4-1-2005

Spaces Of Polygons In The Plane And Morse Theory

Don H. Shimamoto
Swarthmore College, dshimam1@swarthmore.edu

C. Vanderwaart

Spaces of Polygons in the Plane and Morse Theory

Don Shimamoto and Catherine Vanderwaart

1. INTRODUCTION. The side-side-side theorem of high school geometry states that, if the corresponding sides of two triangles have the same lengths, then the triangles are congruent. In other words, after an appropriate sequence of rotations, reflections, and translations, either triangle can be placed exactly on top of the other. If one imagines physically moving around triangles while remaining in the plane, then only the orientation-preserving motions—rotations and translations—are available. For instance, any triangle having side lengths 2, 3, and 4 can be translated and rotated into one of the two positions in Figure 1.

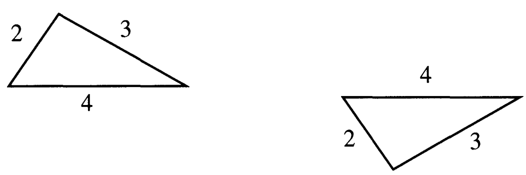

Figure 1. The two triangles of side lengths 2, 3, and 4.

For polygons in the plane having more than three sides, the lengths of the sides no longer determine the polygon. For example, a quadrilateral required to have all sides of equal length can take the shape of an infinite number of noncongruent rhombuses.

In this article we think of polygons having prescribed side lengths as forming a "space." For triangles, the space consists typically of two points, which represent a specific triangle and its mirror image. But for multisided polygons, the space is more complicated and more interesting. Indeed, according to recent results, many (and, in some branches of mathematics, all) important spaces are topologically equivalent to generalized versions of spaces of polygons. (See the last paragraph of the introduction for elaboration on this point.)

To formalize this, let positive real numbers $\ell_1, \ell_2, \dots, \ell_n$ be given, where we always assume that $n \geq 3$. These will be our side lengths. For convenience, we write $\overrightarrow{\ell} = (\ell_1, \ell_2, \dots, \ell_n)$, a vector in $\mathbf{R}^n$. We describe a polygon by listing its vertices. Thus the set of all planar polygons having prescribed side lengths is given by:

$$\mathcal{P}(\overrightarrow{\ell}) = \big\{(p_1, p_2, \dots, p_n) \in (\mathbf{R}^2)^n : |p_2 - p_1| = \ell_1, |p_3 - p_2| = \ell_2, \dots, \\ |p_n - p_{n-1}| = \ell_{n-1}, |p_1 - p_n| = \ell_n\big\}.$$

A representative element of $\mathcal{P}(\overrightarrow{\ell})$ is shown in Figure 2.

The elements of $\mathcal{P}(\overrightarrow{\ell})$ are also called *closed chains* or *polygonal linkages*. The space of polygons $\mathcal{P}(\overrightarrow{\ell})$ is easy to write down and convenient to work with in this form, but the convenience comes at a price. Namely, the term "polygon" must now be

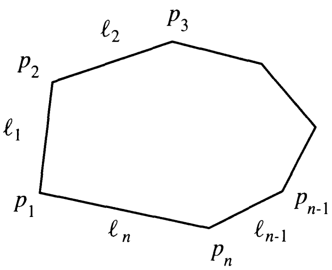

Figure 2. A polygon in the plane with side lengths $\ell_1, \ell_2, \dots, \ell_n$.

taken somewhat loosely, since the definition of $\mathcal{P}(\overrightarrow{\ell})$ allows some of the vertices to coincide and the edges may intersect at points other than vertices.

Two polygons $(p_1, \dots, p_n)$ and $(q_1, \dots, q_n)$ are called *equivalent* if there is a composition σ of rotations and translations of the plane that takes one onto the other, that is, $\sigma(p_i) = q_i$ for all i. One can then consider the set of equivalence classes; this is called the *moduli space* of polygons with side lengths $\ell_1, \dots, \ell_n$. Alternatively, after a translation and rotation, any polygon may be placed so that one vertex, say p_1, is at the origin and an adjacent side, say the one connecting p_1 and p_n, lies along the positive x-axis. Thus the set of equivalence classes is represented by the subset of $\mathcal{P}(\overrightarrow{\ell})$ given by:

$$\mathcal{M}(\overrightarrow{\ell}) = \left\{(p_1, p_2, \dots, p_n) \in \mathcal{P}(\overrightarrow{\ell}) : p_1 = (0,0),\, p_n = (\ell_n, 0)\right\}.$$

It is in this form that we will work with the moduli space.

The purpose of this article is to study the topology of $\mathcal{M}(\overrightarrow{\ell})$. For instance, generically, it is a smooth orientable manifold. The approach is to consider a larger space in which ℓ_n is allowed to vary while $\ell_1, \dots, \ell_{n-1}$ remain fixed. The individual $\mathcal{M}(\overrightarrow{\ell})$ then appear as strata parametrized by ℓ_n. The changes in these strata as ℓ_n varies can be tracked using some basic principles from Morse theory that show how to use critical points of functions of several variables—maxima, minima, saddle points, and the like—to obtain information about the topology of manifolds. The idea to use Morse theory to study configuration spaces of linkages is well known (see particularly Hausmann **[7]** and also Kapovich and Millson **[11]** and Milgram and Trinkle **[17]**). The main general results in this article are contained in these references, but the arguments here have been simplified so that, for the most part, they follow from standard parts of the upper-level undergraduate mathematics curriculum. The paper concludes by applying the results to an example—the topology of spaces of pentagons—in which the answers and methods can be visualized easily. For instance, we recover the fact, observed by several authors, that the space of equilateral pentagons is topologically a surface that is like a torus, only with four holes instead of one.

This material was developed as background for a problem on which we had begun to work in the summer of 2002. The intent of the article is that it be accessible to advanced undergraduates, say those who have studied algebra and analysis, including an introduction to manifolds (e.g., the material that usually precedes the statement of Stokes's theorem). Indeed, part of the appeal of the topic is that it provides a situation in which so many "textbook" techniques can be applied. The only pieces of background material that do lie outside the standard undergraduate curriculum are the principles from Morse theory alluded to earlier, so these are introduced and explained as we go.[1] In general, the paper is written as an extended calculation with expository review material interspersed as necessary.

Finally, to provide some context, it might be worthwhile to mention a few places where related spaces have appeared fairly recently. (The following sampling is not systematic. For a thorough survey, see Connelly and Demaine [**4**].) For instance, suppose that one restricts one's attention to honest polygons, that is, the vertices are required to be distinct and edges can intersect only at common endpoints. Then there is the recent theorem of Connelly, Demaine, and Rote that states that any polygon can be continuously deformed (or "unfolded") into a convex one [**5**]. This has implications in robotics. Or, leaving the plane, the moduli space of polygons in $\mathbf{R}^3$ has a symplectic structure, which opens up a stable of new techniques, including some from algebraic geometry. This has been used by Kapovich and Millson [**12**] and Hausmann and Knutson [**8**], among others. One can also study polygons in manifolds other than Euclidean space, where distances between vertices are measured along the manifold. One of the early examples of this is the work of Kirk and Klassen [**15**], who studied certain representations of the fundamental groups of a type of 3-dimensional manifold and showed that these representations correspond to polygons on a 3-dimensional sphere.

Returning to the plane, one might allow linkages that are not necessarily polygonal, in other words, sequences of points where the distances between certain points are prescribed but the edges need not form a polygon. In this setting, Kapovich and Millson proved a universality, or realization, theorem that shows that moduli spaces of linkages are really quite common. To state their result, suppose that X is a compact real algebraic set in $\mathbf{R}^n$ (i.e., the locus of zeroes of a polynomial in n variables with real coefficients). Then there exists a planar linkage whose moduli space is homeomorphic to a union of disjoint copies of X [**13**]. King showed that, if one allows linkages in which the distances between some vertices are permitted to vary up to prescribed maximum values and the positions of some vertices may be fixed, then there exists a linkage whose moduli space is homeomorphic to X on the nose [**14**]. In particular, this is true for any compact manifold X. These results provide a resolution to questions that had remained open for over a hundred years.

2. PRELIMINARY EXAMPLES.

We begin by examining a couple of concrete examples in order to clarify what a moduli space of polygons is and to introduce the types of questions in which we are interested. Triangles were discussed in the introduction, so let us consider here an example involving quadrilaterals, say the moduli space of all quadrilaterals with side lengths $\ell_1 = 5$, $\ell_2 = 3$, $\ell_3 = 1$, and $\ell_4 = 8$. In keeping with our conventions, $p_1 = (0, 0)$ and $p_4 = (8, 0)$ are fixed, so only p_2 and p_3 are allowed to vary. Since $\ell_1 = 5$, p_2 is constrained to be on the circle of radius 5 centered at p_1. It can't be just anywhere on this circle, however: it is at most $\ell_2 + \ell_3 = 4$ units away from p_4. Hence p_2 is restricted to lie on an arc subtended by an angle going from, say, α radians above the positive x-axis to $-\alpha$ below. If p_2 happens to be one of the endpoints of this arc, then p_2, p_3, and p_4 are collinear. For all other locations of p_2, there are exactly two choices of p_3 satisfying $\ell_2 = 3$ and $\ell_3 = 1$, situated symmetrically about the line through p_2 and p_4. One might imagine p_2 sliding from the top of the arc to the bottom and then back up, tracing with it a sequence of quadrilaterals as in Figure 3. (In the figure, the two possibilities for p_3 for a given location of p_2 appear horizontally opposite one another.)

A few observations are immediate. First, in constructing these quadrilaterals, there is essentially one degree of freedom—as we have presented it, it is the location of p_2

[1] The amount of Morse theory needed is small. A reader seeking a fuller treatment of this material, including proofs, can find it in the first thirteen pages of Milnor [**18**].

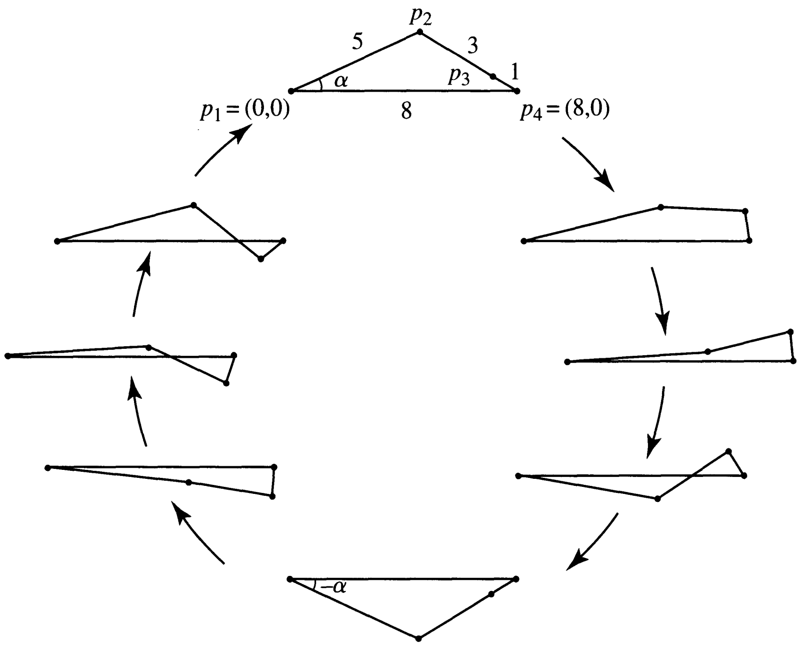

Figure 3. The moduli space of quadrilaterals with $\ell_1 = 5$, $\ell_2 = 3$, $\ell_3 = 1$, $\ell_4 = 8$.

along the arc. In other words, the moduli space $\mathcal{M}(\vec{\ell})$ is one-dimensional. Second, by interpolating between the quadrilaterals shown in Figure 3, any quadrilateral in the moduli space can be continuously deformed into any other. Since each quadrilateral represents a point in $\mathcal{M}(\vec{\ell})$, this means that, as a topological space, $\mathcal{M}(\vec{\ell})$ is connected. Finally, every quadrilateral in the moduli space appears somewhere along the continuum suggested by the figure, so that the entire moduli space consists of a loop's worth of quadrilaterals. More precisely, $\mathcal{M}(\vec{\ell})$ is homeomorphic to a circle. We have determined the topological type of $\mathcal{M}(\vec{\ell})$ completely.

We might hope to use our answer in this example to obtain information about moduli spaces for other combinations of side lengths. It seems reasonable to expect that, if the side lengths change only slightly, the moduli space won't change very much either. For example, if we decrease ℓ_4 from 8 to 7.9, the specific quadrilaterals are different, but the overall structure of a loop of quadrilaterals remains the same. As a general rule, this sort of reasoning is correct. The danger is that, after a while, small changes can build up into big ones, and it seems reckless to assert that the moduli space stays the same no matter what.

For instance, suppose that we keep $\ell_1 = 5$, $\ell_2 = 3$, and $\ell_3 = 1$, but let $\ell_4 = 4$. An example of such a quadrilateral is shown in Figure 4.

Again, there is one degree of freedom in the placement of p_2, but now the triangle inequality implies that p_2 cannot lie anywhere on the x-axis. Therefore, by the intermediate value theorem, no quadrilateral with p_2 above the x-axis can be continuously deformed into one with p_2 below the x-axis. In other words, the moduli space has become disconnected. It now has two components, each homeomorphic to a circle. Somewhere along the line, as ℓ_4 decreases from 8 to 4, something happens. We might try to find out what it is and where it occurs.

Towards that end, in sections 3 and 4 we establish some basic facts about the topology of a moduli space of n-gons. In particular, we find its dimension. Then in sections 5 and 6 we study techniques that provide information about how the topology

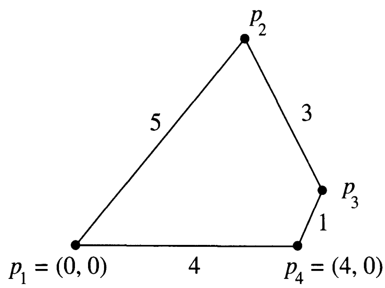

Figure 4. A quadrilateral with $\ell_1 = 5$, $\ell_2 = 3$, $\ell_3 = 1$, $\ell_4 = 4$.

can change as the side lengths vary. In section 7 we apply this to spaces of pentagons, where it is not so easy to determine intuitively what the moduli spaces should be, yet the answers turn out to be tangible geometric objects. We'll see examples in which we can predict how the moduli spaces morph into one another as the side lengths change. We encourage the reader to follow that discussion by returning to the present section to work through the corresponding details for quadrilaterals.

Before we can do any of this, however, we must develop the appropriate theoretical background. It is to this task that we now turn.

3. POLYGONS WITH ONE VARIABLE SIDE: THE SPACE OF ARMS. The techniques described in this article apply to smooth maps. In this section and the next we show that the spaces in which we are interested are smooth manifolds, so that the techniques may be used. (Roughly speaking, a manifold M of dimension m is a space that can be identified locally with $\mathbf{R}^m$. That is, M can be covered by open sets, called *coordinate patches*, each of which is homeomorphic to an open subset of $\mathbf{R}^m$. The homeomorphisms assign m-dimensional coordinates to the associated patches. A manifold M is called *smooth* if whenever two patches overlap the coordinates in one depend smoothly on the coordinates in the other. The degree of smoothness here can be made part of the structure as well, but for the sake of simplicity we always assume smooth maps to be of class C^∞.)

We begin by looking at a space larger than $\mathcal{M}(\overrightarrow{\ell})$, namely, the moduli space of all n-gons in $\mathbf{R}^2$ having one flexible side. These are polygons satisfying the conditions that:

(1) the first vertex p_1 is fixed at the origin;

(2) the first $n-1$ sides have prescribed lengths $\ell_1, \dots, \ell_{n-1}$, respectively; and

(3) the remaining side is constrained to lie along the x-axis, although no restriction is imposed on its (nonzero) length.

In other words, let

$$W = \big\{(p_2, p_3, \dots, p_n) \in (\mathbf{R}^2)^{n-1} : |p_2| = \ell_1, |p_3 - p_2| = \ell_2, \dots, \\ |p_n - p_{n-1}| = \ell_{n-1}, p_n = (x_n, 0), x_n \neq 0\big\}. \qquad (3.1)$$

The polygon space $\mathcal{M}(\overrightarrow{\ell})$ sits inside W as the subset $x_n = \ell_n$. We refer to the elements of W as *arms* and to W as the *space of arms* associated to $\mathcal{M}(\overrightarrow{\ell})$. An example of an arm is given in Figure 5. We prove in Proposition 1 that W is an orientable manifold, so let us review briefly the most common approach for obtaining such a result.

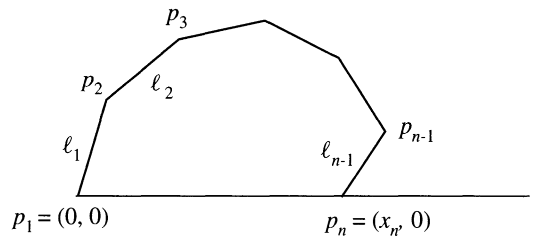

Figure 5. An arm.

Regular Value Theorem. *Let U be an open subset of $\mathbf{R}^m$, and let $f: U \to \mathbf{R}^n$ be a smooth map. Suppose that x in $\mathbf{R}^n$ is a point such that for each p in $f^{-1}(x)$ the Jacobian matrix $Df(p)$ has rank n. Then $f^{-1}(x)$ is a smooth orientable manifold that, if nonempty, is of dimension $m - n$.*

Proof. The part that $f^{-1}(x)$ is a smooth manifold of dimension $m - n$ is completely standard and is a consequence of the implicit function theorem (see, for instance, Munkres [**19**, p. 207 and Exercise 2, pp. 208–209] or A. Browder [**3**, p. 259], among analysis texts currently in use). The part about orientability is no secret either, but it tends to be omitted from introductory treatments, so we indicate the argument here.

One way to orient $M = f^{-1}(x)$ is to orient the tangent space T_pM at every point p in M (i.e., to find a consistent system of ordered bases for each T_pM). Let $N_pM = (T_pM)^{\perp}$ denote the space of normal vectors to M at p. It suffices to find an orienting basis $w_1, \ldots, w_n$ for each of these normal spaces, since we can then define an orientation on T_pM by declaring that an ordered basis $v_1, \ldots, v_{m-n}$ for T_pM belongs to the orientation if and only if $v_1, \ldots, v_{m-n}, w_1, \ldots, w_n$ gives the standard orientation of $\mathbf{R}^m$.

The normal space is easier to work with than the tangent space. The function f is constant along M, hence the Jacobian $Df(p)$ sends the entire $(m - n)$-dimensional vector space T_pM to 0. Since $Df(p)$ has rank n, it must map the complementary space N_pM isomorphically onto $\mathbf{R}^n$. Thus for $j = 1, \ldots, n$ there is a unique w_j in N_pM such that $Df(p) \cdot w_j$ is the standard basis vector $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)$ in $\mathbf{R}^n$. The vectors $w_1, \ldots, w_n$ define an orientation on N_pM in a canonical way. ■

A point x in $\mathbf{R}^n$ satisfying the assumptions of the theorem is called a *regular value* of f. This includes the vacuous case where $f^{-1}(x)$ is empty. All other x in $\mathbf{R}^n$ are called *critical values* of f. Similarly, over on the domain side a point p of U is called a *regular point* if $Df(p)$ has rank n and a *critical point* if not.

The regular value theorem is true for smooth maps between manifolds $f: M \to N$ as well, only now one must work locally, choosing local coordinates to convert f to a map between Euclidean spaces and computing the Jacobian in terms of these coordinates. The value of an individual partial derivative depends on the choice of coordinates, but the rank of $Df(p)$ does not. Moreover, $f^{-1}(x)$ is again orientable if M is orientable.

Proposition 1. *The space of arms W is a smooth orientable manifold of dimension $n - 2$.*

Proof. Consider the open subset U of $(\mathbf{R}^2)^{n-1} = \mathbf{R}^{2n-2}$ consisting of all sequences $(p_2, p_3, \ldots, p_n)$, where $p_j = (x_j, y_j)$ and $x_n \neq 0$. The conditions (3.1) defining W as a subset of U determine a function $F: U \to \mathbf{R}^n$ given by:

$$
\begin{aligned}
F(p_2, \dots, p_n) & \\
&= \left(|p_2|^2, |p_3 - p_2|^2, \dots, |p_n - p_{n-1}|^2, y_n\right) \\
&= \left(x_2^2 + y_2^2, (x_3 - x_2)^2 + (y_3 - y_2)^2, \dots, (x_n - x_{n-1})^2 + (y_n - y_{n-1})^2, y_n\right), \quad (3.2)
\end{aligned}
$$

so $W = F^{-1}(\ell_1^2, \ell_2^2, \dots, \ell_{n-1}^2, 0)$. (The function F is related to a function known in other contexts as the *rigidity map*; see Asimow and Roth [**2**], for example.) The proposition follows from the regular value theorem provided that $(\ell_1^2, \ell_2^2, \dots, \ell_{n-1}^2, 0)$ is a regular value of F. Hence it suffices to show that $DF(p)$ has rank n for each $p = (p_2, \dots, p_n)$ in W.

Regarding the Jacobian, let us fix some notation. We view F as a function of the variables $(x_2, y_2, x_3, y_3, \dots, x_n, y_n)$ in that order. Thus $DF(p)$ is the $n \times (2n-2)$ matrix whose columns are the partial derivatives of F with respect to these variables. From (3.2) we infer that $DF(p)$ has the form:

$$
\begin{bmatrix}
2x_2 & 2y_2 & 0 & 0 & \dots & 0 & 0 \\
-2(x_3 - x_2) & -2(y_3 - y_2) & 2(x_3 - x_2) & 2(y_3 - y_2) & \dots & 0 & 0 \\
 & \vdots & & & \ddots & & \vdots \\
0 & 0 & 0 & 0 & \dots & 2(x_n - x_{n-1}) & 2(y_n - y_{n-1}) \\
0 & 0 & 0 & 0 & \dots & 0 & 1
\end{bmatrix}. \quad (3.3)
$$

Here it is convenient to group the columns into pairs and decompose (3.3) into 1×2 blocks. In this form, $DF(p)$ is the $n \times (n-1)$ matrix whose (i, j)th entry is the block

$$
\begin{bmatrix} \dfrac{\partial F_i}{\partial x_{j+1}} & \dfrac{\partial F_i}{\partial y_{j+1}} \end{bmatrix},
$$

where F_i denotes the ith component of F. Then the expression for $DF(p)$ in (3.3) becomes:

$$
2\begin{bmatrix}
p_2 & 0 & 0 & \dots & 0 & 0 \\
-(p_3 - p_2) & p_3 - p_2 & 0 & \dots & 0 & 0 \\
0 & -(p_4 - p_3) & p_4 - p_3 & \dots & 0 & 0 \\
 & \vdots & & \ddots & & \vdots \\
0 & 0 & 0 & \dots & -(p_n - p_{n-1}) & p_n - p_{n-1} \\
0 & 0 & 0 & \dots & 0 & \frac{1}{2} e_2
\end{bmatrix}, \quad (3.4)
$$

where $e_2 = [0 \ 1]$.

If there were a linear dependence relation among the rows of (3.4),

$$
\left(\omega_1 p_2 - \omega_2(p_3 - p_2), \omega_2(p_3 - p_2) - \omega_3(p_4 - p_3), \dots, \omega_{n-1}(p_n - p_{n-1}) + \frac{1}{2}\omega_n e_2\right) = (0, 0, \dots, 0)
$$

would hold for some scalars $\omega_1, \dots, \omega_n$. Working component-by-component from left to right, it would follow that $p_2, p_3 - p_2, p_4 - p_3, \dots, p_n - p_{n-1}$, and e_2 were all scalar multiples of one another, in which case the entire arm would lie along the y-axis. This would imply that $x_n = 0$, contradicting the fact that p is an element of U. Thus the rows of $DF(p)$ are independent, and $DF(p)$ has rank n, as desired. ■

4. CRITICAL POINTS OF THE ENDPOINT MAP. As noted earlier, the moduli space $\mathcal{M}(\vec{\ell})$ sits inside the space of arms W as the subspace with $p_n = (\ell_n, 0)$. Another way to get at this is to make use of the "endpoint map" $g\colon W \to \mathbf{R}$ defined by:

$$g(p_2, \dots, p_n) = x_n$$

(= the x-coordinate of p_n). That is, g records the position along the x-axis of the terminal point of an arm. Then $\mathcal{M}(\vec{\ell})$ is the level set $g^{-1}(\ell_n)$, so $\mathcal{M}(\vec{\ell})$ is a manifold whenever ℓ_n is a regular value of g.

Another reason to focus on regular and critical values comes from a basic principle of Morse theory. This says that, as far as level sets are concerned, the critical points and critical values are the only places where something interesting happens. To explain, suppose that we start with a large value of x and study what happens to $g^{-1}(x)$ as x decreases in a continuous way. If $x > \ell_1 + \cdots + \ell_{n-1}$, then $g^{-1}(x)$ is empty. If $x = \ell_1 + \cdots + \ell_{n-1}$ (= the maximum value of g), then $g^{-1}(x)$ is a single point, corresponding to the arm's being fully extended to the right. For x immediately below the maximum, $g^{-1}(x)$ turns out to be homeomorphic to an $(n-3)$-dimensional sphere, as we shall see. As x continues to decrease, the level sets may become more complicated, but Morse theory says that they remain diffeomorphic to one another except when x crosses through a critical value. In particular, between any two consecutive critical values, the level sets $g^{-1}(x)$ are all topologically the same. Right at a critical value, a singularity appears, which is what makes changes in the level sets possible. Even then, however, each change is localized near a critical point. Therefore, the type of a particular $g^{-1}(x)$ depends on the critical values through which a point passes on its way to x. We discuss the specific nature of the transitions at critical values later, but for now this may be enough motivation for going ahead and finding the critical points of g.

Definition. An arm $c = (c_2, c_3, \dots, c_n)$ is called a *straight line configuration* if all the c_i lie on the x-axis.

Proposition 2. *The critical points of g are precisely the straight line configurations.*

Proof. First, we show that every straight line configuration $c = (c_2, \dots, c_n)$ is a critical point. To do this, we choose local coordinates for W in a neighborhood of c and compute the Jacobian of g in terms of these coordinates. Since g is real-valued, being a critical point is the same as the Jacobian's having rank zero (i.e., being the zero matrix). Thus we want to show that all the partial derivatives of $g = x_n$ are zero at c.

To find local coordinates, we apply the implicit function theorem to the function F in (3.2) that defines W. Write $c_i = (a_i, 0)$. As in (3.3), $DF(c)$ can be written in the form:

$$\begin{bmatrix} 2a_2 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ -2(a_3 - a_2) & 0 & 2(a_3 - a_2) & 0 & \dots & 0 & 0 & 0 & 0 \\ & \vdots & & & \ddots & & \vdots & & \\ 0 & 0 & 0 & 0 & \dots & -2(a_n - a_{n-1}) & 0 & 2(a_n - a_{n-1}) & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}.$$

(This is the nonblock form, where the entries are individual partial derivatives.) The point is that columns $2, 4, \dots, 2(n-2)$ are all zero. Since $DF(c)$ has rank n, the remaining n columns must be linearly independent. These are the columns corresponding to the partial derivatives with respect to the variables $x_2, x_3, \ldots, x_{n-1}, x_n, y_n$. Now recall that W is the level set $F^{-1}(\ell_1^2, \ldots, \ell_{n-1}^2, 0)$. So the implicit function theorem says that, in a neighborhood of c within W, one can solve for $x_2, x_3, \ldots, x_{n-1}, x_n, y_n$ in terms of the other variables $y_2, y_3, \ldots, y_{n-1}$. In other words, $y_2, y_3, \ldots, y_{n-1}$ may be used as local coordinates to describe W near c. We must show that, at the point c, $D_j x_n = 0$ for $j = 2, \ldots, n-1$, where D_j denotes $\partial/\partial y_j$.

In fact, we have:

Lemma 3. *At c, $D_j x_i = 0$ for $i = 2, \ldots, n$ and $j = 2, \ldots, n-1$.*

Proof. Fix j, and proceed by induction on i. For any point in W, $x_2^2 + y_2^2 = \ell_1^2$, so differentiating with respect to y_j gives

$$2x_2\, D_j x_2 + 2y_2\, D_j y_2 = 0.$$

For the straight line configuration c, $x_2 = a_2 \neq 0$ and $y_2 = 0$. Hence $D_j x_2 = 0$.

Similarly, when $i > 2$ consider the $(i-1)$th edge. It satisfies

$$(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 = \ell_{i-1}^2,$$

which implies that

$$2(x_i - x_{i-1})(D_j x_i - D_j x_{i-1}) + 2(y_i - y_{i-1})(D_j y_i - D_j y_{i-1}) = 0.$$

Again, at the point c, $a_i \neq a_{i-1}$ and $y_i = y_{i-1} = 0$, leaving $D_j x_i = D_j x_{i-1}$. The latter is zero by induction. This proves the lemma. ■

Returning to the proof of Proposition 2, we now know that c is a critical point. Conversely, we must show that, if $p = (p_2, \ldots, p_n)$ in W is not a straight line configuration, then it is not a critical point. We do this by showing that x_n can be taken as one of the local coordinates for W in a neighborhood of p. It then follows that

$$\frac{\partial g}{\partial x_n} = \frac{\partial x_n}{\partial x_n} = 1 \neq 0$$

at p, hence that p is not a critical point.

To find local coordinates near p, we follow the same strategy as earlier; namely, we examine $DF(p)$ and look for n linearly independent columns. The variables corresponding to the columns *not* chosen may then be taken as local coordinates. In choosing independent columns, we thus wish to avoid the column corresponding to x_n, which is the next-to-last one.

Suppose that it is not possible to do so. Then removing the next-to-last column necessarily drops the rank of $DF(p)$ to $n-1$. The resulting matrix S has the form

$$S = 2\begin{bmatrix} p_2 & 0 & \cdots & 0 & 0 \\ -(p_3 - p_2) & p_3 - p_2 & \cdots & 0 & 0 \\ 0 & -(p_4 - p_3) & \cdots & 0 & 0 \\ & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & p_{n-1} - p_{n-2} & 0 \\ 0 & 0 & \cdots & -(p_n - p_{n-1}) & y_n - y_{n-1} \\ 0 & 0 & \cdots & 0 & \frac{1}{2} \end{bmatrix}.$$

This is the block form (3.4), except that the entries in the last column here are scalars rather than vectors in $\mathbf{R}^2$. By assumption, the n rows of S satisfy a nontrivial linear dependence relation:

$$\Big(\omega_1 p_2 - \omega_2(p_3 - p_2), \ldots, \omega_{n-2}(p_{n-1} - p_{n-2}) - \omega_{n-1}(p_n - p_{n-1}),$$

$$\omega_{n-1}(y_n - y_{n-1}) + \frac{1}{2}\omega_n\Big) = (0, \ldots, 0, 0)$$

for some scalars $\omega_1, \ldots, \omega_n$. If all of $\omega_1, \ldots, \omega_{n-1}$ are 0, then so is ω_n. Since the relation is nontrivial, this cannot be. Accordingly, at least one of $\omega_1, \ldots, \omega_{n-1}$ is nonzero. As in the proof of Proposition 1, it follows that $p_2, p_3 - p_2, \ldots, p_n - p_{n-1}$ are all scalar multiples of one another; in other words, p is a straight line configuration, contrary to assumption. The contradiction shows that, even with the x_n-column removed, $DF(p)$ still has rank n. Hence x_n can be taken as one of the local coordinates near p, completing the proof of Proposition 2. ∎

It is now immediate that, generically, our moduli spaces of polygons are manifolds.

Corollary 4. *Let $\mathcal{M}(\overrightarrow{\ell})$ be the moduli space of polygons of side lengths $\ell_1, \ldots, \ell_n$. Assume that $\mathcal{M}(\overrightarrow{\ell})$ is nonempty and contains no straight line configurations. Then $\mathcal{M}(\overrightarrow{\ell})$ is a compact smooth orientable manifold of dimension $n-3$.*

For instance, this formalizes our intuition from section 2 that a moduli space of quadrilaterals is one-dimensional.

Proof. The assumption says that $\mathcal{M}(\overrightarrow{\ell})$ contains no critical points of g. Since $\mathcal{M}(\overrightarrow{\ell}) = g^{-1}(\ell_n)$, this means that ℓ_n is a regular value. In light of the regular value theorem, $\mathcal{M}(\overrightarrow{\ell})$ is an orientable $(n-3)$-manifold. Moreover, $\mathcal{M}(\overrightarrow{\ell})$ is compact, since it is a closed bounded subset of $(\mathbf{R}^2)^{n-1} = \mathbf{R}^{2n-2}$: it is closed as the preimage of a point and bounded because every $p = (p_2, \ldots, p_n)$ in $\mathcal{M}(\overrightarrow{\ell})$ satisfies $|p_i| \le \ell_1 + \cdots + \ell_{n-1}$ for all i. ∎

We introduce some notation that will prove its value in what follows. Suppose that $c = (c_2, c_3, \ldots, c_n)$ is a straight line configuration, where $c_i = (a_i, 0)$. For convenience, set $c_1 = (0, 0)$. Each of the arm edges of c lies along the x-axis. We wish to distinguish between those that go to the right and those that go to the left.

Definition. Suppose that c is a straight line configuration. For $i = 1, 2, \ldots, n-1$ define

$$\epsilon_i = \begin{cases} +1 & \text{if } a_{i+1} > a_i, \\ -1 & \text{if } a_{i+1} < a_i. \end{cases}$$

(See Figure 6 for an example.)

In this notation, $a_{i+1} - a_i = \epsilon_i \ell_i$. Thus $g(c) = a_n = \sum_{i=1}^{n-1}(a_{i+1} - a_i) = \sum_{i=1}^{n-1} \epsilon_i \ell_i$. The nonzero numbers of this form are the critical values of g. In addition, $\mathcal{M}(\overrightarrow{\ell})$

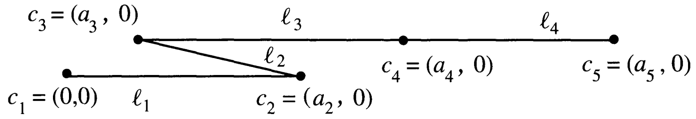

Figure 6. A straight line configuration: $\epsilon_1 = \epsilon_3 = \epsilon_4 = +1$, while $\epsilon_2 = -1$.

contains no straight line configurations if and only if ℓ_n is not of the form $\sum_{i=1}^{n-1} \epsilon_i \ell_i$ for any choices of $\epsilon_i = \pm 1$.

5. THE HESSIAN. We now introduce the other principle from Morse theory that we shall use. Let $f: U \to \mathbf{R}$ be a smooth real-valued function defined in an open subset U of $\mathbf{R}^m$, and let c be a critical point of f. In calculus, one typically learns, at least for functions of one or two variables, that information about the type of critical point (e.g., maximum, minimum, or saddle) is provided by the second-order derivatives. If f is a function of more than two variables, this generalizes in the following way.

Let $H(c)$ denote the $m \times m$ matrix whose (i, j)th entry is the second-order partial derivative $(D_{ij} f)(c)$. The matrix $H(c)$ is called the *Hessian* of f at c. Assume that $H(c)$ is nonsingular, in which event c is a *nondegenerate* critical point. A result known as the "Morse Lemma" states that, near c, f has a particularly simple canonical form; namely, it is a sum and difference of squares. More precisely, there exist local coordinates $u_1, \dots, u_m$ in a neighborhood of c with respect to which c corresponds to the origin and f has the form

$$f(u_1, \dots, u_m) = f(c) \pm u_1^2 \pm u_2^2 \pm \cdots \pm u_m^2. \tag{5.1}$$

The number of coefficients equal to -1 in this representation is called the *index* of c. As a result, near c, the level sets of f are easy to describe—they are quadric hypersurfaces—and their specific shape is determined by the index.

The Morse lemma is a differentiable version of the familiar linear theorem that every real symmetric matrix (or, equivalently, every real symmetric bilinear form) can be diagonalized. (The appropriate matrix and bilinear form here are the Hessian $H(c)$ and the quadratic terms in the Taylor approximation to f at c, respectively.) To review the algebra, let A be an $m \times m$ real symmetric matrix. Then there exists a nonsingular matrix P such that PAP^t is diagonal:

$$PAP^t = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m \end{bmatrix}. \tag{5.2}$$

(In fact, P can be taken to be an orthogonal matrix, in which case (5.2) is known as the spectral theorem, and the diagonal entries are the eigenvalues of A.) We call the number of d_i that are negative the *index* of A. The connection with (5.1) is that, for a nondegenerate critical point c, the index of c is equal to the index of $H(c)$ [**18**, pp. 5–8].

In practice, it is often simpler to work with $H(c)$ than to find a diagonalizing coordinate system $u_1, \dots, u_m$. Thus in this section we compute $H(c)$ for a critical point of the endpoint map $g = x_n$. In the next section we discuss how to find the index, and then we find it.

So suppose that $c = (c_2, \dots, c_n)$ is a straight line configuration. As in the proof of Proposition 2, $y_2, y_3, \dots, y_{n-1}$ may be used as local coordinates for W in a neighborhood of c. The Hessian $H(c)$ is then the $(n-2) \times (n-2)$ matrix whose entries are the various

$$(D_{ij}g)(c) = (D_{ij}x_n)(c) = \frac{\partial^2 x_n}{\partial y_j \partial y_i}(c),$$

where $2 \le i \le n-1$ and $2 \le j \le n-1$. (More precisely, this partial derivative is the entry in row $i-1$ and column $j-1$.) We determine these partials by direct calculation.

The main idea is to proceed inductively by implicitly differentiating the equation

$$(x_{k+1} - x_k)^2 + (y_{k+1} - y_k)^2 = \ell_k^2$$

for the kth edge of the arm, where as usual $(x_1, y_1) = (0, 0)$ is fixed. Differentiating first with respect to y_i, then with respect to y_j, gives

$$(x_{k+1} - x_k) \cdot (D_i x_{k+1} - D_i x_k) + (y_{k+1} - y_k) \cdot (D_i y_{k+1} - D_i y_k) = 0$$

and

$$\begin{aligned}(D_j x_{k+1} - D_j x_k) \cdot (D_i x_{k+1} - D_i x_k) + (x_{k+1} - x_k)(D_{ij} x_{k+1} - D_{ij} x_k) \\ + (D_j y_{k+1} - D_j y_k) \cdot (D_i y_{k+1} - D_i y_k) + (y_{k+1} - y_k)(D_{ij} y_{k+1} - D_{ij} y_k) = 0.\end{aligned} \tag{5.3}$$

Now $y_2, \dots, y_{n-1}$ are the independent variables, so $D_i y_k = 0$ unless $i = k$, whereas $D_{ij} y_k = 0$ always holds. Moreover, at the critical point c, $D_i x_k = 0$ by Lemma 3, and $x_{k+1} - x_k = a_{k+1} - a_k = \epsilon_k \ell_k$. At c, therefore, (5.3) simplifies to

$$\epsilon_k \ell_k (D_{ij} x_{k+1} - D_{ij} x_k) + (D_j y_{k+1} - D_j y_k) \cdot (D_i y_{k+1} - D_i y_k) = 0,$$

giving the inductive relation

$$D_{ij} x_{k+1} = -\frac{\epsilon_k}{\ell_k}(D_j y_{k+1} - D_j y_k) \cdot (D_i y_{k+1} - D_i y_k) + D_{ij} x_k. \tag{5.4}$$

(Note that $\epsilon_k = \pm 1$, so $\epsilon_k = 1/\epsilon_k$.)

Since the product on the right-hand side of (5.4) vanishes unless each of i and j is k or $k+1$, this splits up into essentially three cases, as follows:

(a) if $(i, j) = (k, k)$ or $(k+1, k+1)$, then $D_{ij} x_{k+1} = -(\epsilon_k / \ell_k) + D_{ij} x_k$;

(b) if $(i, j) = (k, k+1)$ or $(k+1, k)$, then $D_{ij} x_{k+1} = (\epsilon_k / \ell_k) + D_{ij} x_k$;

(c) if i or j is different from both k and $k+1$, then $D_{ij} x_{k+1} = D_{ij} x_k$.

Using (a)–(c), we can calculate $D_{ij} x_n$ by backtracking until we reach $D_{ij} x_1 = 0$. Again, there are three cases to consider:

Case 1: $j = i$.

$$\begin{aligned} D_{ii} x_n &= D_{ii} x_{n-1} = \cdots = D_{ii} x_{i+1} && \text{(by (c))} \\ &= -\frac{\epsilon_i}{\ell_i} + D_{ii} x_i && \text{(by (a))} \\ &= -\frac{\epsilon_i}{\ell_i} - \frac{\epsilon_{i-1}}{\ell_{i-1}} + D_{ii} x_{i-1} && \text{(by (a))} \end{aligned}$$

$$= -\frac{\epsilon_i}{\ell_i} - \frac{\epsilon_{i-1}}{\ell_{i-1}} + D_{ii}x_1 \qquad \text{(by (c))}$$
$$= -\frac{\epsilon_i}{\ell_i} - \frac{\epsilon_{i-1}}{\ell_{i-1}}.$$

Case 2: $j = i \pm 1$ (i.e., $|i - j| = 1$).

$$\begin{aligned} D_{i,i+1}x_n &= D_{i,i+1}x_{n-1} = \cdots = D_{i,i+1}x_{i+1} && \text{(by (c))} \\ &= \frac{\epsilon_i}{\ell_i} + D_{i,i+1}x_i && \text{(by (b))} \\ &= \frac{\epsilon_i}{\ell_i} + D_{i,i+1}x_1 && \text{(by (c))} \\ &= \frac{\epsilon_i}{\ell_i}, \end{aligned}$$

and likewise $D_{i,i-1}x_n = \epsilon_{i-1}/\ell_{i-1}$.

Case 3: $|i - j| \geq 2$.

$$D_{ij}x_n = D_{ij}x_{n-1} = \cdots = D_{ij}x_1 = 0 \qquad \text{(by (c))}.$$

This covers all the second-order partials. The end result is that $H(c)$, the Hessian of g at c, is the following $(n-2) \times (n-2)$ tridiagonal matrix:

$$\begin{bmatrix} -\frac{\epsilon_1}{\ell_1} - \frac{\epsilon_2}{\ell_2} & \frac{\epsilon_2}{\ell_2} & 0 & \cdots & 0 & 0 \\ \frac{\epsilon_2}{\ell_2} & -\frac{\epsilon_2}{\ell_2} - \frac{\epsilon_3}{\ell_3} & \frac{\epsilon_3}{\ell_3} & \cdots & 0 & 0 \\ 0 & \frac{\epsilon_3}{\ell_3} & -\frac{\epsilon_3}{\ell_3} - \frac{\epsilon_4}{\ell_4} & \cdots & 0 & 0 \\ & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{\epsilon_{n-3}}{\ell_{n-3}} - \frac{\epsilon_{n-2}}{\ell_{n-2}} & \frac{\epsilon_{n-2}}{\ell_{n-2}} \\ 0 & 0 & 0 & \cdots & \frac{\epsilon_{n-2}}{\ell_{n-2}} & -\frac{\epsilon_{n-2}}{\ell_{n-2}} - \frac{\epsilon_{n-1}}{\ell_{n-1}} \end{bmatrix}. \tag{5.5}$$

6. INDICES OF CRITICAL POINTS. To apply the Morse lemma (5.1) to the level sets of g (i.e., the moduli spaces $\mathcal{M}(\overrightarrow{\ell})$), we compute the index of the Hessian (5.5). As before, let $c = (c_2, \ldots, c_n)$ be a straight line configuration with $c_i = (a_i, 0)$, so $a_n = g(c) = \sum_{i=1}^{n-1} \epsilon_i \ell_i$ is the corresponding critical value.

Proposition 5. *The point c is a nondegenerate critical point of g. Indeed, let N denote the number of edges of c that go to the right (i.e., the number of ϵ_i among $\epsilon_1, \ldots, \epsilon_{n-1}$ that are $+1$). Then the index of c is given by the formula*

$$\operatorname{index}(c) = \begin{cases} N - 1 & \textit{if } a_n > 0, \\ N & \textit{if } a_n < 0. \end{cases}$$

Before proving this, we present some general preliminaries on how to compute the index of an $m \times m$ symmetric matrix A. Recall that the index is defined to be the number of negative diagonal entries d_i after A has been diagonalized as in (5.2).

If all the diagonal entries are positive (in which case the index is zero), A is called *positive definite*. There is a standard test for determining whether A fits into this category. Namely, if A_i denotes the $i \times i$ submatrix consisting of the first i rows and columns, then d_i is positive for all i if and only if $\det A_i$ is positive for all i. It turns out that the determinants of the A_i can be used for other values of the index as well, though this seems to be less familiar. Perhaps the simplest way to state it is as follows, under the assumption that the A_i are nonsingular.

Lemma 6. *Let A be an $m \times m$ real symmetric matrix such that A_i is nonsingular for all i. Then there exists a nonsingular matrix P such that the following hold:*

(a) *$Q = PAP^t$ is diagonal as in* (5.2)*;*

(b) *$d_i \neq 0$ for all i; and*

(c) *$\det A_i$ and $\det Q_i$ have the same sign for all i.*

The proof for the positive definite case still works, so we do not repeat it here (see, for example, Artin [**1**, pp. 242, 247]). ■

We comment at the end of this section about the case when some of the A_i are singular, but for now let us carry forward the assumption that they are not. Since $\det Q_i = d_1 \cdots d_i = d_i \det Q_{i-1}$, we see that d_i is negative if and only if $\det Q_i$ and $\det Q_{i-1}$ have opposite signs. By the lemma this is equivalent to $\det A_i$ and $\det A_{i-1}$ having opposite signs. In other words, we consider the sequence:

$$1, \det A_1, \det A_2, \ldots, \det A_m = \det A. \tag{6.1}$$

Corollary 7. *Under the assumption that A_i is nonsingular for all i, the index of the symmetric matrix A is the number of sign changes in the sequence* (6.1). ■

We will apply this to the Hessian $H(c)$ of our straight line configuration momentarily, but first we need to know something about the determinants of its submatrices. This is provided by the next calculation.

Proposition 8. *Let B denote the following $m \times m$ symmetric tridiagonal matrix:*

$$B = \begin{bmatrix} b_1 + b_2 & -b_2 & 0 & \ldots & 0 & 0 \\ -b_2 & b_2 + b_3 & -b_3 & \ldots & 0 & 0 \\ 0 & -b_3 & b_3 + b_4 & \ldots & 0 & 0 \\ & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & b_{m-1} + b_m & -b_m \\ 0 & 0 & 0 & \ldots & -b_m & b_m + b_{m+1} \end{bmatrix}.$$

Then

$$\det B = \sum_{i=1}^{m+1} b_1 \cdots \widehat{b_i} \cdots b_{m+1},$$

where "$\widehat{\ }$" means "omit."

Proof. We proceed by induction on m. The result is clear when $m = 1$ and $m = 2$. For $m > 2$, expand the determinant along the last column and use induction:

$$\det B = \left((b_m + b_{m+1}) \sum_{i=1}^{m} b_1 \cdots \widehat{b_i} \cdots b_m \right)$$
$$+ b_m \det \begin{bmatrix} b_1 + b_2 & -b_2 & \dots & 0 & 0 \\ -b_2 & b_2 + b_3 & \dots & 0 & 0 \\ & \vdots & \ddots & & \vdots \\ 0 & 0 & \dots & b_{m-2} + b_{m-1} & -b_{m-1} \\ 0 & 0 & \dots & 0 & -b_m \end{bmatrix}.$$

Then expand along the bottom row of the second term, again using induction, and simplify:

$$\begin{aligned} \det B &= \left((b_m + b_{m+1}) \sum_{i=1}^{m} b_1 \cdots \widehat{b_i} \cdots b_m \right) - b_m^2 \sum_{i=1}^{m-1} b_1 \cdots \widehat{b_i} \cdots b_{m-1} \\ &= b_m \sum_{i=1}^{m} b_1 \cdots \widehat{b_i} \cdots b_m + \sum_{i=1}^{m} b_1 \cdots \widehat{b_i} \cdots b_{m+1} - b_m \sum_{i=1}^{m-1} b_1 \cdots \widehat{b_i} \cdots b_m \\ &= b_1 \cdots b_m + \sum_{i=1}^{m} b_1 \cdots \widehat{b_i} \cdots b_{m+1} \\ &= \sum_{i=1}^{m+1} b_1 \cdots \widehat{b_i} \cdots b_{m+1}. \end{aligned}$$
■

This paves the way for computing the index of the Hessian (5.5).

Proof of Proposition 5. Let H_i denote the submatrix $H(c)_i$. For the time being, we continue to assume that all the H_i are nonsingular. Recall that the size of $H(c)$ is $(n-2) \times (n-2)$. We prove that the conclusion of Proposition 5 predicts the index of H_i correctly for $i = 1, \dots, n-2$. Proposition 5 itself is then the case $i = n - 2$.

We argue by induction on i, counting the number of sign changes in the sequence $1, \det H_1, \det H_2, \dots, \det H_i$. By Proposition 8 and (5.5) the determinants of these various submatrices are given by

$$\begin{aligned} \det H_j &= \sum_{k=1}^{j+1} \left(-\frac{\epsilon_1}{\ell_1} \right) \cdots \widehat{\left(-\frac{\epsilon_k}{\ell_k} \right)} \cdots \left(-\frac{\epsilon_{j+1}}{\ell_{j+1}} \right) \\ &= (-1)^j \left(\frac{\epsilon_1}{\ell_1} \right) \cdots \left(\frac{\epsilon_{j+1}}{\ell_{j+1}} \right) \sum_{k=1}^{j+1} \frac{\ell_k}{\epsilon_k} \\ &= (-1)^j \frac{\epsilon_1 \cdots \epsilon_{j+1}}{\ell_1 \cdots \ell_{j+1}} \sum_{k=1}^{j+1} \epsilon_k \ell_k \\ &= (-1)^j \frac{\epsilon_1 \cdots \epsilon_{j+1}}{\ell_1 \cdots \ell_{j+1}} a_{j+2}. \end{aligned} \tag{6.2}$$

(We have again made use of the fact that $\epsilon_k = 1/\epsilon_k$.)

If $i = 1$, then

$$\det H_1 = -\frac{\epsilon_1 \epsilon_2}{\ell_1 \ell_2} a_3.$$

Hence the 1×1 matrix H_1 has index 1 or 0 according as $\epsilon_1 \epsilon_2 a_3$ is positive or negative. If $a_3 > 0$, then either ϵ_1 and ϵ_2 are both $+1$, in which case $\mathrm{index}(H_1) = 1$ and $N = 2$, or one of ϵ_1, ϵ_2 is $+1$ and the other is -1, in which case $\mathrm{index}(H_1) = 0$ and $N = 1$. (The case $\epsilon_1 = \epsilon_2 = -1$ is impossible: both edges can't go the left when $a_3 > 0$.) Similarly, if $a_3 < 0$, either $\epsilon_1 = \epsilon_2 = -1$, so $\mathrm{index}(H_1) = N = 0$, or ϵ_1 and ϵ_2 have opposite signs and $\mathrm{index}(H_1) = N = 1$. The upshot is that the formula in Proposition 5 gives the correct index in each of these cases.

Assume inductively that the proposition is correct for H_{i-1}. The question of whether the sign changes from $\det H_{i-1}$ to $\det H_i$ depends on the recurrence relation between these determinants. By (6.2) this is:

$$\det H_i = -\frac{\epsilon_{i+1}}{\ell_{i+1}} \frac{a_{i+2}}{a_{i+1}} \det H_{i-1}. \tag{6.3}$$

Thus, in going from H_{i-1} to H_i, the index will either go up by 1 or stay the same according as $\epsilon_{i+1} a_{i+2}/a_{i+1}$ is positive or negative. Recall that ϵ_{i+1} records whether the edge from a_{i+1} to a_{i+2} goes to the right or left.

Again we break everything down into cases depending upon the signs of a_{i+1} and a_{i+2}:

Case 1: $a_{i+1} > 0$ and $a_{i+2} > 0$. In this case, by (6.3) the index goes up if and only if $\epsilon_{i+1} = +1$ (i.e., if and only if N increases by 1). Since the proposition is correct for H_{i-1}, it remains correct for H_i.

Case 2: $a_{i+1} < 0$ and $a_{i+2} < 0$. This is identical to Case 1.

Case 3: $a_{i+1} < 0$ and $a_{i+2} > 0$. Here ϵ_{i+1} must be $+1$, so by (6.3) the index does not change. If there were N edges going to the right before, there are $N + 1$ of them afterward, and the index in both cases is N, as it should be.

Case 4: $a_{i+1} > 0$ and $a_{i+2} < 0$. This is similar to Case 3, only now $\epsilon_{i+1} = -1$, the index goes up by 1, and the number of edges going to the right remains the same. If there are N such edges, the index would be equal to $N - 1$ before by induction, then would increase to N afterwards, again in accordance with Proposition 5.

Thus Proposition 5 is correct in all cases. This completes the proof, except for the assumption that H_i is nonsingular for all i. By (6.2) this is equivalent to assuming that $a_3, a_4, \dots, a_n$ are all nonzero. In fact, the terminal endpoint a_n is required to be nonzero by definition of an arm. Thus $H_{n-2} = H(c)$ is always nonsingular. In particular, this justifies the assertion that c is nondegenerate.

It is possible, however, that some of the intermediate a_i are zero. If this happens, at least two fixes are available. On the algebraic side, one can use the fact that, if some of the determinants in the sequence (6.1) are zero, but never two in a row, then the zero terms can be omitted and the number of sign changes still equals the index. This is stated in Gantmacher [**6**, pp. 303–304], where it is attributed to Gundenfinger. The drawback to this approach is that one can no longer simply cite the standard proof for positive definiteness. In any case, for a straight line configuration c the edges of the arm have nonzero length, so no two consecutive a_i are ever zero. Hence the modified criterion for computing the index applies. After further bookkeeping (i.e., more case-by-case analysis), one can check that the previous arguments still go through.

Perhaps a more appealing alternative is that W is invariant, up to diffeomorphism, under permutations of the prescribed lengths. Specifically, any point $p =$

$(p_2, p_3, \dots, p_n)$ of W determines edge vectors $v_1 = p_2$, $v_2 = p_3 - p_2, \dots, v_{n-1} = p_n - p_{n-1}$. But conversely, these vectors determine p as well, since $p_i = \sum_{j=1}^{i-1} v_j$ for all i. Thus W may be described as the set of all $(v_1, v_2, \dots, v_{n-1})$ in $(\mathbf{R}^2)^{n-1}$ such that $|v_1| = \ell_1, |v_2| = \ell_2, \dots, |v_{n-1}| = \ell_{n-1}$, and $\sum_{j=1}^{n-1} v_j$ is a nonzero vector along the x-axis. This last condition is preserved under permutations of the indices.

In other words, suppose that σ is a permutation of $1, 2, \dots, n-1$, and consider the side lengths $\ell_{\sigma(1)}, \dots, \ell_{\sigma(n-1)}$ in that order. If W' denotes the space of arms associated with this sequence, then permutation of the v_j gives a diffeomorphism $\widetilde{\sigma}: W \to W'$. Moreover, W' has its own endpoint map $g': W' \to \mathbf{R}$, which is invariant under $\widetilde{\sigma}$ (i.e., $g = g' \circ \widetilde{\sigma}$). It follows that any critical point c of g corresponds under $\widetilde{\sigma}$ to a critical point c' of g' having the same index.

Now let $c = (c_2, \dots, c_n)$ be a critical point of g, possibly with some c_i with $i \neq n$ at the origin. If $a_n > 0$, then there is a net displacement to the right along the arm. Permute the edges so that all the ones to the right come first, followed by those going to the left. If $a_n < 0$, put all the edges to the left first. In this way, we obtain a new critical point c' in W', none of whose intermediate vertices c_i' are at the origin. By the arguments already given, Proposition 5 is valid for c' and, since permutations affect neither the index nor the number of edges pointing to the right, it is valid for c as well. ∎

7. SPACES OF PENTAGONS. We apply the preceding results to study a particular class of polygons, namely, pentagons. The reason for singling out this case is that, as long as it contains no straight line configurations, a moduli space of pentagons is a compact orientable manifold of dimension two (i.e., a surface). This follows from Corollary 4. Since there is a classification theorem for surfaces, the Morse-theoretic information that we have presented can be interpreted in precise topological ways. (For polygons with more than five sides, the Morse theory approach is still useful, but more sophisticated tools from topology are needed to interpret the information.)

Thus let five positive numbers $\ell_1, \ell_2, \dots, \ell_5$ be given, and assume that ℓ_5 is not of the form $\sum_{i=1}^{4} \epsilon_i \ell_i$ for any choice of $\epsilon_i = \pm 1$. Let W be the associated space of arms, and let $g: W \to \mathbf{R}$ denote the endpoint map. Then W is a 3-dimensional manifold, ℓ_5 is a regular value of g, and $\mathcal{M}(\vec{\ell}) = g^{-1}(\ell_5)$.

As discussed in section 4, the strategy is to examine the level sets $g^{-1}(x)$ as x decreases from some large value to ℓ_5. For this, we may assume that x is always positive. The topological type of the level sets can change only when x crosses through a critical value. So suppose that y is a critical value, say $y = g(c) = \sum_{i=1}^{4} \epsilon_i \ell_i$ for some critical point $c = (c_2, \dots, c_5)$ in W. By the Morse lemma (5.1) there are local coordinates u_1, u_2, and u_3 in a neighborhood of c in which c is the origin, g has the form

$$g = y \pm u_1^2 \pm u_2^2 \pm u_3^2,$$

and the number of negative squares is the index. Hence, for values x near y, the level set $g^{-1}(x)$ looks like $y \pm u_1^2 \pm u_2^2 \pm u_3^2 = x$, or

$$\pm u_1^2 \pm u_2^2 \pm u_3^2 = x - y, \tag{7.1}$$

locally near c.

For instance, suppose that c has index 3, so that $g = y - u_1^2 - u_2^2 - u_3^2$ near c. Clearly, y is a local maximum. For this choice of signs, the level set (7.1) is empty when $x > y$, a point when $x = y$, and a sphere when $x < y$. In other words, in descending through a critical value of index 3 the change that occurs in the level sets is that a sphere appears. Similarly, a critical point of index 0 signifies a local minimum, and crossing through the associated critical value corresponds to a sphere disappearing.

Suppose that c has index 2. Then the level sets $-u_1^2 - u_2^2 + u_3^2 = x - y$ are quadric surfaces of the type shown in Figure 7. Of course, this figure displays what the level sets look like only within a neighborhood of c inside W. Away from the critical point, however, the level sets remain essentially unchanged. Thus a more global view of how the level sets are changing might look something like the picture in Figure 8. Referring to this picture, one says that, in going from left to right, a handle has been attached.

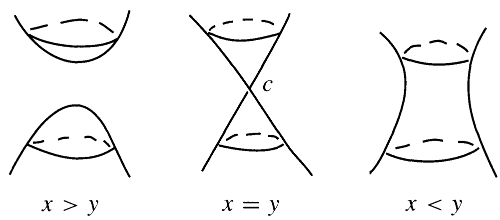

Figure 7. The level sets $-u_1^2 - u_2^2 + u_3^2 = x - y$.

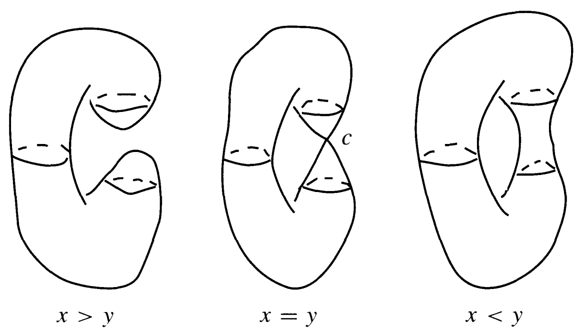

Figure 8. The level sets $g^{-1}(x)$ as x passes through a critical value of index 2.

By similar reasoning, passing through a critical value of index 1 corresponds to removing a handle. These results are summarized in Figure 9.

Index of critical point	Effect on level set
3	Sphere appears.
2	Attach a handle.
1	Remove a handle.
0	Sphere disappears.

Figure 9. Effects of crossing through a critical value.

The various critical values $g(c) = \sum_{i=1}^{4} \epsilon_i \ell_i$ are easily described in terms of $\ell_1, \ldots, \ell_4$. Let $T = \sum_{i=1}^{4} \ell_i$. By Proposition 5 a critical point of index 3 with positive critical value must have all four ϵ_i equal to $+1$. The corresponding critical value is T. A critical point of index 2 has three ϵ_i equal to $+1$ and the remaining one equal to -1.

The corresponding critical value has the form $(\ell_1 + \cdots + \ell_4) - 2\ell_i = T - 2\ell_i$ for some i. Similarly, critical points of index 1 have values of the form $T - 2(\ell_i + \ell_j)$ for distinct i and j, and those of index 0 have values of the form $T - 2(\ell_i + \ell_j + \ell_k)$ for distinct i, j, and k.

The topology of a given $\mathcal{M}(\overrightarrow{\ell})$ depends on where ℓ_5 is in relation to these critical values. By keeping track of the critical values that go by as x decreases to ℓ_5, one obtains a description of $\mathcal{M}(\overrightarrow{\ell})$ in terms of attaching a certain number of handles to spheres. This would determine $\mathcal{M}(\overrightarrow{\ell})$ completely, except for one detail: moduli spaces need not be connected. Thus, for example, attaching a handle might mean increasing the number of holes in the surface by one, or it might mean joining two previously disconnected components. Figure 10 provides an illustration.

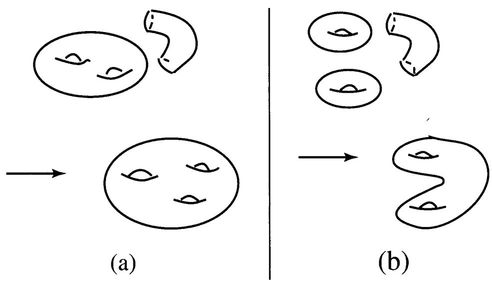

Figure 10. Two situations for attaching a handle.

Fortunately, there is a simple criterion for deciding when a moduli space of polygons is connected that depends on the relation between the second and third longest sides and the total perimeter. An elementary ad hoc proof can be given in the special case of pentagons by adapting the ideas of Walker [**21**, chap. 2], but the techniques are unrelated to what we have been doing. So instead we state without proof the result in the general case, which is due to Kapovich and Millson [**11**, p. 431] and Lenhart and Whitesides [**16**].

Proposition 9. *Arrange the n terms $\ell_1, \dots, \ell_n$ in nonincreasing order (i.e., from largest to smallest). Choose the second and third terms in this list; call them b and c, respectively. Let $P = \sum_{i=1}^{n} \ell_i$. Then the moduli space $\mathcal{M}(\overrightarrow{\ell})$ is connected if and only if $b + c \leq P/2$.* ■

For completeness, we note that when $\mathcal{M}(\overrightarrow{\ell})$ is disconnected it has exactly two components and that the elements of one are precisely the reflections in the x-axis of the elements of the other. Also, as stated, Proposition 9 allows the degenerate possibility that $\mathcal{M}(\overrightarrow{\ell})$ is connected because it is empty. There is a criterion that addresses this issue; namely, if a denotes the largest number among $\ell_1, \dots, \ell_n$, then $\mathcal{M}(\overrightarrow{\ell})$ is nonempty if and only if $a \leq P/2$ (again see [**11**] or [**16**]).

To observe the preceding ideas in action, consider pentagons with $\ell_1 = 9$, $\ell_2 = 8$, $\ell_3 = 4$, and $\ell_4 = 3$ for various values of ℓ_5. For the associated endpoint map g, we have $T = 24$. This is the maximum value of g, corresponding to a critical point of index 3. Following the process described earlier for finding critical values, the positive critical values of index 2 are $y = 18$, 16, 8, and 6, while those of index 1 are $y = 10$ and 2. There are no positive critical values of index 0. So, for instance, when ℓ_5 first decreases below 24, the moduli space becomes homeomorphic to a sphere. It stays that way until ℓ_5 drops past 18, at which point a handle is attached (i.e., it turns into a torus). Continuing in this fashion and making use of Figure 9 and Proposition 9, one obtains the evolution in the moduli spaces portrayed in Figure 11(a).[2] (To elaborate a bit further on the case $8 < \ell_5 < 10$, we can repeat an argument given in section 2 to see why removing a handle when ℓ_5 drops below 10 leaves the two components shown. Namely, if $8 < \ell_5 < 10$, vertex p_2 of the pentagon cannot lie on the x-axis because of the triangle inequality. Thus those pentagons with p_2 above the x-axis and those with p_2 below the x-axis lie in different components, and reflection in the x-axis maps each component homeomorphically onto the other. Alternatively, this is predicted by the comments following Proposition 9.)

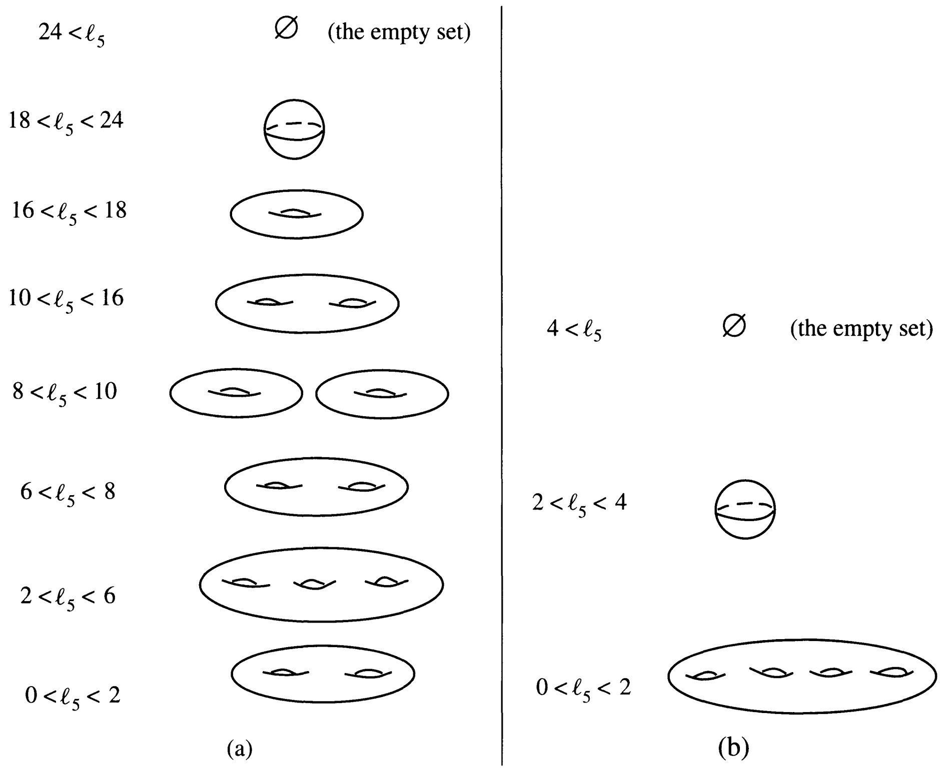

Figure 11. Two families of moduli spaces of pentagons for various ℓ_5: (a) $\ell_1 = 9$, $\ell_2 = 8$, $\ell_3 = 4$, $\ell_4 = 3$; (b) $\ell_1 = \ell_2 = \ell_3 = \ell_4 = 1$.

If information of this type fails to help the reader get through the day, perhaps an example that comes to mind more naturally is the case of equilateral pentagons, say with $\ell_1 = \cdots = \ell_5 = 1$. Here, $T = 4$, a critical value of index 3, as usual. The only other positive critical value is $y = 2$, which has index 2. This comes from four distinct critical points, one for each choice of which three of the first four edges point to the right. In order to reach $\ell_5 = 1$, one must pass through all these critical levels. Furthermore, the moduli space is connected ($1 + 1 \leq 5/2$). Thus the moduli space of equilateral pentagons is a sphere with four handles attached (i.e., a four-holed torus) as shown in Figure 11(b). This reproduces a result that seems to have appeared in print first in the work of Havel [**9**], where Morse theory is applied to the critical points of an oriented area function.

[2]For a different method of obtaining results like these, see Thurston and Weeks [**20**, pp. 113–116] or Walker [**21**, chap. 2].

For pentagons more generally, the techniques described here yield at most four critical points of index 2. As a result, a moduli space of pentagons can be a torus having at most four holes. According to the universality theorem of Kapovich and Millson, a torus with more holes can still be realized as the moduli space of some linkage; however, this linkage will not be polygonal. For a discussion of how to obtain such higher genus surfaces, see Jordan and Steiner [**10**].

ACKNOWLEDGMENTS. We thank Thomas Hunter for his perceptive comments on a draft of this article, Helene Shapiro for providing the reference to the book by Gantmacher, and the referees for suggesting improvements to the exposition. Also, Catherine thanks Swarthmore College for providing financial support during the summer in which this work was carried out, and Don thanks Connie Hungerford for nudging him to do something, possibly involving this material.

REFERENCES

1. M. Artin, *Algebra*, Prentice-Hall, Upper Saddle River, NJ, 1991.
2. L. Asimow and B. Roth, The rigidity of graphs, *Trans. Amer. Math. Soc.* **245** (1978) 279–289.
3. A. Browder, *Mathematical Analysis: An Introduction*, Springer-Verlag, New York, 1996.
4. R. Connelly and E. D. Demaine, Geometry and topology of polygonal linkages, in *Handbook of Discrete and Computational Geometry*, 2nd ed., J. E. Goodman and J. O'Rourke, eds., Chapman & Hall/CRC, Boca Raton, FL, 2004, pp. 197–218.
5. R. Connelly, E. Demaine, and G. Rote, Straightening polygonal arcs and convexifying polygonal cycles, *Discrete Comput. Geom.* **30** (2003) 205–239.
6. F. R. Gantmacher, *The Theory of Matrices*, vol. 1, Chelsea, New York, 1960.
7. J.-C. Hausmann, Sur la topologie des bras articules, in *Algebraic Topology, Poznán, 1989*, Lecture Notes in Mathematics, no. 1474, Springer-Verlag, Berlin, 1991, pp. 146–159.
8. J.-C. Hausmann and A. Knutson, The cohomology ring of polygon spaces, *Ann. Inst. Fourier (Grenoble)* **48** (1998) 281–321.
9. T. F. Havel, Some examples of the use of distances as coordinates for Euclidean geometry, *J. Symbolic Comput.* **11** (1991) 579–593.
10. D. Jordan and M. Steiner, Compact surfaces as configuration spaces of mechanical linkages, *Israel J. Math.* **122** (2001) 175–187.
11. M. Kapovich and J. Millson, On the moduli space of polygons in the Euclidean plane, *J. Differential Geom.* **42** (1995) 430–464.
12. ———, The symplectic geometry of polygons in Euclidean space, *J. Differential Geom.* **44** (1996) 479–513.
13. ———, Universality theorems for configuration spaces of planar linkages, *Topology* **41** (2002) 1051–1107.
14. H. C. King, Planar linkages and algebraic sets, in Proceedings of 6th Gökova Geometry-Topology Conference, *Turkish J. Math.* **23** (1999) 33–56.
15. P. A. Kirk and E. P. Klassen, Representation spaces of Seifert fibered homology spheres, *Topology* **30** (1991) 77–95.
16. W. J. Lenhart and S. H. Whitesides, Reconfiguring closed polygonal chains in Euclidean d-space, *Discrete Comput. Geom.* **13** (1995) 123–140.
17. R. J. Milgram and J. C. Trinkle, The geometry of configuration spaces for closed chains in two and three dimensions, *Homology Homotopy Appl.* **6** (2004) 237–267.
18. J. Milnor, *Morse Theory*, Princeton University Press, Princeton, 1969.
19. J. R. Munkres, *Analysis on Manifolds*, Westview, Boulder, CO, 1991.
20. W. P. Thurston and J. R. Weeks, The mathematics of three–dimensional manifolds, *Scientific American* **251** (July, 1984) 108–120.
21. K. Walker, *Configuration Spaces of Linkages*, undergraduate thesis, Princeton University, 1985.

DON SHIMAMOTO is on the faculty at Swarthmore College. He grew up in California: born in Berkeley, elementary and secondary education in El Cerrito and Richmond, and college at Stanford. Moving east, he did his graduate work at Brandeis, where his dissertation advisor was Ed Brown, then held short-term positions at Bowdoin and Haverford before starting at Swarthmore. His main research interests have been in algebraic

and differential topology, though he is quite taken by the many connections of linkages to other areas of mathematics and hopes to continue following where they lead.
Department of Mathematics and Statistics, Swarthmore College, 500 College Ave., Swarthmore, PA 19081
dshimam1@swarthmore.edu

CATHERINE VANDERWAART studied linkages during the summer after her junior year at Swarthmore College, where she graduated with High Honors in mathematics and philosophy in 2003. Since then, she has been living in Philadelphia and working at a brokerage firm, where her mathematics background has aided her in analyzing sales figures and market data.
Department of Mathematics and Statistics, Swarthmore College, 500 College Ave., Swarthmore, PA 19081
cvanderwaart@alum.swarthmore.edu

Four Limes

To color a globe to perfection
That passes a rigid inspection,
    However you choose
    You need four different hues
And a circular sense of direction.

They say that space ends in a curve,
And whatever goes straight there will swerve.
    It's mainly a menace
    To those who play tennis,
Since no one can hit back a serve.

Prodigious divisions of fractions—
Though better than horrid subtractions—
    For you, if you fidget
    When facing a digit,
They're not one of life's main attractions.

The dx/dy of a sine
Equals plus a cosine—that's fine.
    But dx/dy
    Of a cosine is sly,
With a sign that appears out of line.

——Submitted by Bob Scher, Mill Valley, California
(By the author's definition a "lime" is a "clean limerick.")