Five Point Energy Minimization 1: Energy Lemma

Richard Evan Schwartz

April 12, 2024

Abstract

This is Paper 1 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 presents the general energy estimate we prove, which allows us to estimate the energy of a continuum of configurations from a finite calculation.

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 $\boldsymbol{\mathfrak{w}} \in (15, 15_+)$ such that:*

1. *For $s \in (0, \boldsymbol{\mathfrak{w}})$ the TBP is the unique minimizer for R_s.*
2. *For $s = \boldsymbol{\mathfrak{w}}$ the TBP and some FP are the two minimizers for R_s.*
3. *For each $s \in (\boldsymbol{\mathfrak{w}}, 15_+)$ 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 deals with the energy bound which underpins this divide-and-conquer approach.

1.3 The Result

We first introduce the concepts we need in order to state out main result.

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:

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

Σ^{-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 $\mathcal{E}_F(p_0, p_1, p_2, p_3)$ when we mean the F-potential of the corresponding 5-point configuration.

Energy Hybrids: We introduce the energy potential

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

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

$$F = \sum_{k=1}^{m} c_k G_k, \qquad G_k(r) = (4 - r^2)^k, \qquad c_1 \in \boldsymbol{Q}, \qquad c_2, ..., c_k \in \boldsymbol{Q}_+. \tag{8}$$

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

$$\square_r := [0, r] \times [-r, r]^r \times [-r, r]^r \times [-r, r]^2. \tag{9}$$

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

$$B = Q_0 \times Q_1 \times Q_2 \times Q_3 \subset \square_{3/2}, \tag{10}$$

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 $\boldsymbol{R}^2$.

Rational Block Calculations: 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. The integer tells us how to subdivide our block.

Let ξ_0 denote the avatar of the TBP. Here is the main result of this paper.

Theorem 1.2 (Energy) *For any function F given by Equation 8, there exists a rational block computation C_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 $C_F(B)$ is an integer in $\{0,1,2,3\}$.

The point of the Energy Theorem is not to say that there is an abstract calculation which does the job. What we prove is a concrete result that we implement in practice. In the next chapter we will state precisely what the calculation is. Experimentally, I can see that the Energy Theorem is pretty sharp. It also has the advantage of consisting entirely of rational functions of the vertices. In principle, all our calculations could therefore be done with exact integer arithmetic. However, such a calculation would produce enormous integers. The inverval arithmetic calculations we do are a way of controlling the growth of the integers involved in the calculations.

The calculation we give for the Energy Theorem is specially tailored for the cube $\square_{3/2}$. We could get a similar result for a calculation in any $\square_r$, but the constants involved in the calculation would depend on r. The bigger the value of r, the worse the estimate. This is one reason why I worked hard in the monograph to prove that the only minimizers for the potential functions of interest like in $\square_{3/2}$. See the Avatar Theorem below.

1.4 How this Result Fits In

With G_k as above, define the energy hybrids

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

In Paper 4, we prove the following result.

Theorem 1.3 (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 $\mathcal{E}_{R_s}(T_0) < \mathcal{E}_{R_s}(T)$.*

Paper 2 contains the following corollary.

Corollary 1.4 *Let F be any of the functions in the Interpolation Theorem. Then $\zeta \in \boldsymbol{R} \times (\boldsymbol{R}^2)^3$ is an avatar for an F-minimizer only if $\zeta \in \square_{3/2}$.*

The Interpolation Theorem allows us to make calculations with certain of the energy hybrids rather than the Riesz potentials. The corollary implies that we can confine all our calculations to the domain $\square$, and this is exactly the domain on which the Energy Theorem applies. The Energy Theorem will allow us to show, with a finite calculation, that most of $\square$ has higher F-energy than the avatar ξ_0 representing the TBP.

1.5 Paper Organization

In §2 we explain the calculation we do for the Energy Theorem. In §3-4 we prove that it has the desired properties. The proofs in this paper are not computer-assisted. Everything involves traditional mathematics.

2 Description of the Energy Computation

2.1 The Main Calculation

We only work with acceptable blocks. If our calculation encounters an unacceptable block we return the lowest index for which the condition is violated. Henceforth we assume that we have an acceptable block.

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 Σ^{-1} be inverse stereographic projection, as in Equation 5. Given $Q \in \mathcal{Q}$ we define

$$Q^{\bullet} = \text{Convex Hull}(\Sigma^{-1}(v(Q)). \tag{11}$$

$Q^{\bullet}$ 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

$$\widehat{Q} = \Sigma^{-1}(Q). \tag{12}$$

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

$$\chi(D,d) = \frac{d^2}{4D} + \frac{(d^2)^2}{4D^3}. \tag{13}$$

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

$$\delta(Q) = \chi(2, \|\widehat{q}_1 - \widehat{q}_2\|). \tag{14}$$

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

$$\delta(Q) = \max(s_0, s_2) + \max(s_1, s_3), \qquad s_j = \chi(1, \|q_j - q_{j+1}\|). \tag{15}$$

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

$$Q_1 \cdot Q_2 = \max_{i,j}(\widehat{q}_{1i} \cdot \widehat{q}_{2j}) + (\tau) \times (\delta_1 + \delta_2 + \delta_1 \delta_2). \tag{16}$$

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

$$T(Q_1, Q_2) = 2 + 2(Q_1 \cdot Q_2). \tag{17}$$

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

$$\epsilon_k(Q_1, Q_2) = \frac{1}{2}k(k-1)T^{k-2}d_1^2 + 2kT^{k-1}\delta_1, \tag{18}$$

where

$$d_1 = d_{\bullet}(Q_1), \qquad \delta_1 = \delta(Q_1), \qquad T = T(Q_1, Q_2).$$

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

$$\mathbf{ERR}_k(B) = \sum_{i=0}^{N} \mathbf{ERR}_k(B, i), \qquad \mathbf{ERR}_k(B, i) = \sum_{j \neq i} \epsilon(Q_i, Q_j). \tag{19}$$

More generally, when $F = \sum c_k G_k$ is as in Equation 8, we define

$$\mathbf{ERR}_F(B) = \sum_{k=0}^{N} \mathbf{ERR}_F(B, i), \qquad \mathbf{ERR}_F(B, i) = \sum |c_k| \; \mathbf{ERR}_k(B, i) \tag{20}$$

2.2 The Main Result

Now we state the main error estimate, proved in subsequent chapters. For the most part we only care about the (+) case of the lemma. We only need the (−) case when we deal with the potential $G_5 - 25G_1$.

Lemma 2.1 (E) *Let B be a acceptable block. Let $F = G_k$ for any $k \geq 1$ or $F = -G_1$. Then*

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

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

$$\min_{p \in v(B)} \mathcal{E}_F(v) - \mathbf{ERR}_k(B) \geq [F] + 2^{-50} \tag{21}$$

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.

Remark Lemma E is true more generally for $F = \pm G_k$ but we do not need the general result and so (in the interest of simplicity) we ignore it.

2.3 Logic Tree

The rest of the chapter is devoted to proving Lemma E. This lemma has the same name as in the monograph. Here is a logic tree for the proof.

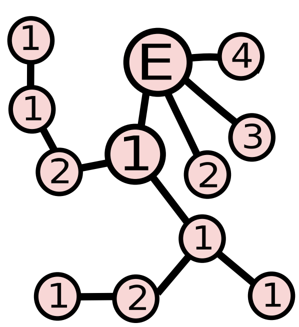

Figure 2: The logic tree for the proof.

This tree tells us that Lemma E is a consequence of Lemmas E1 – E4. Lemma E1 is a consequence of Lemma E11 and Lemma E12. Lemma E12 is a consequence of Lemma E121. And so on.

3 Proof of Lemma E

3.1 Guide to the Proof

Our proof of Lemma E 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 part, done in the next chapter, 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 Lemma E under the assumption such such an averaging system exists. In the next chapter we will prove the existence of the desired averaging system.

3.2 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 18. 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 3.1 (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

$$\sum_{i=1}^{4} \lambda_i(z_1) F(\|\widehat{q}_{1,i} - \widehat{z}_2\|) - F(\|\widehat{z}_1 - \widehat{z}_2\|) \leq \epsilon(Q_1, Q_2). \tag{22}$$

See §4 for the proof.

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:

$$\mathcal{E}(I) = \mathcal{E}(q_{0,i_0}, ..., q_{N,i_N}) \tag{23}$$

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

$$\lambda_I(z) = \prod_{i=0}^{N} \lambda_{i_j}(z_j). \tag{24}$$

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 Lemma E1.

Lemma 3.2 (E2) *Let $z = (z_0, ..., z_N) \in B$. Then*

$$\sum_{I} \lambda_I(z) \mathcal{E}(I) - \mathcal{E}(z) \leq \sum_{i=0}^{N} \sum_{j=0}^{N} \epsilon(Q_i, Q_j). \tag{25}$$

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 Lemma E from Lemma E2. Notice that

$$\sum_{I} \lambda_I(z) = \prod_{j=0}^{N} \left(\sum_{a=1}^{4} \lambda_a(z_j) \right) = 1. \tag{26}$$

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

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

$$\sum_I \lambda_I(z)\mathcal{E}(I) - \mathcal{E}(z) \leq \sum_{i=0}^N \sum_{j=0}^N \epsilon(Q_i, Q_j). \tag{27}$$

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 $Q_4 = \infty$.

3.3 From Local to Global

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

Lemma 3.3 (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

$$F(\|z_0 - z_1\|) \geq \sum_{\alpha=1}^4 \lambda_\alpha(z_0) F(\|q_{0\alpha} - z_1\|) - \epsilon_{01}. \tag{28}$$

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

$$F(\|z_1 - q_{0\alpha}\|) \geq \sum_{\beta=1}^4 \lambda_\beta(z_1) F(\|q_{1\beta} - q_{0\alpha}\|) - \epsilon_{10}. \tag{29}$$

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

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

$$\sum_{\alpha,\beta} \lambda_\alpha(z_0)\lambda_\beta(z_1) F(\|q_{1\beta} - q_{0\alpha}\|) - (\epsilon_{10} + \epsilon_{01}). \tag{30}$$

Equation 30 is equivalent to Equation 25 when $N = 1$. ♠

Now we do the general case.

Lemma 3.4 (E22) *Lemma E2 holds when $N \geq 2$.*

Proof: We rewrite Equation 30 as follows:

$$F(\|z_0-z_1\|) \geq \sum_A \lambda_{A_0}(z_0)\lambda_{A_1}(z_1)\ F(\|q_{0A_0}-q_{1A_1}\|)-(\epsilon_{01}+\epsilon_{10}). \qquad (31)$$

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

We also observe that

$$\sum_{I'} \lambda_{I'}(z')=1, \qquad z'=(z_2,...,z_N). \qquad (32)$$

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

$$\lambda_{A_0}(z_0)\lambda_{A_1}(z_1)=\sum_{I''}\lambda_{I''}(z). \qquad (33)$$

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

$$F(\|z_0-z_1\|) \geq \sum_I \lambda_I(z)F(\|q_{0I_0}-q_{1I_1}\|)-(\epsilon_{01}+\epsilon_{10}). \qquad (34)$$

The same argument works for other pairs of indices, giving

$$F(\|z_i-z_j\|) \geq \sum_I \lambda_I(z)F(\|q_{iI_i}-q_{jI_j}\|)-(\epsilon_{ij}+\epsilon_{ji}). \qquad (35)$$

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 25. When we sum Z_{ij} over all $i<j$ we get the third term in Equation 25. When we sum X_{ij} over all $i<j$ we get

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

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

This is the first term in Equation 25. This proves Lemma E2. ♠

4 Proof of Lemma E1

4.1 The Efficient Averaging System

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 $z^\bullet \in Q^\bullet$

$$z^\bullet = \sum_{i=1}^{4} \lambda_i(z^\bullet)\widehat{q}_i. \tag{36}$$

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

$$\lambda_j(z) = \lambda_j(z^\bullet), \tag{37}$$

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 we have just defined.

4.2 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 4.1 (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*

$$\sum_{i=1}^{4} \lambda_i(z_1) F(\|\widehat{q}_{1,i} - \widehat{z}_2\|) - F(\|\widehat{z}_1 - \widehat{z}_2\|) \leq \frac{1}{2}k(k-1)T^{k-2}d_1^2 + 2kT^{k-1}\delta_1. \tag{38}$$

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

$$T = 2 + 2(Q_1 \cdot Q_1), \qquad Q_1 \cdot Q_2 = \max_{i,j}(\widehat{q}_{1,i} \cdot \widehat{q}_{2,j}) + (\tau) \times (\delta_1 + \delta_2 + \delta_1\delta_2). \tag{39}$$

$\tau = 0$ 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

$$X_\bullet = F(z_1^\bullet - \widehat{z}_2) = (2 + 2z_1^\bullet \cdot \widehat{z}_2)^k \quad \text{or} \quad -2 - 2z_1^\bullet \cdot \widehat{z}_2. \tag{40}$$

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

Lemma 4.2 (E11) $\sum_{i=1}^4 \lambda_i(z_1) F(\|\widehat{q}_{1,i} - \widehat{z}_2\|) - X_\bullet \leq \frac{1}{2}k(k-1)T_\bullet^{k-2} d_1^2$.

Lemma 4.3 (E12) $X_\bullet - F(\|\widehat{z}_1 - \widehat{z}_2\|) \leq 2kT^{k-1}\delta$.

4.3 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 37 and the (bi) linearity of the dot product.

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

Lemma 4.4 (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 $\gamma_1 \cdot \gamma_2 \leq Q_1 \cdot Q_2$.*

Lemma 4.5 (E112) *Let $M \geq 2$ and $k = 1, 2, 3....$ Suppose*

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

Then

$$0 \leq \sum_{i=1}^M \lambda_i x_i^k - \left(\sum_{I=1}^M \lambda_i x_i\right)^k \leq \frac{1}{8}k(k-1)x_M^{k-2}\ (x_M - x_1)^2. \tag{41}$$

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

$$x_i = 4 - \|\widehat{q}_{1,i} - \widehat{z}_2\|^2 = 2 + 2\widehat{q}_{1,i} \cdot \widehat{z}_2, \qquad i = 1, 2, 3, 4. \tag{42}$$

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

$$\sum_{i=1}^{4} \lambda_i(z) F(\|q_{1,i} - z_2\|) = \sum_{i=1}^{4} \lambda_i x_i^k, \tag{43}$$

$$X_\bullet = (2 + 2z_1^\bullet \cdot \widehat{z}_2)^k = \left(\sum_{i=1}^{4} \lambda_i \times (2 + \widehat{q}_i \cdot \widehat{z}_2)\right)^k = \left(\sum_{i=1}^{4} \lambda_i x_i\right)^k. \tag{44}$$

By Equation 43, Equation 44, and the case $M = 4$ of Lemma E112, we have

$$\sum_{i=1}^{4} \lambda_i(z) F(\|q_{1,i} - z_2\|) - X_\bullet = \sum_{i=1}^{4} \lambda_i x_i^k - \left(\sum_{i=1}^{4} \lambda_i x_i\right)^k \leq \frac{1}{8} k(k-1) x_4^{k-2} (x_4 - x_1)^2. \tag{45}$$

By Lemma E111

$$x_4 = 2 + 2(\widehat{q}_4 \cdot \widehat{z}_2) \leq T. \tag{46}$$

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

$$x_4 - x_1 = 2\widehat{z}_2 \cdot (\widehat{q}_4 - \widehat{q}_1) \leq 2\|\widehat{q}_4 - \widehat{q}_1\| \leq 2d_1 \tag{47}$$

Plugging Equations 46 and 47 into Equation 45, we get Lemma E12.

4.4 Proof of Lemma E12

Let $\delta(Q)$ be the hull approximation constant for $Q \in \mathcal{Q}$, as defined (depending on Q) in Equation 14 or Equation 15.

Lemma 4.6 (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

$$f(t) = \left(2 + 2\widehat{z}_2 \cdot \gamma_1(t)\right)^k \quad \text{or} \quad -2 - 2\widehat{z}_2 \cdot \gamma_1(t), \tag{48}$$

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

$$X_\bullet - F(\|\widehat{z}_1 - \widehat{z}_2\|) = f(0) - f(L), \qquad L \leq \delta_1. \tag{49}$$

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

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

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

In short

$$|f'(t)| \leq 2kT^{k-1}. \tag{50}$$

Lemma E13 follows Equation 50, Equation 49, and integration.

4.5 Proof of Lemma E111

See Equation 39 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,

$$q_1^\bullet \cdot q_2^\bullet \leq \max_{i,j}(\widehat{q}_{1i} \cdot \widehat{q}_{2j}). \tag{51}$$

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

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

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

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

This combines with Equation 51 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.

4.6 Proof of Lemma E112

Lemma 4.7 (E1121) *Suppose $a, x \in [0,1]$ and $k \geq 2$. Then $f(x) \leq g(x)$, where*

$$f(x) = (ax^k + 1 - a) - (ax + 1 - a)^k; \qquad g(x) = \frac{1}{8}k(k-1)(1-x)^2. \tag{52}$$

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

$$\frac{f(x)}{g(x)} = \frac{f'(y)}{g'(y)} = \frac{f''(z)}{g''(z)} = 4az^{k-2}\left[1 - a\left(a + \frac{1-a}{z}\right)^{k-2}\right] \leq 4a(1-a) \leq 1. \tag{53}$$

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.

4.7 Proof of Lemma E121

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

$$\chi(D,d) = \frac{d^2}{4D} + \frac{d^4}{4D^3}, \qquad \chi^*(D,d) = \frac{1}{2}(D - \sqrt{D^2 - d^2}). \tag{54}$$

Lemma 4.8 (E1211) *$\chi^*(D,d) \leq \chi(D,d)$ 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 $|y| \leq 3/2$. 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.

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.