Convergence of the restricted Nelder–Mead algorithm in two dimensions

Jeffrey C. Lagarias*
Bjorn Poonen†
Margaret H. Wright‡

April 2, 2011

Abstract

The Nelder–Mead algorithm, a longstanding direct search method for unconstrained optimization published in 1965, is designed to minimize a scalar-valued function f of n real variables using only function values, without any derivative information. Each Nelder–Mead iteration is associated with a nondegenerate simplex defined by $n+1$ vertices and their function values; a typical iteration produces a new simplex by replacing the worst vertex by a new point. Despite the method's widespread use, theoretical results have been limited: for strictly convex objective functions of one variable with bounded level sets, the algorithm always converges to the minimizer; for such functions of two variables, the diameter of the simplex converges to zero, but examples constructed by McKinnon show that the algorithm may converge to a nonminimizing point.

This paper considers the *restricted* Nelder–Mead algorithm, a variant that does not allow expansion steps. In two dimensions we show that, for any nondegenerate starting simplex and any twice-continuously differentiable function with positive definite Hessian and bounded level sets, the algorithm always converges to the minimizer. The proof is based on treating the method as a discrete dynamical system, and relies on several techniques that are non-standard in convergence proofs for unconstrained optimization.

1 Introduction

Since the mid-1980s, interest has steadily grown in *derivative-free* methods (also called *non-derivative* methods) for solving optimization problems, unconstrained and constrained. Derivative-free methods that adaptively construct a local model of relevant nonlinear functions are often described as "model-based", and derivative-free methods that do not explicitly involve such a model tend to be called "direct search" methods. See [5] for a recent survey of derivative-free methods; discussions focusing on direct search methods include, for example, [31, 12, 16, 14, 22].

The Nelder–Mead (NM) simplex method [20] is a direct search method. Each iteration of the NM method begins with a nondegenerate simplex (a geometric figure in n dimensions of nonzero volume that is the convex hull of $n+1$ vertices), defined by its vertices and the associated values

*Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109 (lagarias@umich.edu). The work of this author is partially supported by National Science Foundation grant DMS-0801029. The author also received support through the Mathematics Research Center at Stanford University.

†Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 (poonen@math.mit.edu). The work of this author is supported by National Science Foundation grant DMS-0841321.

‡Computer Science Department, New York University, New York, New York 10012 (mhw@cs.nyu.edu). The work of this author is partially supported by Department of Energy grant DE-FG02-88ER25053.

of f. One or more trial points are computed, along with their function values, and the iteration produces a new (different) simplex such that the function values at its vertices typically satisfy a descent condition compared to the previous simplex.

The NM method is appealingly simple to describe (see Figure 2), and has been widely used (along with numerous variants) for more than 45 years, in many scientific and engineering applications. But little mathematical analysis of any kind of the method's performance has appeared, with a few exceptions such as [30, 10] (from more than 20 years ago) and (more recently) [9]. As we discuss in more detail below, obtaining even limited convergence proofs for the original method has turned out to be far from simple. The shortage of theory, plus the discovery of low-dimensional counterexamples (see (1.1)) have made the NM method an outlier among modern direct search methods, which are deliberately based on a rigorous mathematical foundation. (See, for example, [6, 2, 14, 1], as well as more recent publications about direct search methods for constrained problems.) Nevertheless the NM method retains importance because of its continued use and availability in computer packages (see [23, 17, 7]) and its apparent usefulness in some situations.

In an effort to develop positive theory about the original NM algorithm, an analysis of its convergence behavior was initiated in [15] in 1998, along with resolution of ambiguities in [20] about whether function comparisons involve "greater than" or "greater than or equal" tests.[1] In what follows we use the term *Nelder-Mead algorithm* to refer generically to one of the precisely specified procedures in [15]; these contain a number of adjustable parameters (coefficients), and the *standard coefficients* represent an often-used choice. For strictly convex objective functions with bounded level sets, [15] showed convergence of the most general form of the NM algorithm to the minimizer in one dimension. For the NM algorithm with standard coefficients in dimension two, where the simplex is a triangle, it was shown that the function values at the simplex vertices converge to a limiting value, and furthermore that the diameter of the simplices converges to zero. But it was not shown that the simplices always converge to a limiting point, and up to now this question remains unresolved.

Taking the opposite perspective, McKinnon [18] devised a family of two-dimensional counterexamples consisting of strictly convex functions with bounded level sets and a specified initial simplex, for which the NM simplices converge to a nonminimizing point. In the smoothest McKinnon example, the objective function is

$$f_m(x,y) = \begin{cases} 2400|x|^3 + y + y^2 & \text{if } x \le 0 \\ 6x^3 + y + y^2 & \text{if } x \ge 0, \end{cases} \tag{1.1}$$

when the vertices of the starting simplex are $(0,0)$, $(1,1)$ and $((1+\sqrt{33})/8,\ (1-\sqrt{33})/8))$. Note that f_m is twice-continuously differentiable and that its Hessian is positive definite except at the origin, where it is singular. As shown in Figure 1, the NM algorithm converges to the origin (one of the initial vertices) rather than to the minimizer $(0,-\frac{1}{2})$, performing an infinite sequence of inside contractions (see Section 2) in which the best vertex of the initial triangle is never replaced.

Functions proposed by various authors on which the NM algorithm fails to converge to a minimizer are surveyed in [18], but counterexamples in the McKinnon family illustrated by (1.1) constitute the "nicest" functions for which the NM algorithm converges to a non-stationary point.

An algorithmic flaw that has been observed is that the iterations "stagnate" or "stall", often because the simplex becomes increasingly close to degenerate (as depicted in Figure 1). Previously proposed corrective strategies include: placing more restrictions on moves that decrease the size of the simplex; imposing a "sufficient decrease" condition (stronger than simple decrease) for accepting a new vertex; and resetting the simplex to one that is "nice". See, for example, [25, 30, 29, 11, 24, 19, 5], a small selection of the many papers that include convergence results for modifications of Nelder–Mead.

Our object in this paper is to fill in additional theory for the NM algorithm in the two-dimensional case, which remains of interest in its own right. As noted by McKinnon [18, page 148], it is not even known whether the NM algorithm converges for the prototypically nice function $f(x,y) = x^2 + y^2$.

[1] Resolution of these ambiguities can have a noticeable effect on the performance of the algorithm; see [8].

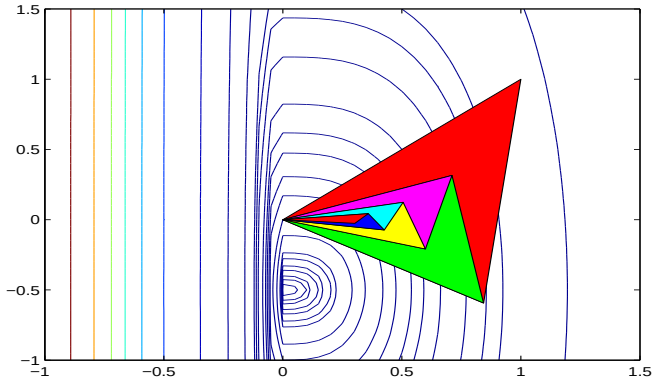

Figure 1: The NM algorithm's failure on the McKinnon counterexample (1.1).

Here we answer this question affirmatively for a simplified variant of the NM algorithm, where the simplification reduces the number of allowable moves rather than attempting to "fix" the method. In the original NM algorithm (see Section 2), the allowable moves are reflection, expansion, outside contraction, inside contraction, and shrink; an expansion doubles the volume of an NM simplex, while all other moves either leave the volume the same or decrease it. An expansion is tried only after the reflection point produces a strict improvement in the best value of f; the motivation is to allow a longer step along an apparently promising direction. The *restricted* Nelder–Mead (RNM) algorithm defined in Section 2 does not allow expansion steps. Thus we are in effect considering a "small step" NM algorithm.

Our analysis applies to the following class of functions:

Definition 1.1. Let $\mathcal{F}$ denote the class of twice-continuously differentiable functions $f\colon \mathbb{R}^2 \to \mathbb{R}$ with bounded level sets and everywhere positive definite Hessian.

The class $\mathcal{F}$ is a subclass of those considered in [15], where there is no requirement of differentiability.

The contribution of this paper is to prove convergence of the restricted Nelder-Mead algorithm for functions in $\mathcal{F}$:

Theorem 1.2. (appears again as Theorem 3.17) *If the RNM algorithm is applied to a function $f \in \mathcal{F}$, starting from any nondegenerate triangle, then the algorithm converges to the unique minimizer of f.*

Remark 1.3. Theorem 1.2 immediately implies a generalization to a larger class of functions. Namely, if $f \in \mathcal{F}$, and $g\colon \mathbb{R} \to \mathbb{R}$ is a strictly increasing function, then the RNM algorithm applied to $\tilde{f} := g \circ f$ converges, because the RNM steps for $\tilde{f}$ are identical to those for f.

Remark 1.4. Because the NM iterations in the McKinnon examples include no expansion steps, the RNM algorithm also will fail to converge to a minimizer on these examples. It follows that, in order to obtain a positive convergence result, additional assumptions on the function over those in [15] must be imposed. In particular, the positive-definiteness condition on the Hessian in Theorem 1.2 rules out the smoothest McKinnon example (1.1), in which the Hessian is singular at the origin (the nonminimizing initial vertex to which the NM algorithm converges).

An interesting general property of the Nelder–Mead algorithm is the constantly changing shape of the simplex as the algorithm progresses. Understanding the varying geometry of the simplex seems crucial to explaining how the algorithm behaves. Our proof of Theorem 1.2 analyzes the RNM algorithm as a discrete dynamical system, in which the shapes of the relevant simplices (with a proper scaling) form a phase-space for the algorithm's behavior. The imposed hypothesis on the

Hessian, which is stronger than strict convexity, allows a crucial connection to be made between a (rescaled) local geometry and the vertex function values. We analyze the algorithm's behavior in a transformed coordinate system that corrects for this rescaling.

The proof of Theorem 1.2 establishes convergence by contradiction, by showing that the algorithm can find no way not to converge. We make, in effect, a "Sherlock Holmes" argument: Once you have eliminated the impossible, whatever remains, however improbable, must be the truth.[2] We show that, in order not to converge to the minimizer, the triangles would need to flatten out according to a particular geometric scaling, but there is no set of RNM steps permitting this flattening to happen. This result is confirmed through an auxiliary potential function measuring the deviation from scaling. One can almost say that the RNM algorithm converges in spite of itself.

2 The restricted Nelder–Mead algorithm

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a function to be minimized, and let $\mathbf{p}_1, \ldots, \mathbf{p}_{n+1}$ be the vertices of a nondegenerate simplex in $\mathbb{R}^n$. One iteration of the RNM algorithm (with standard coefficients) replaces the simplex by a new one according to the following procedure.

One iteration of the standard RNM algorithm.

1. **Order.** Order and label the $n+1$ vertices to satisfy $f(\mathbf{p}_1) \le f(\mathbf{p}_2) \le \cdots \le f(\mathbf{p}_{n+1})$, using appropriate tie-breaking rules such as those in [15].

2. **Reflect.** Calculate $\bar{\mathbf{p}} = \sum_{i=1}^{n} \mathbf{p}_i/n$, the average of the n best points (omitting $\mathbf{p}_{n+1}$). Compute the *reflection point* $\mathbf{p}_\mathrm{r}$, defined as $\mathbf{p}_\mathrm{r} = 2\bar{\mathbf{p}} - \mathbf{p}_{n+1}$, and evaluate $f_\mathrm{r} = f(\mathbf{p}_\mathrm{r})$. If $f_\mathrm{r} < f_n$, accept the reflected point $\mathbf{p}_\mathrm{r}$ and terminate the iteration.

3. **Contract.** If $f_\mathrm{r} \ge f_n$, perform a *contraction* between $\bar{\mathbf{p}}$ and the better of $\mathbf{p}_{n+1}$ and $\mathbf{p}_\mathrm{r}$.

   a. Outside contract. If $f_n \le f_\mathrm{r} < f_{n+1}$ (i.e., $\mathbf{p}_\mathrm{r}$ is strictly better than $\mathbf{p}_{n+1}$), perform an *outside contraction*: calculate the outside contraction point $\mathbf{p}_\mathrm{out} = \frac{1}{2}(\bar{\mathbf{p}} + \mathbf{p}_\mathrm{r})$, and evaluate $f_\mathrm{out} = f(\mathbf{p}_\mathrm{out})$. If $f_\mathrm{out} \le f_\mathrm{r}$, accept $\mathbf{p}_\mathrm{out}$ and terminate the iteration; otherwise, go to Step 4 (perform a shrink).

   b. Inside contract. If $f_\mathrm{r} \ge f_{n+1}$, perform an *inside contraction*: calculate the inside contraction point $\mathbf{p}_\mathrm{in} = \frac{1}{2}(\bar{\mathbf{p}} + \mathbf{p}_{n+1})$, and evaluate $f_\mathrm{in} = f(\mathbf{p}_\mathrm{in})$. If $f_\mathrm{in} < f_{n+1}$, accept $\mathbf{p}_\mathrm{in}$ and terminate the iteration; otherwise, go to Step 4 (perform a shrink).

4. **Perform a shrink step.** Evaluate f at the n points $\boldsymbol{v}_i = \frac{1}{2}(\mathbf{p}_1 + \mathbf{p}_i)$, $i = 2, \ldots, n+1$. The (unordered) vertices of the simplex at the next iteration consist of $\mathbf{p}_1$, $\boldsymbol{v}_2$, $\ldots$, $\boldsymbol{v}_{n+1}$.

The result of an RNM iteration is either: (1) a single new vertex—the *accepted point*—that replaces the worst vertex $\mathbf{p}_{n+1}$ in the set of vertices for the next iteration; or (2) if a shrink is performed, a set of n new points that, together with $\mathbf{p}_1$, form the simplex at the next iteration.

Starting from a given nondegenerate simplex, let $\mathbf{p}_1^{(k)}$, $\ldots$, $\mathbf{p}_{n+1}^{(k)}$ be the vertices at the *start* of the k^{th} iteration. Let $\mathbf{z} \in \mathbb{R}^n$ be a point. We say that the RNM algorithm converges to $\mathbf{z}$ if $\lim_{k\to\infty} \mathbf{p}_i^{(k)} = \mathbf{z}$ for every $i \in \{1, \ldots, n+1\}$.

Remark 2.1. In two dimensions, a reflect step performs a 180° rotation of the triangle around $\bar{\mathbf{p}}$, so the resulting triangle is congruent to the original one. But in higher dimensions, the reflected simplex is not congruent to the original.

Remark 2.2. Shrink steps are irrelevant in this paper because we are concerned only with strictly convex objective functions, for which shrinks cannot occur (Lemma 3.5 of [15]). It follows that, at each NM iteration, the function value at the new vertex is strictly less than the worst function value at the previous iteration.

[2] A. Conan Doyle, "The Sign of the Four", *Lippincott's Monthly Magazine*, February 1890.

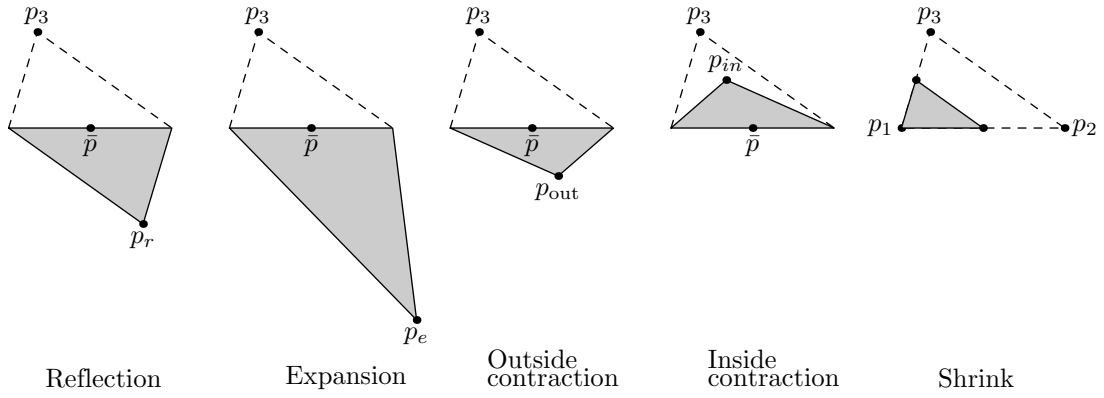

Figure 2: The five possible moves in the original NM algorithm are shown. The original simplex is surrounded by a dashed line, and its worst vertex is labeled $\mathbf{p}_3$. The point $\bar{\mathbf{p}}$ is the average of the two best vertices. The shaded figures are NM simplices following reflection, expansion, outside contraction, inside contraction, and shrink, respectively. (In the "shrink" figure, the best vertex is labeled $\mathbf{p}_1$.) The "expansion" step is omitted in the RNM algorithm.

Remark 2.3. The original Nelder–Mead algorithm differs from the above in Step 2. Namely, if $\mathbf{p}_\mathrm{r}$ is better than all $n+1$ of the vertices, the original NM algorithm tries evaluating f at the *expansion* point $\mathbf{p}_\mathrm{e} := \bar{\mathbf{p}} + \chi(\bar{\mathbf{p}} - \mathbf{p}_{n+1})$ for a fixed expansion coefficient $\chi > 1$, and the worst vertex $\mathbf{p}_{n+1}$ is then replaced by the better of $\mathbf{p}_\mathrm{e}$ and $\mathbf{p}_\mathrm{r}$. In fact, Nelder and Mead proposed a family of algorithms, depending on coefficients for reflection, contraction, and shrinkage in addition to expansion. A complete, precise definition of an NM iteration is given in [15], along with a set of tie-breaking rules. Instances of the moves in the original NM algorithm are shown in Figure 2.

Remark 2.4. One feature of the RNM algorithm that makes it easier to analyze than the original algorithm is that the volume of the simplex is non-increasing at each step. The volume thus serves as a Lyapunov function.[3]

We henceforth consider the RNM algorithm *in dimension two*, for which it is known that the simplex diameter converges to zero.

Lemma 2.5. *Suppose that the RNM algorithm is applied to a strictly convex 2-variable function with bounded level sets. Then for any nondegenerate initial simplex, the diameters of the RNM simplices (triangles) produced by the algorithm converge to* 0.

Proof. The proof given in [15, Lemma 5.2] for the original NM algorithm applies even when expansion steps are disabled. □

3 Convergence

3.1 The big picture

Because the logic of the convergence proof is complicated, we begin with an overview of the argument. Each $f \in \mathcal{F}$ is strictly convex, so by Lemma 2.5 we know that the evolution of any triangle under the RNM algorithm has the diameter of the triangle converging to zero. (We do not yet know that the triangles converge to a limit point.) The convergence proof proceeds by contradiction, making an initial hypothesis (Hypothesis 1 in Section 3.4) that the (unique) minimizer of f is not a limit point of the RNM triangles. Under this condition, all three RNM vertices must approach a level set

[3] See Definition 1.3.4 in [27, page 23].

corresponding to a function value strictly higher than the optimal value. By our assumptions on f, this level set is a strictly convex closed curve with a continuously differentiable tangent vector.

The RNM triangle must become small as it approaches this bounding level set. Therefore, from the viewpoint of the triangle, blown up to have (say) unit diameter, the level set flattens out to a straight line. The heuristic underlying our argument is that, in order for this to happen, the triangle must itself have its shape flatten out, with its width in the level set direction (nearly horizontal, as seen from the triangle) being roughly the square root of its height in the perpendicular direction. In particular, its width becomes proportionally much larger than its height. A local coordinate frame (Section 3.3) is defined in order to describe this phenomenon.

At the *start* of iteration k, we measure area and width in a local coordinate frame, and define a quantity called "flatness" by $\Gamma_k := \text{area}_k/\text{width}_k^3$. If a reflection is taken during iteration k and the same coordinate frame is retained, the area and width of the RNM triangle at iteration $k+1$ remain the same, so $\Gamma_{k+1} = \Gamma_k$. Hence, in order for the diameter to converge to zero (Lemma 2.5), there must be infinitely many contraction steps. We show that, at a sufficiently advanced iteration k of the RNM algorithm, a necessary condition for a contraction to occur is that $\Gamma_k \leq 10$; we also show that the value of Γ eventually unavoidably increases as the algorithm proceeds. A contradiction thus arises because no combination of the permitted reflection and contraction steps allows the needed square root rate of decrease.

The argument is complicated because the local coordinate frame changes at every step. Near the end of the proof (in Proposition 3.15), we analyze sequences of no more than 14 steps, beginning with a contraction, in an advanced phase of the algorithm. Using a coordinate frame defined by a vertex of the first triangle in the sequence, we show that switching to a new coordinate system defined via the final triangle in the sequence makes only a small change in the flatness. This allows us to show that the flatness is inflated by a factor of at least 1.01 after at most 14 steps, which eventually means that a contraction cannot be taken. Since the triangle cannot reflect forever, our contradiction hypothesis must have been false; i.e., the method must converge.

3.2 Notation

Points in two dimensions are denoted by boldface lower-case letters, but a generic point is often called $\boldsymbol{p}$, which is treated as a column vector and written as $\boldsymbol{p} = (x, y)^T$. We shall also often use an affinely transformed coordinate system with generic point denoted by $\widetilde{\boldsymbol{p}} = (\widetilde{x}, \widetilde{y})^T$. To stress the (x, y) coordinates of a specific point, say $\boldsymbol{b}$, we write $\boldsymbol{b} = (b_x, b_y)^T$.

For future reference, we explicitly give the formulas for the reflection and contraction points in two dimensions:

$$\boldsymbol{p}_{\text{r}} = \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{p}_3 \qquad \text{(2-d reflection);} \tag{3.1}$$

$$\boldsymbol{p}_{\text{out}} = \tfrac{3}{4}(\boldsymbol{p}_1 + \boldsymbol{p}_2) - \tfrac{1}{2}\boldsymbol{p}_3 \qquad \text{(2-d outside contraction);} \tag{3.2}$$

$$\boldsymbol{p}_{\text{in}} = \tfrac{1}{4}(\boldsymbol{p}_1 + \boldsymbol{p}_2) + \tfrac{1}{2}\boldsymbol{p}_3 \qquad \text{(2-d inside contraction).} \tag{3.3}$$

Given the three vertices of a triangle, the reflection and contraction points depend only on which (one) vertex is labeled as "worst".

3.3 A changing local coordinate system

The type of move at each RNM iteration is governed by a discrete decision, based on comparing values of f. Heuristically, for a very small triangle near a point $\boldsymbol{b}$, the result of the comparison is usually unchanged if we replace f by its degree-2 Taylor polynomial centered at $\boldsymbol{b}$. If $\boldsymbol{b}$ is a nonminimizing point, then we can simplify the function further by making an affine transformation into a new coordinate system $\widetilde{\boldsymbol{p}} = (\widetilde{x}, \widetilde{y})$ (depending on $\boldsymbol{b}$) in which the Taylor polynomial has the form

$$\text{constant} + \widetilde{y} + \tfrac{1}{2}\widetilde{x}^2.$$

This motivates the following lemma, which is a version of Taylor's theorem.

Lemma 3.1. (Definition of local coordinate frame.) *Let $f \in \mathcal{F}$. Given a point $\boldsymbol{b}$ and a nonsingular 2×2 matrix M, we may define an affine transformation*

$$\widetilde{\boldsymbol{p}} = M^{-1}(\boldsymbol{p} - \boldsymbol{b}) \tag{3.4}$$

(with inverse map $\boldsymbol{p} = M\widetilde{\boldsymbol{p}} + \boldsymbol{b}$).

(i) *For each point $\boldsymbol{b}$ that is not the minimizer of f, there exists a unique M with $\det M > 0$ such that when the function f of $\boldsymbol{p} = (x, y)^T$ is re-expressed in the new coordinate system $\widetilde{\boldsymbol{p}} = (\widetilde{x}, \widetilde{y})^T$ above, the result has the form*

$$f(\boldsymbol{p}) = f(\boldsymbol{b}) + \widetilde{y} + \tfrac{1}{2}\widetilde{x}^2 + r(\widetilde{x}, \widetilde{y}), \tag{3.5}$$

where r is an error term satisfying

$$r(\widetilde{x}, \widetilde{y}) = \tfrac{1}{2}\alpha\widetilde{y}^2 + o(\max(|\widetilde{x}|^2, |\widetilde{y}|^2)), \tag{3.6}$$

as $(\widetilde{x}, \widetilde{y}) \to \mathbf{0}$ (i.e., as $(x, y)^T \to \boldsymbol{b}$), for some $\alpha > 0$.

(ii) *The function r in* (i) *satisfies $dr/d\widetilde{x} = o(\max(|\widetilde{x}|, |\widetilde{y}|))$ and $dr/d\widetilde{y} = o(|\widetilde{x}|) + O(|\widetilde{y}|)$, and the rate at which the $o(\cdot)$ terms approach zero and the bounds implied by $O(\cdot)$ can be made uniform for $\boldsymbol{b}$ in any compact set not containing the minimizer of f.*

(iii) *As $\boldsymbol{b}$ varies over a compact set not containing the minimizer of f, the matrices M and M^{-1} are bounded in norm and uniformly continuous.*

Proof. Let $\boldsymbol{g} = \nabla f(\boldsymbol{b})$ and $H = \nabla^2 f(\boldsymbol{b})$ denote, respectively, the gradient and Hessian matrix of f at $\boldsymbol{b}$. Since f is strictly convex, its gradient can vanish only at the unique minimizer, so $\boldsymbol{g} \neq 0$. Because f is twice-continuously differentiable, we can expand it in Taylor series around $\boldsymbol{b}$:

$$f(\boldsymbol{p}) = f(\boldsymbol{b}) + \boldsymbol{g}^T(\boldsymbol{p} - \boldsymbol{b}) + \tfrac{1}{2}(\boldsymbol{p} - \boldsymbol{b})^T H(\boldsymbol{p} - \boldsymbol{b}) + o(\|\boldsymbol{p} - \boldsymbol{b}\|^2) \tag{3.7}$$

$$= f(\boldsymbol{b}) + \boldsymbol{g}^T M\widetilde{\boldsymbol{p}} + \tfrac{1}{2}\widetilde{\boldsymbol{p}}^T M^T H M\widetilde{\boldsymbol{p}} + o(\|\boldsymbol{p} - \boldsymbol{b}\|^2). \tag{3.8}$$

The Taylor expansion (3.8) has the desired form if

$$\boldsymbol{g}^T M = (0 \;\; 1) \quad \text{and} \quad M^T H M = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}.$$

for some $\alpha > 0$. In terms of the columns $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ of M, these conditions say

$$\boldsymbol{g}^T\boldsymbol{m}_1 = 0, \quad \boldsymbol{g}^T\boldsymbol{m}_2 = 1, \quad \boldsymbol{m}_1^T H\boldsymbol{m}_1 = 1, \quad \boldsymbol{m}_1^T H\boldsymbol{m}_2 = 0,$$

and then we may set $\alpha := \boldsymbol{m}_2^T H\boldsymbol{m}_2$, which will be positive since H is positive definite and since the conditions above force $\boldsymbol{m}_2$ to be nonzero.

Since $\boldsymbol{g} \neq 0$, the condition $\boldsymbol{g}^T\boldsymbol{m}_1 = 0$ says that $\boldsymbol{m}_1$ is a multiple of the vector $\widehat{\boldsymbol{g}}$ obtained by rotating $\boldsymbol{g}$ by $90°$ clockwise: $\boldsymbol{m}_1 = \xi_1\widehat{\boldsymbol{g}}$ for some ξ_1. The condition $\boldsymbol{m}_1^T H\boldsymbol{m}_1 = 1$ implies that $\boldsymbol{m}_1 \neq \mathbf{0}$. The condition $\boldsymbol{m}_1^T H\boldsymbol{m}_2 = 0$ says that $H\boldsymbol{m}_2$ is a multiple of $\boldsymbol{g}$. Since H is positive definite, H is nonsingular, so the equation $H\mathbf{w} = \boldsymbol{g}$ has the unique solution $\mathbf{w} = H^{-1}\boldsymbol{g}$, and then $\boldsymbol{m}_2 = \xi_2\mathbf{w}$ for some ξ_2. The normalizations $\boldsymbol{g}^T\boldsymbol{m}_2 = 1$ and $\boldsymbol{m}_1^T H\boldsymbol{m}_1 = 1$ are equivalent to

$$\xi_2 = \frac{1}{\boldsymbol{g}^T\mathbf{w}} = \frac{1}{\mathbf{w}^T H\mathbf{w}} \quad \text{and} \quad \xi_1^2 = \frac{1}{\widehat{\boldsymbol{g}}^T H\widehat{\boldsymbol{g}}}; \tag{3.9}$$

the denominators are positive since H is positive definite and $\mathbf{w}$ and $\widehat{\boldsymbol{g}}$ are nonzero. These conditions determine M uniquely up to the choice of sign of its first column, i.e., the sign of ξ_1, but we have not yet imposed the condition $\det M > 0$. We claim that it is the positive choice of ξ_1 that makes $\det M > 0$: since $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are then positive multiples of $\widehat{\boldsymbol{g}}$ and $\mathbf{w}$, respectively, the condition $\det M > 0$ is equivalent to $\boldsymbol{g}^T\mathbf{w} > 0$, or equivalently, $\mathbf{w}^T H^T\mathbf{w} > 0$, which is true since the matrix $H^T = H$ is positive definite. This proves (i).

Since f is twice-continuously differentiable, $\boldsymbol{g}$ and H vary continuously as $\boldsymbol{b}$ varies within a compact set not containing the minimizer of f. Hence M and M^{-1} vary continuously as well. This proves (ii) and (iii). ∎

Remark 3.2. If H is positive semidefinite and singular, then the equation $H\mathbf{w} = \boldsymbol{g}$ continues to have a solution provided that $\boldsymbol{g} \in \text{range}(H)$, as in the McKinnon example (1.1). But in this case, $H\widehat{\boldsymbol{g}} = 0$, so $H\boldsymbol{m}_1 = 0$, which contradicts $\boldsymbol{m}_1^T H \boldsymbol{m}_1 = 1$, and no matrix M exists.

Remark 3.3. As $\boldsymbol{b}$ approaches the minimizer of f, we have $\boldsymbol{g} \to \mathbf{0}$, and the formulas obtained in the proof of Lemma 3.1 show that $\boldsymbol{m}_1$ remains bounded while $\boldsymbol{m}_2$ and the value of α "blow up", so M becomes unbounded in norm with an increasing condition number.

The local coordinate frame defined in Lemma 3.1 depends on the base point $\boldsymbol{b}$, the gradient vector $\boldsymbol{g}$, and the Hessian matrix H. In the rest of this section, we use $\mathfrak{F}(\boldsymbol{b})$ (with a nonminimizing point $\boldsymbol{b}$ as argument) to denote the local coordinate frame with base point $\boldsymbol{b}$. In the context of a sequence of RNM iterations, $\mathfrak{F}_k$ (or $\mathfrak{F}(\Delta_k)$, with a subscripted RNM triangle as argument) will mean the coordinate frame defined with a specified base point in RNM triangle Δ_k.

3.3.1 Width, height, area, and flatness.

This section collects some results about transformed RNM triangles.

Definition 3.4. (Width, height, and flatness.) Let $f \in \mathcal{F}$, and let Δ denote a nondegenerate triangle that lies in a compact set $\mathcal{Q}$ not containing the minimizer of f. Assume that we are given a base point $\boldsymbol{b}$ in $\mathcal{Q}$, along with the coordinate frame defined at $\boldsymbol{b}$ as in Lemma 3.1.

- The (transformed) *width* of Δ, denoted by $\widetilde{w}(\Delta)$, is the maximum absolute value of the difference in $\widetilde{x}$-coordinates of two vertices of Δ;
- The (transformed) *height*, denoted by $\widetilde{h}(\Delta)$, is the maximum absolute value of the difference of $\widetilde{y}$-coordinates of two vertices of Δ;
- The *flatness* of Δ, denoted by $\Gamma(\Delta)$, is

$$\Gamma(\Delta) := \frac{\widetilde{A}(\Delta)}{\widetilde{w}(\Delta)^3}, \tag{3.10}$$

  where $\widetilde{A}(\Delta)$ is the (positive) area of Δ measured in the transformed coordinates.

The argument Δ may be omitted when it is obvious.

Lemma 3.5. (Effects of a reflection) *The (transformed) height and width of an RNM triangle are the same as those of its reflection, if the same base point is used to define the local coordinate frame for both triangles.*

Proof. The new triangle is a 180° rotation of the old triangle. □

The next lemma bounds the change in three quantities arising from small changes in the base point used for the local coordinate frames. In (iii), we need a hypothesis on the width and height since for a tall thin triangle, a slight rotation can affect its flatness dramatically.

Lemma 3.6. (Consequences of close base points.) *Assume that $f \in \mathcal{F}$ and that $\mathcal{Q}$ is a compact set that does not contain the minimizer of f. Let $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ denote two points in $\mathcal{Q}$, and Δ denote an RNM triangle contained in $\mathcal{Q}$. For $i \in \{1, 2\}$, let $\widetilde{w}_i$, $\widetilde{h}_i$, and Γ_i be the transformed width, height, and flatness of Δ measured in the local coordinate frame $\mathfrak{F}(\boldsymbol{b}_i)$ associated with $\boldsymbol{b}_i$, and let M_i be the matrix of Lemma 3.1 associated with $\mathfrak{F}(\boldsymbol{b}_i)$.*

(i) *Given $\epsilon > 0$, there exists $\delta > 0$ (independent of $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$) such that if $\|\boldsymbol{b}_2 - \boldsymbol{b}_1\| < \delta$, then*

$$\|M_2 M_1^{-1} - I\| < \epsilon.$$

(ii) *Given $\epsilon > 0$, there exists $\delta > 0$ (independent of $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and Δ) such that if $\|\boldsymbol{b}_1 - \boldsymbol{b}_2\| < \delta$, then*

$$(1 - \epsilon)\widetilde{A}_1 < \widetilde{A}_2 < (1 + \epsilon)\widetilde{A}_1. \tag{3.11}$$

(iii) *Given $C, \epsilon > 0$, there is $\delta > 0$ (independent of $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and Δ) such that if $\|\boldsymbol{b}_1 - \boldsymbol{b}_2\| < \delta$ and $\widetilde{w}_1 > C\widetilde{h}_1$, then*

$$(1-\epsilon)\Gamma_1 < \Gamma_2 < (1+\epsilon)\Gamma_1. \tag{3.12}$$

Proof.

(i) We have

$$\|M_2 M_1^{-1} - I\| = \|M_2(M_1^{-1} - M_2^{-1})\| \le \|M_2\|\,\|M_1^{-1} - M_2^{-1}\|.$$

By Lemma 3.1(iii), the first factor $\|M_2\|$ is uniformly bounded, and M^{-1} is uniformly continuous as a function of $\boldsymbol{b} \in \mathcal{Q}$, so the second factor $\|M_1^{-1} - M_2^{-1}\|$ can be made as small as desired by requiring $\|\boldsymbol{b}_2 - \boldsymbol{b}_1\|$ to be small.

(ii) Letting $\widetilde{\boldsymbol{p}}_2$ and $\widetilde{\boldsymbol{p}}_1$ denote the transformed versions of a point $\boldsymbol{p}$ in $\mathcal{Q}$ using $\mathfrak{F}(\boldsymbol{b}_1)$ and $\mathfrak{F}(\boldsymbol{b}_2)$, we have

$$\widetilde{\boldsymbol{p}}_2 = M_2^{-1} M_1 \widetilde{\boldsymbol{p}}_1 + M_2^{-1}(\boldsymbol{b}_1 - \boldsymbol{b}_2), \tag{3.13}$$

so that $\widetilde{\boldsymbol{p}}_2$ and $\widetilde{\boldsymbol{p}}_1$ are related by an affine transformation with matrix $M_2^{-1}M_1$. When an affine transformation with nonsingular matrix B is applied to the vertices of a triangle, the area of the transformed triangle is equal to the area of the original triangle multiplied by $|\det(B)|$ [13, page 144]. Applying this result to Δ gives

$$\widetilde{A}_2 = \widetilde{A}_1\,|\det(M_2^{-1}M_1)|. \tag{3.14}$$

Since $|\det B|$ is a continuous function of B, the result follows from (i).

(iii) Because of (ii), it suffices to prove the analogous inequalities for width instead of flatness. Fixing two vertices of Δ, we let $\mathbf{v}_i$ denote the vector from one to the other measured in $\mathfrak{F}(\boldsymbol{b}_i)$, and let $x(\mathbf{v}_i)$ denote the corresponding x-component. Then $|\mathbf{v}_i| \le \widetilde{w}_1 + \widetilde{h}_1 = O(\widetilde{w}_1)$, since $\widetilde{w}_1 > C\widetilde{h}_1$. By (3.13), $\mathbf{v}_2 = M_2^{-1}M_1\mathbf{v}_1$, so

$$|x(\mathbf{v}_2) - x(\mathbf{v}_1)| \le |\mathbf{v}_2 - \mathbf{v}_1| = |(M_2^{-1}M_1 - I)\mathbf{v}_1| = O(\|M_2^{-1}M_1 - I\| \cdot |\widetilde{w}_1|).$$

This bounds the change in x-component of each vector of the triangle in passing from $\mathfrak{F}(\boldsymbol{b}_1)$ to $\mathfrak{F}(\boldsymbol{b}_2)$, and it follows that

$$|\widetilde{w}_2 - \widetilde{w}_1| = O(\|M_2^{-1}M_1 - I\| \cdot |\widetilde{w}_1|).$$

Finally, by (i), $\|M_2^{-1}M_1 - I\|$ can be made arbitrarily small. □

3.4 The contradiction hypothesis and the limiting level set

Our proof of Theorem 1.2 is by contradiction. Therefore we assume the following hypothesis for the rest of Section 3 and hope to obtain a contradiction.

Hypothesis 1. *Assume that the RNM algorithm is applied to $f \in \mathcal{F}$ and a nondegenerate initial triangle, and that it does not converge to the minimizer of f.*

We begin with a few easy consequences of Hypothesis 1. Let Δ_k be the RNM triangle at the start of the k^{th} iteration. Let $\widetilde{\Delta}_k$ be that triangle in the coordinate frame determined by any one of its vertices, and define its width $\widetilde{w}_k$, height $\widetilde{h}_k$, and flatness Γ_k as in Definition 3.4.

Lemma 3.7. *Assume Hypothesis 1. Then:*

(a) *The diameter of Δ_k tends to 0.*

(b) *The RNM triangles have at least one limit point $\boldsymbol{p}^\dagger$.*

(c) *The function values at the vertices of Δ_k are greater than or equal to $f(\boldsymbol{p}^\dagger)$, and they tend to $f(\boldsymbol{p}^\dagger)$.*

(d) *If $\mathcal{Q}$ is a neighborhood of the level set of $\boldsymbol{p}^\dagger$, then all the action of the algorithm is eventually inside $\mathcal{Q}$.*

(e) *We may choose $\mathcal{Q}$ to be a compact neighborhood not containing the minimizer of f; then there is a positive lower bound on the smallest eigenvalue of the Hessian in $\mathcal{Q}$.*

(f) *The diameter of $\widetilde{\Delta}_k$ tends to zero.*

(g) *We have $\widetilde{w}_k \to 0$ and $\widetilde{h}_k \to 0$.*

Proof.

(a) This follows from Lemma 2.5, even without Hypothesis 1.

(b) Lemma 3.3 of [15] states that the best, next-worst, and worst function values in each successive triangle cannot increase, and that at least one of them must strictly decrease at each iteration. Because level sets are bounded, compactness guarantees that there is a limit point $\boldsymbol{p}^\dagger$.

(c) This follows from the monotonic decrease in function values, the shrinking of the diameter to zero, and the continuity of f.

(d) Since the level sets are compact, there is a compact neighborhood I of $f(\boldsymbol{p}^\dagger)$ such that $f^{-1}(I)$ is a compact set contained in the interior of $\mathcal{Q}$. By (c), the triangles are eventually contained in $f^{-1}(I)$. By (a), eventually even the rejected points tested in each iteration lie within $f^{-1}(I)$.

(e) The first statement follows since the minimizer is not on the level set of $\boldsymbol{p}^\dagger$. The second statement follows from uniform continuity of the Hessian.

(f) By Lemma 3.1(iii), the distortion of the triangles is uniformly bounded.

(g) This follows from (f).

□

For the rest of Section 3, we may assume that all our RNM triangles and test points lie in a compact set $\mathcal{Q}$ not containing the minimizer, as in Lemma 3.7(e). In particular, the implied bounds in Lemma 3.1 are uniform.

3.5 Flattening of the RNM triangles

Under Hypothesis 1, we now show that the transformed RNM triangles "flatten out" in the sense that the height becomes arbitrarily small relative to the width. The proof is again a proof by contradiction, showing that, unless the triangles flatten out, there must be a sequence of consecutive reflections in which the value of f at the reflection point is eventually less than $f(\boldsymbol{p}^\dagger)$, contradicting Lemma 3.7(c).

Lemma 3.8. (Flattening of RNM triangles.) *Assume Hypothesis 1. Then* $\lim_{k\to\infty} \widetilde{h}_k/\widetilde{w}_k = 0$.

Proof. Assume that the result of the lemma does not hold. In other words, within the rest of this proof, the following hypothesis is assumed:

Hypothesis 2. *There exists $\rho > 0$ such that for arbitrarily large k we have $\widetilde{h}_k/\widetilde{w}_k > \rho$.*

We may assume also that $\boldsymbol{p}^\dagger$ is a limit point of the triangles Δ_k for which $\widetilde{h}_k/\widetilde{w}_k > \rho$.

Given $\epsilon > 0$, we define a *downward-pointing sector* of points $(\widetilde{x}, \widetilde{y})$ satisfying $\widetilde{y} \le \epsilon - \rho|\widetilde{x}|/10$, and a *truncated sector* of points in the downward sector that also satisfy $\widetilde{y} \ge -\epsilon$: see Figure 3.

We now show that there exists $\epsilon > 0$ (depending on f and ρ) such that, for any sufficiently advanced iteration k_0 for which $\widetilde{h}_{k_0}/\widetilde{w}_{k_0} > \rho$,

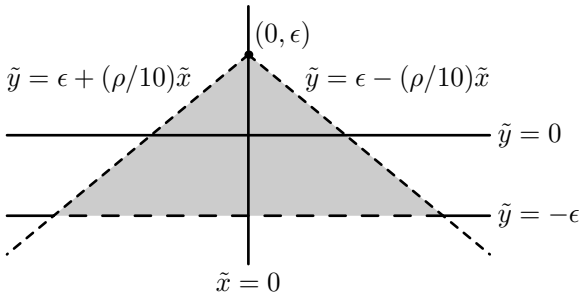

Figure 3: The downward-pointing sector lies between the two finely dashed lines. The truncated sector consists of the shaded area, for $\rho = 8$ and $\epsilon = 0.5$.

(a) $\widetilde{\Delta}_{k_0}$ is contained in the truncated sector.

(b) If $\widetilde{\Delta}$ is *any* RNM triangle in the coordinates $(\widetilde{x}, \widetilde{y})$ of $\mathfrak{F}_{k_0}$ such that $\widetilde{\Delta}$ is contained in the truncated sector and has (transformed) width $\widetilde{w}$ and height $\widetilde{h}$ satisfying $\widetilde{h}/\widetilde{w} > \rho$, then

  (i) One RNM iteration reflects $\widetilde{\Delta}$ to a new triangle $\widetilde{\Delta}'$. (And $\widetilde{\Delta}'$ has the same width and height as $\widetilde{\Delta}$, by Lemma 3.5.)

  (ii) The $\widetilde{y}$-coordinate of the centroid of $\widetilde{\Delta}'$ is at least $88\widetilde{h}/300$ below that of $\widetilde{\Delta}$.

  (iii) $\widetilde{\Delta}'$ is contained in the downward-pointing sector.

  (iv) If $\widetilde{\Delta}'$ is not contained in the truncated sector, then the function value at the new vertex is less than $f(\boldsymbol{p}^\dagger)$.

Starting from (a), applying (b) repeatedly shows that the triangle in the $(\widetilde{x}, \widetilde{y})$ coordinates reflects downward until it exits the truncated sector through the bottom, at which point the function value at the exiting vertex is less than $f(\boldsymbol{p}^\dagger)$, which contradicts Lemma 3.7(c). Thus it remains to prove (a) and (b).

Proof of (a).

By definition of $\mathfrak{F}_{k_0}$, the point $(0, 0)$ is a vertex of $\widetilde{\Delta}_{k_0}$. For any given $\epsilon > 0$, if k_0 is sufficiently large, then Lemma 3.7(f) shows that the diameter of $\widetilde{\Delta}_{k_0}$ is less than the distance from $(0, 0)$ to the boundary of the truncated sector, so $\widetilde{\Delta}_{k_0}$ is entirely contained in the truncated sector.

Proof of (b).

Suppose that $\widetilde{\Delta}$ is contained in the truncated sector and satisfies $\widetilde{h}/\widetilde{w} > \rho$. Its vertices $\widetilde{\boldsymbol{p}}_i = (\widetilde{x}_i, \widetilde{y}_i)$ are the transforms of vertices $\boldsymbol{p}_i$ of some Δ. We will use the notation $f_i = f(\boldsymbol{p}_i)$ for any subscript i, and use similar abbreviations for other functions and coordinates.

We show first that the difference in f values at any two vertices $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ is within $3\widetilde{h}/100$ of the differences of their $\widetilde{y}$-coordinates. Using (3.5), we find that

$$f_i - f_j = \widetilde{y}_i - \widetilde{y}_j + \tfrac{1}{2}(\widetilde{x}_i^2 - \widetilde{x}_j^2) + r_i - r_j. \tag{3.15}$$

The quantity $|\widetilde{x}_i^2 - \widetilde{x}_j^2|$ is bounded by $2\widetilde{w}|\widetilde{x}_i| + \widetilde{w}^2$. If $\epsilon < \rho^2/4000$, then $|\widetilde{x}| < \rho/200$ for any point in the truncated sector. By Lemma (3.7)(f), if k_0 is large enough, then $\widetilde{w} < \rho/100$. It follows that

$$\widetilde{w}|\widetilde{x}| < \frac{\widetilde{w}\rho}{200} < \frac{\widetilde{h}}{200} \qquad \text{and} \qquad \widetilde{w}^2 < \frac{\rho\widetilde{w}}{100} < \frac{\widetilde{h}}{100}, \tag{3.16}$$

so $|\widetilde{x}_i^2 - \widetilde{x}_j^2| < \widetilde{h}/50$. On the other hand, $r_i - r_j$ is the line integral of $(dr/d\widetilde{x}, dr/d\widetilde{y})$ over a path of length at most $\widetilde{w} + \widetilde{h} = O(\widetilde{h})$. Since $dr/d\widetilde{x}$ and $dr/d\widetilde{y}$ are $O(\max(|\widetilde{x}|, |\widetilde{y}|))$, the derivatives can

be made arbitrarily small on the truncated sector by choosing ϵ small enough, and we may assume that $|r_i - r_j| < \widetilde{h}/100$. Now (3.15) yields

$$f_i - f_j = \widetilde{y}_i - \widetilde{y}_j + \zeta, \quad \text{with} \quad |\zeta| < \frac{3\widetilde{h}}{100}. \tag{3.17}$$

(i) Let $\boldsymbol{p}_{\rm best}, \boldsymbol{p}_{\rm next}, \boldsymbol{p}_{\rm worst}$ be the vertices of Δ ordered so that $f_{\rm best} \le f_{\rm next} \le f_{\rm worst}$. Let $\boldsymbol{p}_{\rm low}, \boldsymbol{p}_{\rm mid}, \boldsymbol{p}_{\rm high}$ be the same vertices ordered so that $\widetilde{y}_{\rm low} \le \widetilde{y}_{\rm mid} \le \widetilde{y}_{\rm high}$. Recall that the reflection point $\boldsymbol{p}_{\rm r} = \boldsymbol{p}_{\rm best} + \boldsymbol{p}_{\rm next} - \boldsymbol{p}_{\rm worst}$ is accepted only if $f_{\rm r} < f_{\rm next}$. Equation (3.17) implies that $\widetilde{y}_{\rm best}$, $\widetilde{y}_{\rm next}$, $\widetilde{y}_{\rm worst}$ are within $3\widetilde{h}/100$ of $\widetilde{y}_{\rm low}$, $\widetilde{y}_{\rm mid}$, $\widetilde{y}_{\rm high}$, respectively. Hence the difference

$$\widetilde{y}_{\rm next} - \widetilde{y}_{\rm r} = \widetilde{y}_{\rm worst} - \widetilde{y}_{\rm best}$$

is within $6\widetilde{h}/100$ of $\widetilde{y}_{\rm high} - \widetilde{y}_{\rm low} = \widetilde{h}$. Applying (3.17) to the reflected triangle shows that $f_{\rm next} > f_{\rm r}$, and the reflection point is accepted.

(ii) The reflection decreases the $\widetilde{y}$ coordinate of the reflected vertex by

$$\widetilde{y}_{\rm worst} - \widetilde{y}_{\rm r} = 2\widetilde{y}_{\rm worst} - \widetilde{y}_{\rm best} - \widetilde{y}_{\rm next},$$

which is within $4(3\widetilde{h}/100)$ of

$$2\widetilde{y}_{\rm high} - \widetilde{y}_{\rm low} - \widetilde{y}_{\rm mid} \ge \widetilde{y}_{\rm high} - \widetilde{y}_{\rm low} = \widetilde{h}.$$

Consequently, $\widetilde{y}_{\rm worst} - \widetilde{y}_{\rm r} \ge 88\widetilde{h}/100$, and the centroid drops by at least $88\widetilde{h}/300$.

(iii) Furthermore, $\widetilde{x}_{\rm r}$ differs from $\widetilde{x}_{\rm worst}$ by no more than $2\widetilde{w}$, i.e., $|\widetilde{x}_{\rm r}| \le |\widetilde{x}_{\rm worst}| + 2\widetilde{w}$. Since $\boldsymbol{p}_{\rm worst}$ lies in the truncated sector and $\rho\widetilde{w} < \widetilde{h}$, it follows that

$$\begin{aligned} \widetilde{y}_{\rm r} + \frac{\rho}{10}|\widetilde{x}_{\rm r}| &\le \widetilde{y}_{\rm worst} - \frac{88\widetilde{h}}{100} + \frac{\rho}{10}(|\widetilde{x}_{\rm worst}| + 2\widetilde{w}) < \widetilde{y}_{\rm worst} - \frac{88\widetilde{h}}{100} + \frac{\rho}{10}|\widetilde{x}_{\rm worst}| + \frac{2\widetilde{h}}{10} \\ &< \widetilde{y}_{\rm worst} + \frac{\rho}{10}|\widetilde{x}_{\rm worst}| < \epsilon. \end{aligned}$$

Thus, using the local coordinate frame $\mathfrak{F}_{k_0}$, the reflection point $\boldsymbol{p}_{\rm r}$ lies in the downward-pointing sector, and also lies in the truncated sector as long as $\widetilde{y}_{\rm r} \ge -\epsilon$.

(iv) Let $\boldsymbol{b}$ denote the base point of $\mathfrak{F}_{k_0}$, so $\widetilde{\boldsymbol{b}} = (0,0)$. For $\widetilde{\boldsymbol{p}}$ on the bottom edge of the truncated sector, we have $\widetilde{y} = -\epsilon$ and $\widetilde{x} = O(\epsilon)$ as $\epsilon \to 0$ (similar triangles). Relation (3.5) then implies

$$f(\boldsymbol{p}) = f(\boldsymbol{b}) - \epsilon + O(\epsilon^2). \tag{3.18}$$

Fixing ϵ to be small enough that $f(\boldsymbol{p}) - f(\boldsymbol{b}) < 0$ everywhere on the bottom edge, we can also fix a neighborhood U of the bottom edge and a neighborhood V of $\widetilde{\boldsymbol{b}} = (0,0)$ such that $f(\boldsymbol{p}) < f(\boldsymbol{b}')$ holds whenever $\widetilde{\boldsymbol{p}} \in U$ and $\widetilde{\boldsymbol{b}}' \in V$.

If $\widetilde{\Delta}'$ is not in the truncated sector, its new vertex $\widetilde{\boldsymbol{p}}_{\rm r}$ is within $\widetilde{w} + \widetilde{h}$ of the bottom edge. If k_0 is sufficiently large to make $\widetilde{w} + \widetilde{h}$ small enough, it follows that $\widetilde{\boldsymbol{p}}_{\rm r} \in U$.

By choice of $\boldsymbol{p}^\dagger$ (defined immediately following Hypothesis 2), k_0 can be taken large enough that $\boldsymbol{p}^\dagger$ is arbitrarily close to $\boldsymbol{b}$ in *untransformed* coordinates. By Lemma 3.1(iii), the matrix defining the local coordinate transformation is bounded and nonsingular. Hence we can make $\widetilde{\boldsymbol{p}}^\dagger$ arbitrarily close to $(0,0)$ in transformed coordinates, and in particular we can guarantee that $\widetilde{\boldsymbol{p}}^\dagger$ lies in V.

Thus $f(\boldsymbol{p}_{\rm r}) < f(\boldsymbol{p}^\dagger)$. □

Remark 3.9. An important consequence of Lemma 3.8 is that $\widetilde{w} > \widetilde{h}$ for Δ_k measured in a coordinate frame associated to any one of its vertices, so that Lemma 3.6(iii) can be applied with $C = 1$.

3.6 The distance travelled during a sequence of reflections

We now show that a sequence of valid reflections, starting from a sufficiently advanced iteration, does not move the triangle far. This result limits the possible change in flatness caused by moving the base point of the local coordinate system from the first to last triangle in the series of reflections.

Lemma 3.10. *Assume Hypothesis 1. Given $\kappa > 0$, the following is true for any sufficiently large k_0 and any $k \geq k_0$: if all steps taken by the RNM algorithm from Δ_{k_0} to Δ_k are reflections, then the distance between the transformed centroids of Δ_{k_0} and Δ_k is less than κ (where we use a coordinate frame whose base point is a vertex of Δ_{k_0}).*

Proof. We work in the coordinates $(\widetilde{x}, \widetilde{y})$ of $\mathfrak{F}_{k_0}$. It suffices to show that for sufficiently small positive $\epsilon < \kappa/2$, if k_0 is sufficiently large and $\widetilde{\Delta}$ is a later RNM triangle with centroid in the box $\{|\widetilde{x}| \leq \epsilon, |\widetilde{y}| \leq \epsilon\}$, then the next move does not reflect $\widetilde{\Delta}$ so that its centroid exits the box. More precisely, for suitable ϵ and k_0, the idea is to prove:

(a) The centroid cannot escape out the top of the box (i.e., the $\widetilde{y}$-coodinate cannot increase beyond ϵ) because the function value of the reflection point would exceed the function values of Δ_{k_0} (i.e., the function values near the center of the box).

(b) The centroid cannot escape out the bottom because the function value there would be less than the limiting value $f(\boldsymbol{p}^\dagger)$.

(c) The centroid cannot escape out either side, because the triangle $\widetilde{\Delta}$ will be flat enough that the function values there are controlled mainly by the $\widetilde{x}$-coordinates, which force the triangle to reflect inward towards the line $\widetilde{x} = 0$.

The conditions on ϵ and k_0 will be specified in the course of the proof.

Proof of (a).

We copy the argument used in proving (b)(iv) of Lemma 3.8. Let $\boldsymbol{b}$ be the base point used to define $\mathfrak{F}_{k_0}$. For $\boldsymbol{p}$ along the top edge of the box, by definition $\widetilde{y} = \epsilon$. Thus the same argument that proved (3.18) shows that

$$f(\boldsymbol{p}) = f(\boldsymbol{b}) + \epsilon + O(\epsilon^2),$$

and that if ϵ is sufficiently small, then there are neighborhoods U of the top edge and V of $(0, 0)$ such that $f(\boldsymbol{p}) > f(\boldsymbol{b}')$ holds whenever $\widetilde{\boldsymbol{p}} \in U$ and $\widetilde{\boldsymbol{b}}' \in V$. If k_0 is sufficiently large, and $\widetilde{\Delta}$ is the later triangle whose centroid is about to exit the box through the top, then by Lemma (3.7)(f), $\widetilde{\Delta}_{k_0}$ and $\widetilde{\Delta}$ are small enough that $\widetilde{\Delta}_{k_0} \subset V$ and $\widetilde{\Delta} \subset U$, so the function values at vertices of Δ are greater than those for Δ_{k_0}, which is impossible since function values at vertices of successive RNM triangles are non-increasing.

Proof of (b).

This case is even closer to the proof of (b)(iv) in Lemma 3.8. That argument shows that if ϵ is sufficiently small and k_0 is sufficiently large, then the function values at the vertices of a triangle Δ whose transformed centroid is about to exit through the bottom are strictly less than the value $f(\boldsymbol{p}^\dagger)$ (which is made arbitrarily close to $f(\boldsymbol{b})$ by taking k_0 large). This contradicts Lemma 3.7(c).

Proof of (c).

By symmetry, suppose that $\widetilde{\Delta}$ reflects so that its centroid exits the box through the *right* side. By Lemma 3.7(g) and Lemma 3.8, we may take k_0 large enough that

$$\widetilde{w}_{k_0} < 0.01\epsilon \qquad \text{and} \qquad \widetilde{h}_{k_0} < 0.01\epsilon\widetilde{w}_{k_0}. \tag{3.19}$$

The width $\widetilde{w}$ and height $\widetilde{h}$ of $\widetilde{\Delta}$ are the same as that of $\widetilde{\Delta}_{k_0}$. So all vertices of $\widetilde{\Delta}$ satisfy $0.99\epsilon < \widetilde{x} < 1.01\epsilon$ and $-1.01\epsilon < \widetilde{y} < 1.01\epsilon$. Let $\widetilde{\boldsymbol{v}} = (\widetilde{x}, \widetilde{y})$ and $\widetilde{\boldsymbol{v}}' = (\widetilde{x} + \delta_x, \widetilde{y} + \delta_y)$ be two such vertices.

We claim that if $\delta_x > \widetilde{w}/10$, then $f(\boldsymbol{v}') > f(\boldsymbol{v})$. By (3.5),

$$\begin{aligned} f(\boldsymbol{v}') - f(\boldsymbol{v}) &= \widetilde{x}\delta_x + \tfrac{1}{2}\delta_x^2 + \delta_y + (r(\widetilde{x}+\delta x, \widetilde{y}+\delta y) - r(\widetilde{x},\widetilde{y})) \\ &\geq (0.99\epsilon)(\widetilde{w}/10) + 0 - \widetilde{h} - (o(\epsilon)\widetilde{w} + O(\epsilon)\widetilde{h}) \quad \text{(by integrating Lemma 3.1(ii))} \\ &\geq 0.099\epsilon\widetilde{w} + 0 - \widetilde{h} - 0.001\epsilon\widetilde{w} - \widetilde{h} \quad \text{(if } \epsilon \text{ is sufficiently small)} \\ &= 0.098\epsilon\widetilde{w} - 2\widetilde{h} \\ &> 0 \qquad \text{(by the second inequality in (3.19)).} \end{aligned}$$

Now we can mimic part of the proof of (b) in Lemma 3.8, but in the horizontal rather than the vertical direction. Let $\widetilde{x}_{\text{best}}, \widetilde{x}_{\text{next}}, \widetilde{x}_{\text{worst}}$ be the $\widetilde{x}$-coordinates of the vertices ordered by increasing function value, and let $\widetilde{x}_{\text{left}}, \widetilde{x}_{\text{mid}}, \widetilde{x}_{\text{right}}$ be the same $\widetilde{x}$-coordinates in increasing order. The previous paragraph shows that $\widetilde{x}_{\text{best}}, \widetilde{x}_{\text{next}}, \widetilde{x}_{\text{worst}}$ are within $\widetilde{w}/10$ of $\widetilde{x}_{\text{left}}, \widetilde{x}_{\text{mid}}, \widetilde{x}_{\text{right}}$, respectively. The reflection decreases the $\widetilde{x}$ coordinate of the reflected vertex by

$$\widetilde{x}_{\text{worst}} - \widetilde{x}_{\text{r}} = 2\widetilde{x}_{\text{worst}} - \widetilde{x}_{\text{best}} - \widetilde{x}_{\text{next}},$$

which is within $4(\widetilde{w}/10)$ of

$$2\widetilde{x}_{\text{right}} - \widetilde{x}_{\text{left}} - \widetilde{x}_{\text{mid}} \geq \widetilde{x}_{\text{right}} - \widetilde{x}_{\text{left}} = \widetilde{w},$$

so the $\widetilde{x}$ coordinate of the centroid decreases instead of increasing beyond ϵ as hypothesized. □

3.7 Conditions at an advanced contraction

Assuming Hypothesis 1, we next show that, whenever a contraction step is taken at a sufficiently advanced iteration k, we have $\widetilde{h}_k = O(\widetilde{w}_k^2)$. We stress the assumption that the base of the local coordinate frame at iteration k lies inside Δ_k.

Lemma 3.11. *Assume Hypothesis 1. If k is sufficiently large and a contraction step is taken at iteration k (meaning that the reflection point was not accepted), then the transformed height $\widetilde{h}$ and width $\widetilde{w}$ of Δ_k in a coordinate frame with base point inside Δ_k must satisfy $\widetilde{h} \leq 10\widetilde{w}^2$.*

Proof. Given a base point of the local coordinate frame in Δ_k, Lemma 3.1 shows that the difference in values of f at any two points $\boldsymbol{p}$ and $\mathbf{v}$ is

$$f(\boldsymbol{p}) - f(\mathbf{v}) = \widetilde{y}_{\boldsymbol{p}} - \widetilde{y}_{\mathbf{v}} + \tfrac{1}{2}(\widetilde{x}_{\boldsymbol{p}}^2 - \widetilde{x}_{\mathbf{v}}^2) + r(\widetilde{x}_{\boldsymbol{p}}, \widetilde{y}_{\boldsymbol{p}}) - r(\widetilde{x}_{\mathbf{v}}, \widetilde{y}_{\mathbf{v}}). \tag{3.20}$$

For $i \in \{1,2,3\}$, let $\boldsymbol{p}_i$ be the i^{th} vertex of Δ_k, and let $\widetilde{\boldsymbol{p}}_i$ be its transform in the local coordinate frame. We assume throughout the proof that $\boldsymbol{p}_3$ is the worst vertex. Let $\boldsymbol{p}_{\text{r}} := \boldsymbol{p}_1 + \boldsymbol{p}_2 - \boldsymbol{p}_3$ be the reflect point, and let $\widetilde{\boldsymbol{p}}_{\text{r}}$ be its transform.

The origin of the coordinate frame is inside Δ_k, so $|\widetilde{x}_i| \leq \widetilde{w}$ for $i = 1,2,3$. The RNM triangles are flattening out (Lemma 3.8), and the flatness does not change very much when measured using the coordinate frame with a nearby base point (Lemma 3.6(iii)). Hence, if k is large enough, $\widetilde{h} \leq \widetilde{w}$, so $|\widetilde{y}_i| \leq \widetilde{w}$ for $i = 1,2,3$. Since $\boldsymbol{p}_3$ is the worst vertex, $f(\boldsymbol{p}_3) - f(\boldsymbol{p}_1) \geq 0$. Substituting (3.20) and rearranging yields

$$\widetilde{y}_3 - \widetilde{y}_1 \geq \tfrac{1}{2}(\widetilde{x}_1^2 - \widetilde{x}_3^2) + r(\widetilde{x}_1, \widetilde{y}_1) - r(\widetilde{x}_3, \widetilde{y}_3). \tag{3.21}$$

Because $|\widetilde{x}_i| \leq \widetilde{w}$ and $|\widetilde{x}_j| \leq \widetilde{w}$, we obtain $|\widetilde{x}_i^2 - \widetilde{x}_j^2| \leq \widetilde{w}^2$, so the inequality (3.21) implies

$$\widetilde{y}_3 - \widetilde{y}_1 \geq -\tfrac{1}{2}\widetilde{w}^2 + r(\widetilde{x}_1, \widetilde{y}_1) - r(\widetilde{x}_3, \widetilde{y}_3). \tag{3.22}$$

Next we use the definition of the reflection point to obtain bounds in the other direction. A contraction occurs only when the reflection point is not accepted (see Step 3 of Algorithm RNM in Section 2), which implies that $f(\boldsymbol{p}_{\text{r}}) - f(\boldsymbol{p}_2) \geq 0$. Substituting (3.20) and rearranging yields

$$\widetilde{y}_{\text{r}} - \widetilde{y}_2 \geq \tfrac{1}{2}(\widetilde{x}_2^2 - \widetilde{x}_{\text{r}}^2) + r(\widetilde{x}_2, \widetilde{y}_2) - r(\widetilde{x}_{\text{r}}, \widetilde{y}_{\text{r}}). \tag{3.23}$$

By definition of $\boldsymbol{p}_\mathrm{r}$, we have $\widetilde{y}_\mathrm{r} - \widetilde{y}_2 = \widetilde{y}_1 - \widetilde{y}_3$. Substituting into the left-hand side of (3.23) yields

$$\widetilde{y}_1 - \widetilde{y}_3 \geq \tfrac{1}{2}(\widetilde{x}_2^2 - \widetilde{x}_\mathrm{r}^2) + r(\widetilde{x}_2, \widetilde{y}_2) - r(\widetilde{x}_\mathrm{r}, \widetilde{y}_\mathrm{r}). \tag{3.24}$$

We have $|\widetilde{x}_2| \leq \widetilde{w}$ and $|\widetilde{x}_\mathrm{r} - \widetilde{x}_2| = |\widetilde{x}_1 - \widetilde{x}_3| \leq \widetilde{w}$, so

$$|\widetilde{x}_2^2 - \widetilde{x}_\mathrm{r}^2| = |\widetilde{x}_2 + \widetilde{x}_\mathrm{r}| \cdot |\widetilde{x}_2 - \widetilde{x}_\mathrm{r}| \leq 3\widetilde{w}^2,$$

and substituting into (3.24) yields

$$\widetilde{y}_1 - \widetilde{y}_3 \geq -\tfrac{3}{2}\widetilde{w}^2 + r(\widetilde{x}_2, \widetilde{y}_2) - r(\widetilde{x}_\mathrm{r}, \widetilde{y}_\mathrm{r}). \tag{3.25}$$

If k is sufficiently large, we know from Lemmas 2.5 and 3.1 that, in the smallest box containing a transformed advanced RNM triangle and its reflection point, $|dr/d\widetilde{x}| \leq \widetilde{w}$ and $|dr/d\widetilde{y}| \leq \frac{1}{2}$. Consequently,

$$\begin{aligned} |r(\widetilde{x}_1, \widetilde{y}_1) - r(\widetilde{x}_3, \widetilde{y}_3)| &\leq \widetilde{w}|\widetilde{x}_1 - \widetilde{x}_3| + \tfrac{1}{2}|\widetilde{y}_1 - \widetilde{y}_3| \leq \widetilde{w}^2 + \tfrac{1}{2}|\widetilde{y}_1 - \widetilde{y}_3| \\ |r(\widetilde{x}_2, \widetilde{y}_2) - r(\widetilde{x}_\mathrm{r}, \widetilde{y}_\mathrm{r})| &\leq \widetilde{w}|\widetilde{x}_1 - \widetilde{x}_3| + \tfrac{1}{2}|\widetilde{y}_1 - \widetilde{y}_3| \leq \widetilde{w}^2 + \tfrac{1}{2}|\widetilde{y}_1 - \widetilde{y}_3|. \end{aligned} \tag{3.26}$$

Substituting the equations (3.26) into (3.22) and (3.25), respectively, we obtain

$$\widetilde{y}_3 - \widetilde{y}_1 \geq -\tfrac{3}{2}\widetilde{w}^2 - \tfrac{1}{2}|\widetilde{y}_1 - \widetilde{y}_3| \quad \text{and} \quad \widetilde{y}_1 - \widetilde{y}_3 \geq -\tfrac{5}{2}\widetilde{w}^2 - \tfrac{1}{2}|\widetilde{y}_1 - \widetilde{y}_3|. \tag{3.27}$$

These imply $\widetilde{y}_3 - \widetilde{y}_1 \geq -3\widetilde{w}^2$ and $\widetilde{y}_1 - \widetilde{y}_3 \geq -5\widetilde{w}^2$, so $|\widetilde{y}_1 - \widetilde{y}_3| \leq 5\widetilde{w}^2$. Our numbering of $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ was arbitrary, so $|\widetilde{y}_2 - \widetilde{y}_3| \leq 5\widetilde{w}^2$ too. These two inequalities imply $\widetilde{h} \leq 10\widetilde{w}^2$. □

Remark 3.12. The lemma just proved applies to an RNM triangle not at an arbitrary iteration, but only at a sufficiently advanced iteration k. Even for large k, the condition $\widetilde{h} \leq 10\widetilde{w}^2$ is necessary but not sufficient to characterize an RNM triangle for which a contraction occurs.

Figures 4 and 5 illustrate two cases for the function $\frac{1}{2}\widetilde{x}^2 + \widetilde{y} + \frac{1}{2}\widetilde{y}^2$. The worst vertex is at the origin in each figure. In Figure 4, we have $\widetilde{h} = 1.2 \times 10^{-6}$ and $\widetilde{w} = 2 \times 10^{-4}$, so $\widetilde{h}/\widetilde{w}^2 = 30$; as Lemma 3.11 would predict at an advanced iteration, the triangle reflects instead of contracting. In Figure 5, by contrast, $\widetilde{h} = 3 \times 10^{-8}$ and $\widetilde{w} = 2 \times 10^{-4}$, so $\widetilde{h}/\widetilde{w}^2 = \frac{3}{4}$ and an outside contraction is taken. The vertical scale in each figure is greatly compressed compared to the horizontal, and the vertical scale in Figure 4 differs from that in Figure 5 by two orders of magnitude.

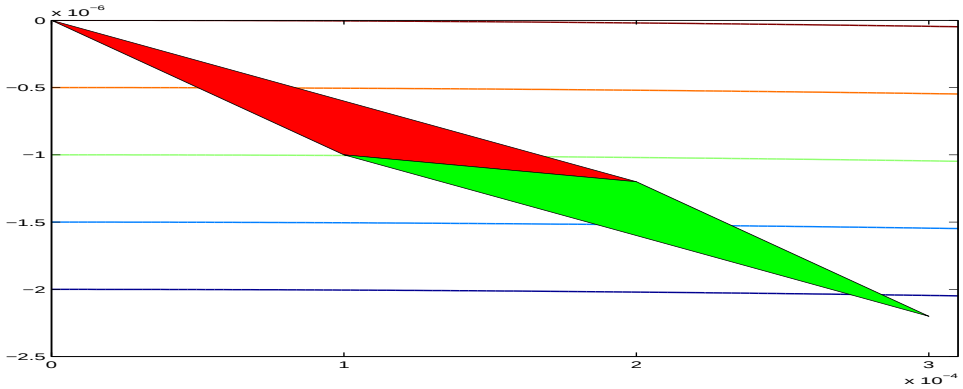

Figure 4: The contours of $\widetilde{y} + \frac{1}{2}\widetilde{x}^2 + \frac{1}{2}\widetilde{y}^2$ are shown along with an RNM triangle with $\widetilde{h}/\widetilde{w}^2 = 30$. The reflection is accepted.

Lemma 3.13. *Under the assumptions of Lemma 3.11, if k is sufficiently large and a contraction step is taken at iteration k, then $\Gamma_k \leq 10$, where Γ_k is the flatness of $\widetilde{\Delta}_k$ as in Definition 3.4.*

Proof. Let $\widetilde{w}$, $\widetilde{h}$, $\widetilde{A}$ be the width, height, and area of Δ_k with respect to the coordinate frame associated by Lemma 3.1 to a vertex of Δ_k. If k is sufficiently large, then Lemma 3.11 implies $\widetilde{h} \leq 10\widetilde{w}^2$. Hence

$$\Gamma_k = \frac{\widetilde{A}}{\widetilde{w}^3} \leq \frac{\widetilde{h}\widetilde{w}}{\widetilde{w}^3} \leq \frac{(10\widetilde{w}^2)\widetilde{w}}{\widetilde{w}^3} = 10.$$ □

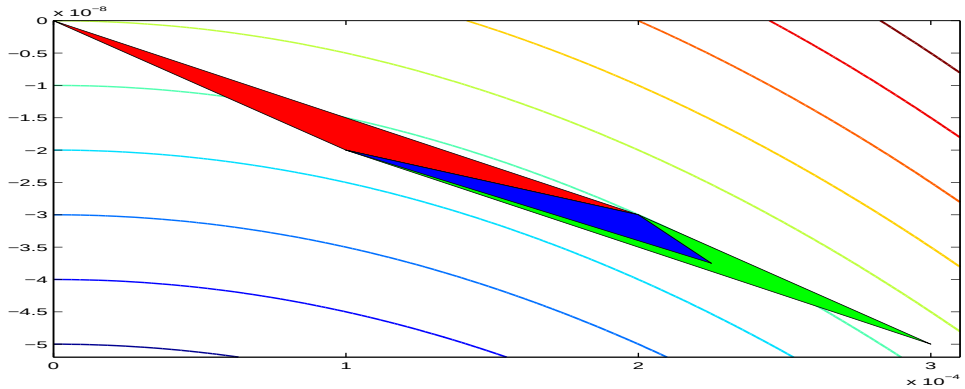

Figure 5: The contours of $\widetilde{y} + \frac{1}{2}\widetilde{x}^2 + \frac{1}{2}\widetilde{y}^2$, are shown along with an RNM triangle with $\widetilde{h}/\widetilde{w}^2 = \frac{3}{4}$. The reflection step is not accepted, and an outside contraction is performed. Note the difference, by four orders of magnitude, between the horizontal and vertical scales.

3.8 Eliminating the impossible: increasing flatness is unavoidable

The final piece of the proof of Theorem 1.2 will show that, for sufficiently advanced iterations, the flatness of the RNM triangles must increase by a factor of at least 1.001 *within a specified number of iterations following a contraction*. To obtain this result, we begin by characterizing the structure of RNM vertices at sufficiently advanced iterations following a contraction, and then defining a related but simpler triangle.

3.8.1 A simpler triangle.

Assume that (i) there is a limit point $\boldsymbol{p}^\dagger$ of the RNM triangles that is not the minimizer of f, (ii) k_0 is sufficiently large, and (iii) iteration k_0 is a contraction. For the RNM triangle Δ_{k_0}, let $\mathfrak{F}_1$ denote the coordinate frame whose base point is the vertex of Δ_{k_0} with the worst value of f:

$$\text{base}(\mathfrak{F}_1) = (\boldsymbol{p}_{\text{worst}})_{k_0}. \tag{3.28}$$

This first coordinate frame is used to identify $\widetilde{\boldsymbol{p}}_{\text{left}}$ and $\widetilde{\boldsymbol{p}}_{\text{right}}$, the transformed vertices of Δ_{k_0} with leftmost and rightmost $\widetilde{x}$ coordinates.

A second coordinate frame, $\mathfrak{F}_2$, is defined next whose base point (measured in frame $\mathfrak{F}_1$) is the midpoint of $[\widetilde{\boldsymbol{p}}_{\text{left}}, \widetilde{\boldsymbol{p}}_{\text{right}}]$:

$$\text{base}(\mathfrak{F}_2) = \tfrac{1}{2}(\widetilde{\boldsymbol{p}}_{\text{left}} + \widetilde{\boldsymbol{p}}_{\text{right}}). \tag{3.29}$$

Unless otherwise specified, the coordinate frame $\mathfrak{F}_2$ is used throughout the remainder of this proof. The base points of $\mathfrak{F}_1$ and $\mathfrak{F}_2$ will be arbitrarily close if k_0 is sufficiently large.

We assume that k_0 is sufficiently large so that the RNM triangles have become tiny in diameter and flattened out (Lemma 3.8). The reason for defining $\mathfrak{F}_2$ is that we can choose a small $\eta > 0$ such that the transformed three vertices of Δ_{k_0}, measured in coordinate frame $\mathfrak{F}_2$, may be expressed as

$$\boldsymbol{a}_0 = \begin{pmatrix} -\eta \\ -u\eta^2 \end{pmatrix}, \quad \boldsymbol{b}_0 = \begin{pmatrix} s\eta \\ t\eta^2 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{c}_0 = \begin{pmatrix} \eta \\ u\eta^2 \end{pmatrix}, \tag{3.30}$$

where vertex $\boldsymbol{a}_0$ corresponds to $\boldsymbol{p}_{\text{left}}$ and vertex $\boldsymbol{c}_0$ to $\boldsymbol{p}_{\text{right}}$.

Without loss of generality the value of s in (3.30) can be taken as nonnegative. The vertices $\boldsymbol{a}_0$ and $\boldsymbol{c}_0$ were leftmost and rightmost when measured in $\mathfrak{F}_1$; by Lemma 3.6(i), the s in (3.30) cannot be too much larger than 1. We assume that k_0 is large enough so that $0 \le s \le 1.00001$.

Because of the form of the vertices in (3.30) and the bounds on s, the transformed width $\widetilde{w}$ of Δ_{k_0} (measured using coordinate frame $\mathfrak{F}_2$) can be no larger than 2.00001η. Iteration k_0 is, by assumption, a contraction, so it follows from Lemma 3.11 that the transformed height of Δ_{k_0} satisfies

$\widetilde{h} \le 10\widetilde{w}^2$, and hence $\widetilde{h} \le 40.0005\eta^2$. Since $\widetilde{h}$ is equal to the larger of $2|u|\eta^2$ or $(|u|+|t|)\eta^2$, it follows that $|u| \le 40.0005$ and $|t| \le 40.0005$ in (3.30).

If Δ and Δ' are any two consecutive RNM triangles in which the same coordinate frame is used, the new vertex of Δ' is a linear combination of the vertices of Δ, with rational coefficients defined by the choice of worst vertex and the nature of the move. (See (3.1)–(3.3).) Furthermore, the values of $\widetilde{w}$ and $\widetilde{h}$ in Δ and Δ' remain the same or decrease, and, if $\boldsymbol{v}$ is any vertex of Δ and $\boldsymbol{v}'$ is any vertex of Δ', then $|\widetilde{x}_{\boldsymbol{v}'} - \widetilde{x}_{\boldsymbol{v}}| \le 2\widetilde{w}$ and $|\widetilde{y}_{\boldsymbol{v}'} - \widetilde{y}_{\boldsymbol{v}}| \le 2\widetilde{h}$. Thus, after $\ell \ge 0$ moves, we reach a triangle $\Delta_{k_0+\ell}$ for which each transformed vertex $\widetilde{\boldsymbol{v}}$ has the form

$$\widetilde{\boldsymbol{v}} = \begin{pmatrix} \lambda\eta \\ \mu\eta^2 \end{pmatrix}, \quad \text{where} \quad |\lambda| \le 1.00001 + 4.00002\ell \ \text{ and } \ |\mu| \le 40.0005(1+2\ell). \tag{3.31}$$

3.8.2 Rescaled inequalities associated with RNM moves.

The next step is to make a rescaling of coordinates to define a triangle $\boldsymbol{\Delta}_\ell$ that is related to $\widetilde{\Delta}_{k_0+\ell}$ by the diagonal affine transformation $\mathrm{diag}(\eta, \eta^2)$. Let $\widetilde{\boldsymbol{p}} = (\lambda\eta, \mu\eta^2)$ be a point in $\widetilde{\Delta}_{k_0+\ell}$ measured in $\mathfrak{F}_2$. Then

$$\widetilde{\boldsymbol{p}} = \begin{pmatrix} \lambda\eta \\ \mu\eta^2 \end{pmatrix} \quad \text{corresponds to} \quad \boldsymbol{P} = \begin{pmatrix} \lambda \\ \mu \end{pmatrix} \ \text{(a point in } \boldsymbol{\Delta}_\ell), \tag{3.32}$$

where λ and μ satisfy the bounds (3.31). The flatness of $\boldsymbol{\Delta}_\ell$, defined as $\mathrm{area}(\boldsymbol{\Delta}_\ell)/(\mathrm{width}(\boldsymbol{\Delta}_\ell))^3$, is equal to the flatness of $\Delta_{k_0+\ell}$ measured in coordinate frame $\mathfrak{F}_2$.

Assume now that $\ell \le 20$; the reason for this limit on ℓ will emerge later in Proposition 3.15. For vertex i of $\Delta_{k_0+\ell}$, equation (3.31) shows that the coefficients in its transformed coordinates satisfy $|\lambda_i| < 82$ and $|\mu_i| < 3000$. By (3.5), (3.6), and (3.31), once k_0 is large enough to make $o(\eta^2)$ sufficiently small, the difference in f values between vertices i and j is

$$\begin{aligned} f(\boldsymbol{v}_i) - f(\boldsymbol{v}_j) &= \eta^2[(\tfrac{1}{2}\lambda_i^2 + \mu_i) - (\tfrac{1}{2}\lambda_j^2 + \mu_j)] + r(\eta\mu_i, \eta^2\lambda_i) - r(\eta\mu_j, \eta^2\lambda_j^2) \\ &= \eta^2[(\tfrac{1}{2}\lambda_i^2 + \mu_i) - (\tfrac{1}{2}\lambda_j^2 + \mu_j)] + o(\eta^2). \end{aligned} \tag{3.33}$$

Let ψ denote the simple quadratic function

$$\psi(\lambda, \mu) := \tfrac{1}{2}\lambda^2 + \mu. \tag{3.34}$$

Then (3.33) shows that, if k_0 is large enough, the following relationships hold between f at vertices of $\Delta_{k_0+\ell}$ and ψ at vertices of $\boldsymbol{\Delta}_\ell$:

$$f(\boldsymbol{v}_i) \ge f(\boldsymbol{v}_j) \quad \text{implies} \quad \psi(\lambda_i, \mu_i) > \psi(\lambda_j, \mu_j) - 10^{-6}, \tag{3.35}$$

where 10^{-6} is not magical, but simply a number small enough so our subsequent results follow.

Example 3.14. For illustration, let $\ell = 0$. Based on (3.30), the vertices of $\boldsymbol{\Delta}_0$ are given by

$$\boldsymbol{A}_0 = \begin{pmatrix} -1 \\ -u \end{pmatrix}, \quad \boldsymbol{B}_0 = \begin{pmatrix} s \\ t \end{pmatrix}, \quad \text{and} \quad \boldsymbol{C}_0 = \begin{pmatrix} 1 \\ u \end{pmatrix}, \tag{3.36}$$

and suppose that $\boldsymbol{a}_0$ is the worst transformed vertex of Δ_{k_0}, i.e. that

$$f(\boldsymbol{a}_0) \ge f(\boldsymbol{b}_0) \quad \text{and} \quad f(\boldsymbol{a}_0) \ge f(\boldsymbol{c}_0).$$

Application of (3.35) gives $\psi(-1,-u) > \psi(s,t) - 10^{-6}$ and $\psi(-1,-u) > \psi(1,u) - 10^{-6}$, i.e.

$$\tfrac{1}{2} - u > \tfrac{1}{2}s^2 + t - 10^{-6} \quad \text{and} \quad 10^{-6} > 2u \ \text{ (a simplification of } \tfrac{1}{2} - u > \tfrac{1}{2} + u - 10^{-6}).$$

In this way, inequalities characterizing the transformed vertices (3.31) of $\Delta_{k_0+\ell}$ when applying the RNM algorithm with function f can be derived in terms of vertices of the simpler triangle $\boldsymbol{\Delta}_\ell$ when applying the RNM algorithm to the function $\psi(\lambda, \mu)$, except that *both possible outcomes of a comparison must be allowed* if the two values of ψ are within 10^{-6}. The importance of (3.35) is that, for $\ell \le 20$, a possible sequence of RNM moves specifying the move type and worst vertex leads to a set of algebraic inequalities in s, t, and u.

3.9 Flatness must increase after no more than 14 steps

In the remainder of this section, we consider the transformed width, area, and flatness of a sequence of RNM triangles, $\Delta_{k_0}, \ldots, \Delta_{k_0+\ell}$, defined using a coordinate frame whose base point is in Δ_{k_0}. Accordingly, notation is needed that separately identifies the RNM triangle being measured and the relevant coordinate frame. The value $\Gamma_k^{(1)}$ will denote the flatness of RNM triangle Δ_k measured in $\mathfrak{F}_1$ of (3.28), and $\Gamma_k^{(2)}$ will denote the flatness of Δ_k measured in $\mathfrak{F}_2$ (3.29), with similar notation for $\widetilde{w}$ and $\widetilde{A}$. Since the base points of coordinate frames $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are in Δ_{k_0}, an essential point is that, when $k > k_0$, the triangle containing the base point of the coordinate frame is different from the triangle being measured.

The result in the following proposition was found using symbolic computation software.

Proposition 3.15. *Assume Hypothesis 1. If k_0 is sufficiently large and a contraction step is taken at iteration k_0, then there exists ℓ with $1 \le \ell \le 14$ such that $\Gamma_{k_0+\ell}^{(2)} > 1.01\, \Gamma_{k_0}^{(2)}$.*

Before giving the proof, we sketch the basic idea. As just described in Section 3.8.2, we are in a situation where two properties apply: (1) the transformed objective function at the scaled point $(\lambda, \mu)^T$ can be very well approximated by the quadratic function $\psi(\lambda, \mu) := \frac{1}{2}\lambda^2 + \mu$ in (3.34), and (2) the RNM move sequences of interest can be analyzed by beginning with an initial simplified (scaled) triangle whose vertices (see (3.36)) involve bounded scalars (s, t, u) that lie in a compact set. Under these conditions, the proof explains how algebraic constraints can be derived that characterize geometrically valid sequences of RNM moves. Further algebraic constraints involving s can also be defined that must be satisfied when the flatness increases by a factor of no more than 1.01.

In principle, one could establish the result of the proposition by numerically checking flatness for all geometrically valid RNM move sequences beginning with the simplified triangle, but this approach is complicated, structureless, and too time-consuming for numerical calculation. Instead, we used Mathematica™ 7.0 to construct symbolic inequalities representing RNM move sequences such that

- s, t, and u are suitably bounded,
- the geometric condition (3.35) for a valid RNM move applies, and
- the flatness increases by a factor of *less than or equal to* 1.01.

Proof of Proposition 3.15. The flatness is not changed by a reflection step as long as the same coordinate frame is retained. Assuming that k_0 is sufficiently large and that the move taken during iteration k_0 is a contraction, we wish to show that there is an index ℓ satisfying $1 < \ell \le 14$ such that the flatness Γ of the RNM triangle $\Delta_{k_0+\ell}$, *measured in coordinate frame* $\mathfrak{F}_2$, must be a factor of at least 1.01 larger than the flatness of Δ_{k_0}, i.e., that

$$\frac{\Gamma_{k_0+\ell}^{(2)}}{\Gamma_{k_0}^{(2)}} = \frac{\widetilde{A}_{k_0+\ell}^{(2)}}{\widetilde{A}_{k_0}^{(2)}} \left(\frac{\widetilde{w}_{k_0}^{(2)}}{\widetilde{w}_{k_0+\ell}^{(2)}} \right)^3 > 1.01. \tag{3.37}$$

Let us prove (3.37) directly for $\ell = 1$ when $\boldsymbol{A}_0$ of (3.36) is the worst vertex of $\boldsymbol{\Delta}_0$ and an inside contraction occurs. In this case, the next triangle $\boldsymbol{\Delta}_1$ has vertices

$$\boldsymbol{A}_1 = \begin{pmatrix} \frac{1}{4}s - \frac{1}{4} \\ \frac{1}{4}t - \frac{1}{4}u \end{pmatrix}, \quad \boldsymbol{B}_1 = \begin{pmatrix} s \\ t \end{pmatrix}, \quad \text{and} \quad \boldsymbol{C}_1 = \begin{pmatrix} 1 \\ u \end{pmatrix}, \tag{3.38}$$

where the first vertex $\boldsymbol{A}_0$ has been replaced. We have two cases:

- If $0 \le s \le 1$, then $\widetilde{w}(\boldsymbol{\Delta}_0) = 2$ and $\widetilde{w}(\boldsymbol{\Delta}_1) = \frac{5}{4} - \frac{1}{4}s \le \frac{5}{4}$, which implies that $\widetilde{w}(\boldsymbol{\Delta}_0)/\widetilde{w}(\boldsymbol{\Delta}_1) \ge \frac{8}{5}$.
- If $1 < s \le 1.00001$, then $\widetilde{w}(\boldsymbol{\Delta}_0) \ge 2$ and $\widetilde{w}(\boldsymbol{\Delta}_1) = \frac{3}{4}s + \frac{1}{4}$, so that $\widetilde{w}(\boldsymbol{\Delta}_1) \le 1.0000075$ and $\widetilde{w}(\boldsymbol{\Delta}_0)/\widetilde{w}(\boldsymbol{\Delta}_1) \ge 1.9999$.

For all s satisfying $0 \le s \le 1.00001$, it follows that $\widetilde{w}(\boldsymbol{\Delta}_0)/\widetilde{w}(\boldsymbol{\Delta}_1) \ge \frac{8}{5}$, and hence that

$$\left(\frac{\widetilde{w}(\boldsymbol{\Delta}_0)}{\widetilde{w}(\boldsymbol{\Delta}_1)}\right)^3 \ge \left(\frac{8}{5}\right)^3 = 4.096.$$

The area of $\boldsymbol{\Delta}_1$ is half the area of $\boldsymbol{\Delta}_0$. Hence the ratio of the flatnesses of $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_0$ satisfies

$$\frac{\Gamma(\boldsymbol{\Delta}_1)}{\Gamma(\boldsymbol{\Delta}_0)} = \frac{\widetilde{A}(\boldsymbol{\Delta}_1)}{\widetilde{A}(\boldsymbol{\Delta}_0)} \left(\frac{\widetilde{w}(\boldsymbol{\Delta}_0)}{\widetilde{w}(\boldsymbol{\Delta}_1)}\right)^3 \ge \tfrac{1}{2}(4.096) > 1.01.$$

The same argument applies when $\boldsymbol{\Delta}_1$ is the result of an *outside* contraction in which vertex $\boldsymbol{A}_0$ is the worst.

But when the sequence of moves begins with a contraction in which vertex $\boldsymbol{B}_0$ or $\boldsymbol{C}_0$ is worst, we must break into further cases, and the analysis becomes too complicated to do by hand. To examine such sequences of RNM moves, we use a Mathematica program that generates inequalities involving vertices of $\boldsymbol{\Delta}_\ell$ and the function ψ of (3.34), as described in Section 3.8.2.

Any sequence of RNM moves (where a move is specified by the worst vertex and the type of move) starting with triangle Δ_{k_0} gives rise to a set of algebraic inequalities in s, t, and u. The i^{th} of these latter inequalities has one of the forms $\phi_i(s) + \nu_i t + \omega_i u > \theta_i$ or $\phi_i(s) + \nu_i t + \omega_i u \ge \theta_i$, where $\phi_i(s)$ is a quadratic polynomial in s with rational coefficients, and ν_i, ω_i, and θ_i are rational constants.

The next step is to determine whether there are acceptable values of s, t, and u for which these inequalities are satisfied. To do so, we begin by treating s as constant (temporarily) and considering the feasibility of a system of *linear* inequalities in t and u, namely the system $Nz \ge d$, where $z = (t \;\; u)^T$, the i^{th} row of N is $(\nu_i \;\; \omega_i)$, and $d_i = \theta_i - \phi_i(s)$. A variant of Farkas' lemma [26, page 89] states that the system of linear inequalities $Nz \ge d$ is feasible if and only if $\gamma^T d \le 0$ for every vector γ satisfying $\gamma \ge 0$ and $N^T\gamma = 0$. If the only nonnegative vector γ satisfying $N^T\gamma = 0$ is $\gamma = 0$, then $Nz \ge d$ is feasible for any d.

The existence (or not) of a nonnegative nonzero γ in the null space of N^T can be determined symbolically by noting that the system $Nz \ge d$ is feasible if and only if it is solvable for every subset of three rows of N. Let $\widehat{N}$ denote the 3×2 matrix consisting of three specified rows of N, with a similar meaning for $\widehat{d}$. To determine the feasibility of $\widehat{N}z \ge \widehat{d}$, we first find a vector $\widehat{\gamma}$ such that $\widehat{N}^T\widehat{\gamma} = 0$.

If $\widehat{N}$ has rank 2, then $\widehat{\gamma}$ is unique (up to a scale factor) and we can write $\widehat{N}^T$ (or a column permutation) so that the leftmost 2×2 submatrix B is nonsingular. Then, with

$$\widehat{N}^T = \begin{pmatrix} \nu_1 & \nu_2 & \nu_3 \\ \omega_1 & \omega_2 & \omega_3 \end{pmatrix} = \begin{pmatrix} B & h \end{pmatrix}, \quad \widehat{\gamma} \text{ is a multiple of } \begin{pmatrix} -B^{-1}h \\ 1 \end{pmatrix},$$

where the components of B^{-1} and h are rational numbers. If (with appropriate scaling) $\widehat{\gamma} \ge 0$ with at least one positive component, then $\widehat{N}^T z \ge \widehat{d}$ is solvable if and only if $\widehat{\gamma}^T\widehat{d} \le 0$. If the components of $\widehat{\gamma}$ do not have the same sign, $\widehat{N}^T z \ge \widehat{d}$ is solvable for any $\widehat{d}$.

If $\widehat{N}$ has rank one, its three rows must be scalar multiples of the same vector, i.e., the i^{th} row is $(\beta_i\nu_1 \;\; \beta_i\omega_1)$, and the null vectors of $\widehat{N}^T$ are linear combinations of $(\beta_2, -\beta_1, 0)^T$, $(0, \beta_3, -\beta_2)^T$, and $(\beta_3, 0, -\beta_1)^T$.

Since the components of d are quadratic polynomials in s and the components of each $\widehat{\gamma}$ are rational numbers, the conditions for feasibility of $Nz \ge d$ (e.g., the conjunction of conditions that $\widehat{\gamma}^T\widehat{d} \le 0$ for each set of three rows of N) can be expressed as a Boolean combination of quadratic inequalities in s with rational coefficients that, for a given value of s, evaluates to "True" if and only if there exist t and u such that these inequalities are satisfied.

To verify the result of the proposition for a given sequence of ℓ RNM moves applied to $\boldsymbol{\Delta}_0$, we need to compute the flatness of $\boldsymbol{\Delta}_\ell$, which is, by construction, equal to the flatness of $\Delta_{k_0+\ell}$ measured in coordinate frame $\mathfrak{F}_2$; see (3.32). We can directly calculate the ratio of the area of $\boldsymbol{\Delta}_\ell$ to the area of $\boldsymbol{\Delta}_0$ by using the number of contractions in the move sequence, since each contraction multiplies the area by $\frac{1}{2}$. The width of $\boldsymbol{\Delta}_\ell$ can be obtained using inequalities and linear polynomials in s, since

the width is determined by the largest and smallest $\widetilde{x}$ coordinates, which are linear polynomials in s. Consequently, the condition that the flatness for each triangle in the sequence is less than 1.01 times the original flatness can be expressed as a Boolean combination of (at most cubic) polynomial inequalities in s, where s is constrained to satisfy $0 \le s \le 1.00001$.

To determine whether there are allowable values of s for which a specified sequence of RNM moves is possible, observe that a Boolean combination of polynomial inequalities in s will evaluate to "True" for s in a certain union of intervals that can be computed as follows. We first find the values of s that are solutions of the polynomial *equations* obtained by replacing any inequalities by equalities. Then, between each adjacent pair of solutions, we choose a test value (e.g., the midpoint) and check whether the associated inequality evaluates to "True" on that interval.

The computation time can be cut in half by considering only sequences that begin with an *inside* contraction, for the following reason. The outside contraction point for an original triangle Δ with vertices $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, and $\boldsymbol{p}_3$ is equal to the inside contraction point for a triangle, denoted by Δ', whose worst vertex $\boldsymbol{p}_3$ is the reflection point $\boldsymbol{p}_{\mathrm{r}}$ of Δ. With exact computation, the conditions for an outside contraction of Δ differ from those for an inside contraction of Δ' if equality holds in some of the comparisons. In particular, if $f(\boldsymbol{p}_3) > f(\boldsymbol{p}_{\mathrm{r}}) \ge f(\boldsymbol{p}_2)$, then Δ will undergo an outside contraction and Δ' will undergo an inside contraction; but if $f(\boldsymbol{p}_3) = f(\boldsymbol{p}_{\mathrm{r}})$, then both Δ' and Δ will undergo inside contractions. Since our inequalities allow for a small error in comparisons, this difference will not change the result, and we may assume that the RNM move at Δ_{k_0} is an inside contraction.

Finally, the definition of the RNM algorithm imposes further constraints on valid move patterns. For example, if a reflection occurs, the reflection point must be strictly better than the second-worst vertex, so this reflection point cannot be the worst point in the new triangle. Such sequences (impossible in the RNM algorithm) would be permitted by the small error allowed in the inequalities, so they are explicitly disallowed in the Mathematica code.

Putting all this together, a program can test each sequence of valid operations that begins with an inside contraction to determine whether there exists an initial triangle for which ratio of the flatnesses, measured in $\mathfrak{F}_2$, is less than 1.01. The results of this computation show that, within no more than 14 RNM moves following a contraction, a triangle is always reached for which the ratio of the flatnesses, measured in the second coordinate frame $\mathfrak{F}_2$, is at least 1.01. We stress that the count of 14 moves includes a mixture of reflections and both forms of contraction. Details of these move sequences can be found in the appendix. There we list the s-values and the associated sequences of 14 or fewer RNM moves for which the ratio of the flatnesses remains less than 1.01. □

Proposition 3.15 used $\mathfrak{F}_2$, but its analogue for $\mathfrak{F}_1$ follows almost immediately with a slightly smaller constant in place of 1.01.

Lemma 3.16. *Under the assumptions of Proposition 3.15, there exists ℓ with $1 \le \ell \le 14$ such that*

$$\Gamma^{(1)}_{k_0+\ell} > 1.001\, \Gamma^{(1)}_{k_0}.$$

Proof. The base point of $\mathfrak{F}_1$ is the worst point of Δ_{k_0}; the base point of $\mathfrak{F}_2$ is the midpoint of the edge of Δ_{k_0} joining the two vertices whose $\widetilde{x}$ coordinates are leftmost and rightmost when measured in $\mathfrak{F}_1$. By choosing k_0 to be large enough, the two base points can be made arbitrarily close. Lemma 3.6(iii) with $\epsilon = 0.0001$ shows that for large enough k_0, the flatnesses of triangles Δ_{k_0} and $\Delta_{k_0+\ell}$ measured in coordinate frames $\mathfrak{F}_1$ and $\mathfrak{F}_2$ satisfy

$$0.9999\, \Gamma^{(1)}_{k_0} \le \Gamma^{(2)}_{k_0} \le 1.0001\, \Gamma^{(1)}_{k_0} \quad \text{and} \quad 0.9999\, \Gamma^{(2)}_{k_0+\ell} \le \Gamma^{(1)}_{k_0+\ell} \le 1.0001\, \Gamma^{(2)}_{k_0+\ell}. \tag{3.39}$$

Now, for ℓ as in Proposition 3.15,

$$\begin{aligned}
\Gamma^{(1)}_{k_0+\ell} &\ge 0.9999\, \Gamma^{(2)}_{k_0+\ell} \\
&> 0.9999(1.01)\Gamma^{(2)}_{k_0} \qquad \text{(by Proposition 3.15)} \\
&\ge 0.9999(1.01)(0.9999)\Gamma^{(1)}_{k_0} \\
&> 1.001\, \Gamma^{(1)}_{k_0}.
\end{aligned}$$

□

3.10 Completion of the proof

The main result of this paper is the following theorem (called Theorem 1.2 in Section 1).

Theorem 3.17. *If the RNM algorithm is applied to a function $f \in \mathcal{F}$, starting from any nondegenerate triangle, then the algorithm converges to the unique minimizer of f.*

Proof. In this proof, $\Gamma_j(\Delta_i)$ denotes the flatness of RNM triangle Δ_i measured in a coordinate frame $\mathfrak{F}_j$ whose base point is the *worst vertex of triangle* Δ_j.

Given a small positive number κ, let k_0 be sufficiently large (we will specify how small and how large as we go along). As mentioned in Section 3.1, the RNM triangle must contract infinitely often, so we may increase k_0 to assume that Δ_{k_0} contracts. Lemma 3.16 shows that the flatness measured in $\mathfrak{F}_{k_0}$ increases by a factor of 1.001 in at most 14 RNM moves; i.e., there exists k_1 with $k_0 < k_1 \le k_0 + 14$ such that

$$\Gamma_{k_0}(\Delta_{k_1}) > 1.001\,\Gamma_{k_0}(\Delta_{k_0}). \tag{3.40}$$

We now switch coordinate frames on the left hand side: Lemma 3.6(iii) and Remark 3.9 show that the flatness of Δ_{k_1} in $\mathfrak{F}_{k_1}$ is close to its flatness in $\mathfrak{F}_{k_0}$. In particular, if k_0 is sufficiently large, then

$$\Gamma_{k_1}(\Delta_{k_1}) \ge 0.9999\,\Gamma_{k_0}(\Delta_{k_1}). \tag{3.41}$$

Let $k_2 \ge k_1$ be the first iteration after (or equal to) k_1 such that Δ_{k_2} contracts. Lemma 3.10 shows that if k_0 is sufficiently large, then from iteration k_1 to the beginning of iteration k_2, the distance travelled by the centroid, measured in $\mathfrak{F}_{k_1}$, is less than κ. During those iterations, the RNM triangle retains its shape and hence its flatness, as measured in $\mathfrak{F}_{k_1}$; that is,

$$\Gamma_{k_1}(\Delta_{k_2}) = \Gamma_{k_1}(\Delta_{k_1}). \tag{3.42}$$

If κ was small enough, Lemma 3.6(iii) and Remark 3.9 again imply

$$\Gamma_{k_2}(\Delta_{k_2}) \ge 0.9999\,\Gamma_{k_1}(\Delta_{k_2}). \tag{3.43}$$

Combining (3.40), (3.41), (3.42), and (3.43) yields

$$\Gamma_{k_2}(\Delta_{k_2}) > (0.9999)^2(1.001)\Gamma_{k_0}(\Delta_{k_0}) > 1.0007\,\Gamma_{k_0}(\Delta_{k_0}).$$

If k_0 is sufficiently large, then repeating the process that led from k_0 to k_2 defines $k_0 < k_2 < k_4 < \cdots$ such that

$$\Gamma_{k_{2n}}(\Delta_{k_{2n}}) > (1.0007)^n\Gamma_{k_0}(\Delta_{k_0})$$

for all n: to know that the same lower bound on k_0 works at every stage, we use that in Lemma 3.6(iii) the number δ is independent of $\boldsymbol{b}_1$, $\boldsymbol{b}_2$, and Δ. Now, if n is sufficiently large, then

$$\Gamma_{k_{2n}}(\Delta_{k_{2n}}) > 10.$$

But $\Delta_{k_{2n}}$ contracts, so this contradicts Lemma 3.13.

Hence the assumption made at the beginning of our long chain of results, Hypothesis 1, must be wrong. In other words, the RNM algorithm *does* converge to the minimizer of f. □

4 Concluding Remarks

4.1 Why do the McKinnon examples fail?

For general interest, we briefly revisit the smoothest McKinnon counterexample (1.1), which consists of a twice-continuously differentiable function f and a specific starting triangle for which the RNM algorithm converges to a nonminimizing point (with nonzero gradient). The Hessian matrix is positive semidefinite and singular at the limit point, but positive definite everywhere else. Thus all the assumptions in our convergence theorem are satisfied except for positive-definiteness of the Hessian, which fails at one point. Hypothesis 1 is valid for this example, and it is enlightening to examine where the proof by contradiction fails.

The McKinnon iterates do satisfy several of the intermediate lemmas in our proof: the RNM triangles not only flatten out (Lemma 3.8), but they do so more rapidly than the rate proved in Lemma 3.11.[4] However, an essential reduction step, Lemma 3.6, fails to hold for the McKinnon example, as discussed below.

Positive-definiteness of the Hessian plays a crucial role in our proof by contradiction because it allows us to uniformly approximate the objective function close to the limit point $\boldsymbol{p}^\dagger$ by its degree-2 Taylor polynomial. Applying a well-defined change of variables, the function $\frac{1}{2}x^2 + y$ for a simple triangle can then be taken as a surrogate, and we can essentially reduce the problem to studying the RNM algorithm for the objective function $\frac{1}{2}x^2+y$ near the non-optimal point $(0, 0)$. In the McKinnon example (1.1), however, the objective function near the limit point $(0, 0)$ cannot be (uniformly) well approximated by $\frac{1}{2}x^2 + y$, even after a change of variable. Although the Hessian of the McKinnon function f remains positive definite at base points in Δ_k as $k \to \infty$, it becomes increasingly close to singular, in such a way that ever-smaller changes in the base point will eventually not satisfy the closeness conditions of Lemma 3.6. In fact, the actual shape of the McKinnon objective function allows a sequence of RNM moves that are forbidden for $\frac{1}{2}x^2 + y$ near the non-optimal point $(0, 0)$. namely an infinite sequence of inside contractions with the best vertex never replaced. In dynamical terms, the McKinnon objective function allows symbolic dynamics forbidden for $\frac{1}{2}x^2 + y$ near $(0, 0)$, and these symbolic dynamics evade the contradiction in our argument.

4.2 An instance of RNM convergence

Most of this paper has been devoted to analysis of situations that we subsequently show cannot occur; this is the nature of arguments by contradiction. For contrast, we present one example where the RNM algorithm will converge, as we have proved, on the strictly convex quadratic function

$$f(x, y) = 2x^2 + 3y^2 + xy - 3x + 5y,$$

whose minimizer is $x^* = (1, -1)^T$. Using starting vertices $(0, 0.5)^T$, $(0.25, -0.75)^T$, and $(-0.8, 0)^T$, after 20 RNM iterations the best vertex is $(0.997986, -1.00128)^T$, and the RNM triangles are obviously converging to the solution. The first nine iterations are depicted in Figure 6.

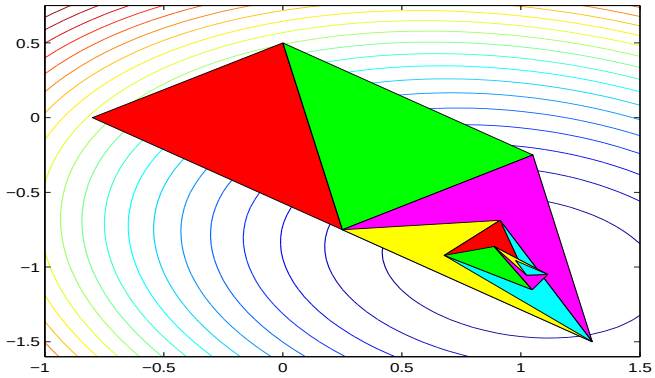

Figure 6: Convergence of the RNM algorithm on a strictly convex quadratic function.

4.3 Significance of the results in this paper

This paper began by noting that very little is known about the theoretical properties of the original Nelder–Mead method, despite 45 years of practice. It is fair to say that proving convergence for an

[4] As $k \to \infty$, the McKinnon triangles satisfy $\widetilde{h}_k \approx \widetilde{w}_k^\theta$ for $\theta = |\lambda_2|(1 + |\lambda_2|)/\lambda_1 \approx 3$, where $\lambda_{1,2} = (1 \pm \sqrt{33})/8$.

RNM algorithm in two dimensions on a restricted class of functions adds only a little more to this knowledge. This contribution seems of interest, however, because of the lack of other results despite determined efforts, and the introduction of dynamical systems methods to the analysis.

Our analysis applies only to a simplified ("small step") version of the original Nelder–Mead method which excludes expansion steps. We have observed that in thousands of computational experiments with functions defined in $\mathbb{R}^n$ ($n \geq 2$) in which the Nelder–Mead method converges to a minimizer, expansion steps are almost never taken in the neighborhood of the optimum. Expansion steps are typically taken early on, forming part of the "adaptation to the local contours" that constituted the motivation for Nelder and Mead when they originally conceived the algorithm [20]. Thus the RNM algorithm appears to represent, to a large extent, the behavior of the original method near the solution. In this direction, it would be valuable if these empirical observations could be rigorously justified under a well-defined set of conditions. The observed good performance of the Nelder–Mead method on many real-world problems remains a puzzle.

This paper applies dynamical systems methods to the analysis of the RNM algorithm. The use of such ideas in the proofs, particularly that of a (rescaled) local coordinate frame in Section 3.8.2, may also be useful in other contexts where it is valuable to connect the geometry of a simplex with the contours of the objective function. The evolving geometric figures of the algorithm remain one of the intuitive appeals of the original Nelder–Mead method, leading to the nickname of "amoeba method" [23]. There may well be other applications, but the latest direct search methods tend to exhibit a less clear connection with geometry.

Finally, our analysis for the RNM algorithm relies in part on the fact that the volume of the RNM simplex is non-increasing at every iteration, thereby avoiding the difficulties associated with expansion steps. Consequently, McKinnon's question remains open: does the original Nelder–Mead algorithm, including expansion steps, always converge for the function $x^2 + y^2$, or more generally for a class of functions like those treated in Theorem 3.17? We hope that further development of the dynamical systems approach could lead to progress on this question.

Appendix: Computation for Proposition 3.15.

This appendix provides details of the symbolic computation performed to prove Proposition 3.15. We regard the coding of moves as a form of *symbolic dynamics* for the RNM iteration. Moves are represented as follows: `1`, `2`, and `3` denote reflections with, respectively, vertex $\boldsymbol{A}$, $\boldsymbol{B}$, or $\boldsymbol{C}$ of (3.36) taken as the worst vertex, i.e. replaced during the move. Similarly, `4`, `5`, and `6` denote inside contractions, and `7`, `8`, `9` denote outside contractions with worst vertex $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, respectively.

We describe a sequence of move numbers as *possible* for a given $s \in [0, 1.00001]$ if there exist $t, u \in [-40.0005, 40.0005]$ such that for the triangle (3.36) described by (s, t, u),

(i) the variables s, t, u satisfy the inequality implied by (3.35) for each RNM move,

(ii) the flatness after each step is less than or equal to 1.01 times the original flatness, and

(iii) no reflection undoes an immediately preceding reflection.

Remark 4.1. Because (3.35) involves a relaxation of 10^{-6}, a sequence characterized as "possible" using the first two properties listed above could be impossible for the RNM algorithm *in exact arithmetic*. This is why the third condition explicitly prohibits sequences in which a reflection undoes the previous move, something that can never happen in the RNM algorithm.

In the proof of Proposition 3.15, we described a symbolic algorithm for computing all possible sequences beginning with an inside contraction. The Mathematica output below lists all these sequences.

```
{5} possible for s in {{0.999999, 1.00001}}
{5, 6} possible for s in {{0.999999, 1.00001}}
```

```
{6} possible for s in {{0.582145, 1.}}
{6, 2} possible for s in {{0.582145, 0.737035}}
{6, 2, 1} possible for s in {{0.582145, 0.695708}}
{6, 2, 1, 3} possible for s in {{0.582145, 0.654949}}
{6, 2, 1, 3, 2} possible for s in {{0.582145, 0.654949}}
{6, 2, 1, 3, 6} possible for s in {{0.582145, 0.654949}}
{6, 2, 1, 3, 6, 2} possible for s in {{0.616769, 0.654949}}
{6, 2, 1, 3, 6, 2, 5} possible for s in {{0.616769, 0.64706}}
{6, 2, 1, 3, 6, 8} possible for s in {{0.582145, 0.64706}}
{6, 2, 1, 3, 6, 8, 4} possible for s in {{0.582145, 0.623495}}
{6, 2, 1, 3, 9} possible for s in {{0.582145, 0.644579}}
{6, 2, 1, 6} possible for s in {{0.582145, 0.695708}}
{6, 2, 1, 9} possible for s in {{0.582145, 0.673138}}
{6, 2, 1, 9, 2} possible for s in {{0.616769, 0.673138}}
{6, 2, 1, 9, 2, 5} possible for s in {{0.616769, 0.64706}}
{6, 2, 1, 9, 8} possible for s in {{0.582145, 0.64706}}
{6, 2, 1, 9, 8, 4} possible for s in {{0.582145, 0.623495}}
{6, 2, 5} possible for s in {{0.582145, 0.737035}}
{6, 2, 5, 4} possible for s in {{0.582145, 0.695708}}
{6, 2, 5, 7} possible for s in {{0.582145, 0.681931}}
{6, 2, 5, 7, 6} possible for s in {{0.582145, 0.635866}}
{6, 2, 5, 7, 9} possible for s in {{0.582145, 0.681931}}
{6, 2, 5, 7, 9, 5} possible for s in {{0.582145, 0.679967}}
{6, 2, 5, 7, 9, 8} possible for s in {{0.582145, 0.663254}}
{6, 2, 5, 7, 9, 8, 4} possible for s in {{0.582145, 0.646912}}
{6, 2, 5, 7, 9, 8, 7} possible for s in {{0.582145, 0.663254}}
{6, 2, 5, 7, 9, 8, 7, 6} possible for s in {{0.582145, 0.663254}}
{6, 2, 5, 7, 9, 8, 7, 6, 5} possible for s in {{0.589537, 0.663254}}
{6, 2, 5, 7, 9, 8, 7, 6, 5, 1} possible for s in {{0.589537, 0.635373}}
{6, 2, 5, 7, 9, 8, 7, 9} possible for s in {{0.582145, 0.65445}}
{6, 2, 5, 7, 9, 8, 7, 9, 5} possible for s in {{0.582145, 0.651784}}
{6, 2, 5, 7, 9, 8, 7, 9, 5, 4} possible for s in {{0.582145, 0.651784}}
{6, 2, 5, 7, 9, 8, 7, 9, 5, 4, 3} possible for s in {{0.582145, 0.651784}}
{6, 2, 5, 7, 9, 8, 7, 9, 8} possible for s in {{0.597869, 0.65445}}
{6, 2, 5, 7, 9, 8, 7, 9, 8, 4} possible for s in {{0.597869, 0.65445}}
{6, 2, 5, 7, 9, 8, 7, 9, 8, 4, 6} possible for s in {{0.597869, 0.65445}}
{6, 2, 5, 7, 9, 8, 7, 9, 8, 4, 6, 2} possible for s in {{0.597869, 0.654004}}
{6, 2, 5, 7, 9, 8, 7, 9, 8, 4, 6, 2, 5} possible for s in {{0.64094, 0.654004}}
{6, 2, 5, 7, 9, 8, 7, 9, 8, 4, 6, 8} possible for s in {{0.64094, 0.65445}}
{6, 2, 8} possible for s in {{0.582145, 0.614711}}
{6, 5} possible for s in {{0.582145, 1.}}
{6, 8} possible for s in {{0.582145, 0.853944}}
{6, 8, 4} possible for s in {{0.582145, 0.810502}}
{6, 8, 7} possible for s in {{0.582145, 0.853944}}
{6, 8, 7, 6} possible for s in {{0.582145, 0.853944}}
{6, 8, 7, 9} possible for s in {{0.582145, 0.818183}}
{6, 8, 7, 9, 5} possible for s in {{0.582145, 0.811611}}
{6, 8, 7, 9, 8} possible for s in {{0.582145, 0.818183}}
{6, 8, 7, 9, 8, 4} possible for s in {{0.582145, 0.818183}}
{6, 8, 7, 9, 8, 4, 6} possible for s in {{0.763168, 0.818183}}
{6, 8, 7, 9, 8, 4, 6, 2} possible for s in {{0.763168, 0.817831}}
{6, 8, 7, 9, 8, 7} possible for s in {{0.582145, 0.777853}}
```

```
{6, 8, 7, 9, 8, 7, 6} possible for s in {{0.582145, 0.777853}}
{6, 8, 7, 9, 8, 7, 6, 5} possible for s in {{0.589537, 0.777853}}
{6, 8, 7, 9, 8, 7, 6, 5, 1} possible for s in {{0.589537, 0.777853}}
{6, 8, 7, 9, 8, 7, 9} possible for s in {{0.582145, 0.751661}}
{6, 8, 7, 9, 8, 7, 9, 5} possible for s in {{0.582145, 0.751661}}
{6, 8, 7, 9, 8, 7, 9, 5, 4} possible for s in {{0.582145, 0.751661}}
{6, 8, 7, 9, 8, 7, 9, 5, 4, 3} possible for s in {{0.582145, 0.751661}}
{6, 8, 7, 9, 8, 7, 9, 8} possible for s in {{0.597869, 0.694824}}
{6, 8, 7, 9, 8, 7, 9, 8, 4} possible for s in {{0.597869, 0.694824}}
{6, 8, 7, 9, 8, 7, 9, 8, 4, 6} possible for s in {{0.597869, 0.694824}}
{6, 8, 7, 9, 8, 7, 9, 8, 4, 6, 2} possible for s in {{0.597869, 0.694824}}
{6, 8, 7, 9, 8, 7, 9, 8, 4, 6, 2, 5} possible for s in {{0.64094, 0.663616}}
{6, 8, 7, 9, 8, 7, 9, 8, 4, 6, 8} possible for s in {{0.64094, 0.663616}}
```

All we need from this computation is that there is no possible sequence of 14 steps or more. In other words, following an inside contraction, the flatness will be greater than 1.01 times the original flatness after no more than 14 steps (including the initial contraction).

Remarks about the list of possible sequences

The remarks in this section are not needed for the proof, but they may give further insight into the behavior of the RNM algorithm as well as clear up some potential ambiguity about the computer output above.

- That the sequence $\{4\}$ is not possible (i.e., that an inside contraction with $\boldsymbol{A}_0$ as worst vertex immediately increases the flatness by at least a factor of 1.01) was shown already near the beginning of the proof of Proposition 3.15.

- The bound 40.0005 on $|t|$ and $|u|$ need not be fed into the program, because the program automatically calculates stronger inequalities that are necessary for a contraction to occur.

- Move sequences that do not appear in the list may still occur in actual runs of the RNM algorithm, but then the flatness must grow by more than a factor of 1.01. Similarly, a move sequence appearing in the list may occur while running the RNM algorithm even if s lies outside the given interval. For example, one can show that there exist triangles with $0 \le s < 0.582145$ on which the RNM algorithm takes move $\{6\}$.

- One cannot predict from the list *which* step causes the flatness to grow beyond the factor of 1.01. For example, using our definition the sequence $\{6, 2, 1, 3, 2\}$ is possible (for a certain range of s), but the extended sequence $\{6, 2, 1, 3, 2, 1\}$ is not. This should not be taken to mean that the last reflection $\{1\}$ caused the increase in flatness, since reflections do not change the flatness (measured in the same coordinate frame). Rather, there may exist a triangle in the given range that for the objective function $f(\lambda, \mu) = \frac{1}{2}\lambda^2 + \mu$ will take the sequence of steps $\{6, 2, 1, 3, 2, 1\}$. What must be the case, however, is that for any such triangle the initial inside contraction $\{6\}$ will have already increased the invariant by a factor at least 1.01.

- One cannot deduce that in every run of the RNM algorithm, every sufficiently advanced sequence of 14 steps involves a contraction. Experiments show that, when omitting any test for flatness, a sequence beginning with $\{6\}$ can legitimately be followed by a very large number of reflect steps during which the flatness does not change. Thus we truly needed Lemma 3.10 in addition to Proposition 3.15 to complete our proof.

- The entire computation took about 11 minutes on an Intel Xeon 3.0 GHz processor.

References

[1] C. Audet (2004). Convergence results for pattern search algorithms are tight, *Optimization and Engineering* 5, 101–122.

[2] C. Audet and J. E. Dennis, Jr. (2003). Analysis of generalized pattern searches, *SIAM Journal on Optimization* 13, 889-903.

[3] C. Audet and J. E. Dennis, Jr. (2006). Mesh adaptive direct search algorithms for constrained optimization, *SIAM Journal on Optimization* 17, 188–217.

[4] D. Bertsekas (2003). *Convex Analysis and Optimization*, Athena Scientific.

[5] A. R. Conn, K. Scheinberg, and L. N. Vicente (2009). *Introduction to Derivative-Free Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

[6] I. D. Coope and C. J. Price (2001). On the convergence of grid-based methods for unconstrained optimization, *SIAM Journal on Optimization* 11, 859–869.

[7] GNU Scientific Library (2011). NLopt, Free Software Foundation, Boston, Massachusetts. `ab-initio.mit.edu/wiki/index.php/NLopt_Algorithms`

[8] A. P. Gurson (2000). "Simplex search behavior in nonlinear optimization", Bachelor's honors thesis, Computer Science Department, College of William and Mary, Williamsburg, Virginia. `www.cs.wm.edu/~va/CS495`

[9] L. Han and M. Neumann (2006). Effect of dimensionality on the Nelder–Mead simplex method, *Optimization Methods and Software* 21, 1–16.

[10] D. Hensley, P. Smith, and D. Woods (1988). Simplex distortions in Nelder–Mead reflections, IMSL Technical Report Series No. 8801, IMSL, Inc., Houston, Texas.

[11] C. T. Kelley (1999). Detection and remediation of stagnation in the Nelder–Mead algorithm using a sufficient decrease condition, *SIAM Journal on Optimization* 10, 43–55.

[12] C. T. Kelley (1999). *Iterative Methods for Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

[13] F. Klein (1939). *Elementary Mathematics from an Advanced Standpoint: Geometry*, Dover Publications: New York. (Reprint of Volume II of English translation of F. Klein, *Elementarmathematik vom Höheren Standpunkte aus*, J. Springer, Berlin 1924–1928.)

[14] T. G. Kolda, R. M. Lewis, and V. Torczon (2003). Optimization by direct search: new perspectives on some classical and modern methods, *SIAM Review* 45, 385–482.

[15] J. C. Lagarias, J. A. Reeds, M. H. Wright and P. E. Wright (1998). Convergence properties of the Nelder–Mead simplex algorithm in low dimensions, *SIAM Journal on Optimization* 9, 112–147.

[16] R. M. Lewis, V. Torczon, and M. W. Trosset (2001). Direct search methods: then and now, in *Numerical Analysis 2000*, Volume 4, 191–207, Elsevier, New York.

[17] *MATLAB™ User's Guide* (2010). R2010b Documentation, The Mathworks, Inc., Natick, Massachusetts. `www.mathworks.com/help/techdoc/ref/fminsearch.html`

[18] K. I. M. McKinnon (1998). Convergence of the Nelder–Mead simplex method to a non-stationary point, *SIAM Journal on Optimization* 9, 148–158.

[19] L. J. Nazareth and P. Tseng (2002). Gilding the lily: A variant of the Nelder–Mead algorithm based on golden section search, *Computational Optimization and Applications* 22, 133–144.

[20] J. A. Nelder and R. Mead (1965). A simplex method for function minimization, *Computer Journal* 7, 308–313.

[21] J. M. Ortega and W. C. Rheinboldt (1970). *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

[22] M. J. D. Powell (1998). Direct search algorithms for optimization calculations, *Acta Numerica* 7, 287–336.

[23] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering (1992). *Numerical Recipes in Fortran: the Art of Scientific Computing* (second ed.), Cambridge University Press, Cambridge, UK.

[24] C. J. Price, I. D. Coope, and D. Byatt (2002). A convergent variant of the Nelder–Mead algorithm, *Journal of Optimization Theory and Applications* 113, 5–19.

[25] A. S. Rykov (1983). Simplex algorithms for unconstrained optimization, *Problems of Control and Information Theory* 12, 195–208.

[26] A. Schrijver (1987). *Theory of Linear and Integer Programming*, John Wiley and Sons, New York.

[27] A. M. Stuart and A. R. Humphries (1996). *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK.

[28] V. Torczon (1997). On the convergence of pattern search algorithms, *SIAM Journal on Optimization* 7, 1–25.

[29] P. Tseng (1999). Fortified-descent simplicial search method: A general approach, *SIAM Journal on Optimization*, 10, No. 1, 269–288.

[30] D. J. Woods (1985). *An Interactive Approach for Solving Multi-Objective Optimization Problems*, PhD thesis, Technical Report 85-5, Department of Computational and Applied Mathematics, Rice University, Houston, Texas.

[31] M. H. Wright (1996). Direct search methods: once scorned, now respectable. in *Numerical Analysis 1995: Proceedings of the 1995 Dundee Biennial Conference in Numerical Analysis*, D. F. Griffiths and G. A. Watson (eds.), 191–208, Addison Wesley Longman, Harlow, UK.