than work with the Riesz potentials, we work with potentials that have a more polynomial flavor.

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

Also define

$$G_5^{\flat} = G_5 - 25G_1, \quad G_{10}^{\sharp\sharp} = G_{10} + 28G_5 + 102G_2, \quad G_{10}^{\sharp} = G_{10} + 13G_5 + 68G_2$$

Here is our first result. This one has a pretty short proof.

**Theorem 1.2 (Containment)** *The following is true:*

1. *Let $F = G_4, G_6, G_{10}^{\sharp}$. If $\xi$ is not isomorphic to any avatar in $\Omega$ then then $\xi$ does not minimize $\mathcal{E}_F$.*

2. *Let $F = G_5^{\flat}$. If $\xi$ is not isomorphic to any avatar in $\Omega^{\flat}$ then then $\xi$ does not minimize $\mathcal{E}_F$.*

Here is the main result of this paper. We state it in a conditional way.

**Theorem 1.3 (Calculation)** *Assume the truth of Lemma E, described in §3. Then 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 $\Upsilon$.*

We give the proof of Lemma E in a separate paper.

Combining these results we get the following corollary.

**Corollary 1.4** *Assume the truth of Lemma E. Then the following true*

1. *The TBP is the unique minimizer for $G_4, G_5^{\flat}, G_6, G_{10}^{\sharp\sharp}$ amongst configurations which are not represented by avatars in $\Omega_0$.*

2. *The TBP is the unique minimizer for $G_{10}^{\sharp}$ among 5-point configurations which have are not represented by avatars in $\Upsilon \cup \Omega_0$.*

**Proof:** The only non-obvious point 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 $\Omega - \Omega_0$. ♠

## 1.4 How This Fits In

In Paper 4 we prove the following result.

**Theorem 1.5 (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 $\mathcal{E}_{R_s}(T_0) < \mathcal{E}_{R_s}(T)$.*
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 $\mathcal{E}_{R_s}(T_0) < \mathcal{E}_{R_s}(T)$.*

The Interpolation Theorem and Corollary 1.4 combine to prove the following result.

**Theorem 1.6** *Assume the truth of Lemma E. Let $T_0$ be the TBP. Then*

1. *Suppose $s \in (0, 13]$ and $T$ is any 5-point configuration not represented by an avatar in $\Omega_0$. Then $\mathcal{E}_{R_s}(T_0) < \mathcal{E}_{R_s}(T)$.*
2. *Suppose $s \in [13, 15^+]$ and $T$ is any 5-point configuration not represented by an avatar in $\Omega_0 \cup \Upsilon$. Then $\mathcal{E}_{R_s}(T_0) < \mathcal{E}_{R_s}(T)$.*

To be clear, what we mean in these results is that when a configuration is not represented by an avatar in a certain set, no isomorphic configuration is so represented.

## 1.5 Paper Organization

In §2 we prove the Containment Theorem. In §3 we describe Lemma E. In §4 we describe our calculation modulo Lemma E and thereby prove the Calculation Theorem. The end of §3 has the details of our calculation.

The Calculation Theorem is a massive computer-assisted calculation, the main one in the monograph. We will explain carefully how we do the calculation and then give a record of it.

## 2 Proof of the Containment Theorem

Let $\xi_0$ be an avatar of the TBP. Let $[F] = \mathcal{E}_F(\xi_0)$. Since the TBP has 6 bonds of length $\sqrt{2}$, and 3 of length $\sqrt{3}$, and 1 of length $\sqrt{4}$, we have

$$[G_k] = 6 \times 2^k + 3. \tag{8}$$

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

$$[G_4] = 99, \qquad [G_6] = 387, \qquad [G_5^\flat] = -180, \qquad [G_{10}^\sharp] = 10518. \tag{9}$$

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

**Lemma 2.1** *Let $F = G_6, G_5^\flat, G_{10}^\sharp$. If $\|p_0\| > 3/2$ then $\xi$ does not minimize $\mathcal{E}_F$. If $F = G_4$ then $\xi$ does not minimize $\mathcal{E}_F$ provided that either $\|p_0\| > 2$ or $\|p_0\|, \|p_j\| > 3/2$ for some $j = 1, 2, 3$.*

**Proof:** Let $\tau_j$ be the term in $\mathcal{E}_F$ corresponding to the pair $(p_j, p_4)$. 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 $[G_5^*] = 120$. When $\|p_0\| > 3/2$ we check that $\tau_0 > 450, 123, 26909$, which respectively exceeds $[G_6], [G_5^*], [G_{10}^\sharp]$. (We check this by computing that the distance involved is at most $d_0 = 4/\sqrt{13}$ and that $F$ is monotone decreasing on $[0, d_0]$. Then we evaluate $F(d_0)$ in each case.) Now we treat the case $F = G_4$. When $\|p_0\| > 2$ we have $\tau_0 > 104 > [G_4]$. When $\|p_0\|, \|p_i\| > 3/2$ we have $\tau_0 + \tau_j > 58 + 58 > [G_4]$. ♠

**Lemma 2.2** *If $\min(p_{1k}, p_{2k}, p_{3k}) > 0$ and $F$ is strictly monotone decreasing, then $\xi$ does not minimize $\mathcal{E}_F$.*

**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 $\partial H$ then we increase at least 2 of the distances in the configuration and keep the rest the same. ♠

Assume $\xi$ is a minimizer for $\mathcal{E}_F$. As we have already discussed in the definition of even and odd avatars, we normalize so that $\xi$ 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 2.1 tells us that, in all cases, $p_{01} \in [0, 2]$ and $p_j \in [-3/2, 3/2]^2$ for $j = 1, 2, 3$. We have also arranged that $p_{02} = 0$. For $F = G_5^\flat$ we have nothing left to check. Otherwise, Lemma 2.2 shows that $\xi$ satisfies $\min(p_{1k}, p_{2k}, p_{3k}) \leq 0$ for $k = 1, 2, 3$.

# 3 Statement of Lemma E

## 3.1 Preliminary Definitions

**Energy Hybrids:** Recall that $G_k(r) = (4 - r^2)^k$. We say that an *energy hybrid* is a potential of the form

$$F = \sum_{k=1}^{m} c_k G_k, \qquad G_k(r) = (4 - r^2)^k, \qquad c_1 \in \boldsymbol{Q}, \qquad c_2, ..., c_k \in \boldsymbol{Q}_+. \quad (10)$$

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

$$\square_r := [0, r] \times [-r, r]^r \times [-r, r]^r \times [-r, r]^2. \quad (11)$$

**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

$$B = Q_0 \times Q_1 \times Q_2 \times Q_3 \subset \square_{3/2}, \quad (12)$$

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. If $B$ is not acceptable we let the *offending index* be the lowest index where the condition fails.

The $k$th subdivision of a block amounts to performing dyadic subdivision to the $k$th factor and leaving the others alone. We call these operations $S_0, S_1, S_2, S_3$. Thus $S_0$ cuts $B$ into two pieces and each other $S_k$ cuts $B$ into 4 pieces for $k = 1, 2, 3$. We let $S_k(B)$ denote the list of the blocks obtained by performing $S_k$ on $B$. All the blocks our algorithm produces come from iterated subdivision of $\square_2$. (We toss out those which do not lie in $\square_{3/2}$.)

## 3.2 The Main Calculation

We only work with acceptable blocks. 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:** Let $\Sigma^{-1}$ be inverse stereographic projection, as in Equation 5. Given $Q \in \mathcal{Q}$ we define

$$Q^\bullet = \text{Convex Hull}(\Sigma^{-1}(v(Q)). \tag{13}$$

$Q^\bullet$ 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

$$\widehat{Q} = \Sigma^{-1}(Q). \tag{14}$$

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

$$\chi(D, d) = \frac{d^2}{4D} + \frac{(d^2)^2}{4D^3}. \tag{15}$$

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

$$\delta(Q) = \chi(2, \|\widehat{q}_1 - \widehat{q}_2\|). \tag{16}$$

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

$$\delta(Q) = \max(s_0, s_2) + \max(s_1, s_3), \qquad s_j = \chi(1, \|q_j - q_{j+1}\|). \tag{17}$$

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

$$Q_1 \cdot Q_2 = \max_{i,j}(\widehat{q}_{1i} \cdot \widehat{q}_{2j}) + (\tau) \times (\delta_1 + \delta_2 + \delta_1\delta_2). \tag{18}$$

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

$$T(Q_1, Q_2) = 2 + 2(Q_1 \cdot Q_2). \tag{19}$$

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

$$\epsilon_k(Q_1, Q_2) = \frac{1}{2}k(k-1)T^{k-2}d_1^2 + 2kT^{k-1}\delta_1, \tag{20}$$

where

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

One of the terms in the error estimate 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 a block $Q_0 \times Q_1 \times Q_2 \times Q_3$ we define

$$\mathbf{ERR}_k(B) = \sum_{i=0}^{N} \mathbf{ERR}_k(B, i), \qquad \mathbf{ERR}_k(B, i) = \sum_{j \neq i} \epsilon(Q_i, Q_j). \tag{21}$$

More generally, when $F = \sum c_k G_k$ is as in Equation 10, we define

$$\mathbf{ERR}_F(B) = \sum_{k=0}^{N} \mathbf{ERR}_F(B, i), \qquad \mathbf{ERR}_F(B, i) = \sum |c_k|\ \mathbf{ERR}_k(B, i) \tag{22}$$

Here is the main result.

**Lemma 3.1 (E)** *Let $B$ be a acceptable block. Let $F = G_k$ for any $k \geq 1$ or $F = -G_1$. Then*

$$\min_{p \in B} \mathcal{E}_F(v) \geq \min_{p \in v(B)} \mathcal{E}_k(v) - \mathbf{ERR}_k(B)$$

# 4 Proof of the Calculation Theorem

## 4.1 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 $\Omega_0$ and $\Upsilon$ 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

$$I_0 = [-2^{-17}, 2^{-17}], \quad I_1 = [1-2^{-17}, 1+2^{-17}] \quad 2^{30} I_{\sqrt{3}/3} = [619916940, 619933323] \tag{23}$$

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

$$\Omega_{00} = (I_1 \times \{0\}) \times (I_0 \times -I_{\sqrt{3}/3}) \times (-I_1 \times I_0) \times (I_0 \times I_{\sqrt{3}/3}). \tag{24}$$

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 $\Omega_{00}$. 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 $B$ is disjoint from the interior of $\Omega$. The same goes for $\Omega^\flat$.

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 $k$th factor. Thus $Q_{j1}$ and $Q_{j2}$ are both line segments in $\boldsymbol{R}$.

We return **yes** for any 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 $Q_{22} \le 0$.
- $Q_{j1} \ge 0$ for $j=1,2,3$ or $Q_{j2} \ge 0$ for $j=1,2,3$.

If any of these things happens, all avatars in $Q$ violate some condition for membership in the interior of $\Omega$. We don't check the last item for $\Omega^\flat$.

**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 $\Upsilon$.

For $C_3^\sharp$ we return **yes** if all the vertices of $B$ lie in $\Upsilon$. For $C_3^{\sharp\sharp}$ we return **yes** if one of the factors of $B$ is disjoint from the corresponding factor of $\Upsilon$. This amounts to checking whether a pair of rational squares in the plane are disjoint. We do this using the projections defined for Lemma A132.

**Ingredient 4:** For any function $F$ given by Equation 10, 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 Lemma E, described in the previous chapter.

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

$$\min_{p \in v(B)} \mathcal{E}_F(v) - \mathbf{ERR}_k(B) \ge [F] + 2^{-50} \tag{25}$$

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. ♠

### 4.2 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 $C_1(B) =$ **yes** or $C_2(B) =$ **yes** we remove $B$ from $L$ and go to Step 2. Here we are eliminating blocks disjoint from the interior of $\Omega$ or else contained in $\Omega_0$.
5. If $F = G_{10}^{\sharp}$ and $C_3^{\sharp}(B) =$ **yes** we remove $B$ from $L$ and go to Step 2. If $F = G_{10}^{\sharp\sharp}$ and $C_3^{\sharp\sharp}(B) =$ **yes** we remove $B$ from $L$ and go to Step 2.
6. If $C_{4,F}(B) =$ **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 $\Omega$ or $\Omega^{\flat}$. 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.

### 4.3 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 $S$. Running the algorithm in parallel over sets which partition $\{0, ..., 7\}$ is logically the same as running the basic algorithm without any parallelization. To be able to do the big calculations in pieces, we run the program for various subsets of $\{0, ..., j\}$, sometimes in parallel.

### 4.4 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. I used a 2017 iMac Pro with a 3.2 GHz Intel Zeon W processor, running the Mojave operating system. I ran the programs using Java 8 Update 201. (The Java version I use is not the latest one. The graphical parts of my program use some methods in the Applet class in a very minor but somehow essential way that I find hard to eliminate.) 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 $\Upsilon$): 28 minutes, 805242 blocks.

# 5 References

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

[**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

[**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.

[**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)

See Paper 0 for an extended bibliography.