Automatic Geometric Theorem Proving

Sam Ratliff

19 March 2021

1 Automatic Geometric Theorem Proving with Gröbner Bases

Many geometric proofs are done without reference to a coordinate system. To some extent, an explicit coordinate system is not necessary in many elementary geometric applications. However, by introducing a coordinate plane, we can more easily represent geometric objects and statements through equations. By doing so, we can reduce geometric problems into polynomial systems of equations. Thus, Gröbner bases appear to be uniquely well-suited to the task of automatic geometric theorem proving.

1.1 Notation and Intuition

In any geometric construction, some decisions are arbitrary, like where we place our origin or how long we draw our first line. However, later in the construction, other decisions will depend on choices we made earlier. After all, if you're drawing a triangle and you've already drawn two sides of it, you're rather limited in the options that you have to draw the last side. In order to reflect this distinction in our proofs, we will use the variables $u_1, \ldots, u_m$ to indicate a variable that has no explicit restriction on its value and we will use the variables $x_1, \ldots, x_n$ to indicate a dependant variable that has a value that is defined by the values of the variables $u_1, \ldots, u_m$.

As an example, consider four equally spaced points on the x-axis of the Cartesian plane. Then, two of these points trisect the line segment created by the first two. Call these points A, B, M_1 and M_2. For the sake of convenience, take $A = (0, 0)$. Then, take $B = (u_1, 0)$. Now, the locations of

M_1 and M_2 have been determined by the selection of B. Thus, we denote M_1 as $(x_1, 0)$ and M_2 as $(x_2, 0)$.

Based on our construction, we know these points are equally spaced. Thus, $AM_1 = M_1M_2 = M_2B$. We can express these relations by writing $x_1 - 0 = x_2 - x_1 = u_1 - x_2$, which, after rearranging, yields the equations below.

$$\begin{aligned} h_1 &= x_2 - 2x_1 = 0 \\ h_2 &= u_1 - 2x_2 + x_1 = 0 \\ h_3 &= u_1 - x_2 - x_1 = 0 \end{aligned}$$

The relations above do not explicitly relate the length of one of the smaller segments like $\overline{AM_1}$ to the length of $\overline{AB}$. However, we can intuitively guess that $\overline{AM_1}$ is one third of the length of $\overline{AB}$.

We can encode this relationship as a polynomial. First we write an equation in terms of the lengths of the line segments, which yields $3 \cdot AM_1 = AB$. When we rewrite this as a polynomial, we get the equation $3(x_1 - 0) = u_1 - 0$. By rearranging, we get the polynomial $g = u_1 - 3x_1 = 0$. Notice that

$$\begin{aligned} h_1 + h_3 &= (u_1 - x_2 - x_1) + (x_2 - x_1) \\ &= u_1 - 3x_1 \\ &= g \end{aligned}$$

Therefore, g is a linear combination of our hypotheses, so if $h_1 = h_2 = h_3 = 0$, then $g = 0$ also holds. Moreover, since ideals are closed under addition, it follows that $g \in \langle h_1, h_2, h_3 \rangle$.

When $g \in \mathbf{I}(V)$, where $V = \mathbf{V}(h_1, h_2, h_3)$, we say that g **follows strictly** from the hypothesis set. We can easily check whether a polynomial is a member of an ideal by dividing it by the Gröbner basis of the ideal. This is a much more reliable method than eyeballing our equations and figuring out whether g can be written as a linear combination of our hypotheses.

1.2 Translating Theorems into Polynomials

As shown in the example above, we can translate many geometric statements into polynomials. This is because many geometric statements are asserting some form of equivalence between two quantities. In the example above, we equated the distances between points in order to represent two line segments

having the same length. Similarly, we can equate slopes of lines to indicate that they are parallel and we can use mathematical objects like dot products and cross products to measure angles between line segments. More complicated geometric statements, like those relating to collinearity, midpoints, circles, and other common geometric constructs, can be synthesized by modifying and combining the previous definitions. For example, three collinear points A, B, and M, where M is between A and B, can be expressed by the relation $AB = AM + MB$, which is a modification of the critereon for two lines having the same length.

1.3 Fun with Parallelograms and Irreducible Varieties

In `6.4.11.m2`, I follow the basic structure of the proof of Example 1 through page 326. I encourange the reader to follow along in the code as they read this section. The proof begins by establishing our hypotheses and conclusions. However, the book and my code both quickly find that g does not strictly follow from our hypotheses. The reason that this occurs is because the variety associated with the ideal of hypotheses is reducible. We can readily see this by factoring the generators of our Gröbner basis.

In order to reduce our Gröbner basis, we factor all of the generators. Then, we compute two new Gröbner bases, where each new Gröbner basis uses all of the same generators as the previous one, except it replaces a factorable generator by one of its factors. The first generator that I factored was f_2, which can be factored into $u_3(x_1 - u_1 - u_2)$. To reduce this basis, we compute new Gröbner bases of the ideals $\langle f_1, x_1 - u_1 - u2, f_3, f_4, f_5, f_6\rangle$ and $\langle f_1, u_3, f_3, f_4, f_5, f_6\rangle$. This yields more generators, which we can factor again. We repeat this process until our generators are not factorable. We can select a set of Gröbner bases from our final collection such that the varieties that they are associated with are the irreducible components of V. This will make our subsequent calculations easier while simultaneously explaining why g did not follow strictly from our hypotheses.

I eventually found that V was composed of four irreducible varieties, V', U_1, U_2, and U_3. Notice that the varieties U_1, U_2, and U_3 all have some polynomial in their generators that depends only on a u variable. Consider U_1. The Gröbner basis of U_1 is $\{x_2, x_4, u_3\}$, which corresponds to the set of equations $x_2 = x_4 = u_3 = 0$. By referring back to the definitions of these variables on pages 319 and 320, we can see that solutions that are in U_1 correspond to the "parallelogram" in which the point C is on the x-axis. Likewise, this forces

the points D and N to be on the x-axis as well. There's nothing mathematically wrong with a solution like this. However, the solutions that it finds do not lead to a parallelogram, which makes it rather difficult to say anything meaningful about the diagonals of said parallelogram.

If we recall our distinction between the u variables and the x variables, this conclusion makes sense. When setting up our geometric system, we specifically chose values our u variables that produce a system we want. Then, the values of our x variables were determined completely by our selection of u variables. Therefore, if we are not careful about our selection of u variables, we may end up with solutions that fail to have meaning about our system. After all, the computer sees all of the variables as essentially identical, regardless of whether they were named with a u or an x. It's up to the programmer to ensure that the results have the expected meaning.

1.4 Degenerate Solutions

We call solutions like those found in U_1 **degenerate**. If we want to find solutions that correspond to the system that we are interested in, we have to exclude degenerate solutions. In the case of U_1, we noticed that u_3 was one of our generators. If any of our generators are composed wholly of u variables, then some of the solutions in that variety are related to an incorrectly set up system. In our case, this improper system will likely be a line. In other systems, a degenerate solution will reflect some other loss of information encoded in the hypotheses.

To avoid degenerate solutions, we look for irreducible varieties that have no generators that depend only on u variables. We say that $u_1, \ldots u_m$ are algebraically independant on an irreducible variety if this is true. In our example, we can see that the only irreducible variety of V that the u variables are algebraically independant over is the variety V'. Thus, we will not have to worry about degenerate solutions appearing on V'. As such, we can use the generators of the Gröbner basis of V' to determine whether g follows from the hypotheses.

If $g \in \mathbf{I}(V')$, where the u variables are algebraically independant over V', we say that g **follows generically** from the hypotheses. Note that the difference between following strictly and following generically from the hypotheses is that if $g \in \mathbf{I}(V)$, it follows strictly, while if $g \in \mathbf{I}(V')$, it follows generically. Since $V' \subseteq V$, it is a stronger condition to state that g follows generically from the hypotheses.

1.5 Optimizations

1.5.1 Better Hypotheses

The hypotheses in this example come in pairs. The first two hypotheses describe the location of the point D, while the second pair describe the point N. The book presents a second set of hypotheses that replace the hypotheses that describe the location of D. Instead of using slopes to locate D, the book instead looks at the diagram and notes that $(u_1 + u_2, u_3) = (x_1, x_2)$. This yields the two hypotheses $h_1' = x_1 - u_1 - u_2 = 0$ and $h_2' = x_2 - u_3 = 0$. These two hypotheses are much simpler than the hypotheses that we used initially. In particular, h_2' does not have a multiplication in it. This simplifies the process of reducing the variety into irreducible components because our reductions required us to factor the generators. By choosing better hypotheses, we avoid having to factor and reduce a variety at least once. In fact, I only had to reduce the variety generated by these new hypotheses twice in order to find V', whereas I had to reduce the variety five times with the orignal set of hypotheses. Thus, picking a good hypothesis set can potentially halve the amount of work that has to be done.

1.5.2 Avoiding Decomposition

There are also techniques that we can use to determine if g is in the algebraically independant region of V without decomposing V into its irreducible components.

Proposition 1. *The conclusion g follows generically from the hypotheses $h_1, \ldots, h_n$ whenever there is some nonzero polynomial $c(u_1, \ldots, u_m) \in \mathbb{R}[u_1, \ldots, u_m]$ such that*

$$c \cdot g \in \sqrt{H},$$

where H is the ideal generated by the set of hypotheses.

The proof for this is rather intuitive. I present an informal rendering of it below.

We know that $c \cdot g \in \sqrt{H}$, so it must hold that $c \cdot g = 0$ at all points on the variety V. Therefore, since $V' \subseteq V$, $c \cdot g = 0$ on V'. We know that the u variables are algebraically independant over V'. Since c depends only on u variables, it must hold that $c \neq 0$ on V'. In order for $c \cdot g = 0$ to hold on V', it must be true that $g = 0$ on V'. Thus, g follows generically from the hypotheses.

There are two other equivalent characterizations of g that we can use in order to determine whether g follows generically from the set of hypotheses. One nice feature of these alternate characterizations is that they do not rely on determining the existance of a polynomial c.

Corollary 2. *The following statements are equivalent:*

1. *There is a nonzero polynomial $c \in \mathbb{R}[u_1, \ldots, u_m]$ such that $c \cdot g \in \sqrt{H}$.*
2. *$g \in \sqrt{\tilde{H}}$, where $\tilde{H}$ is the ideal generated by the hypotheses in $\mathbb{R}(u_1, \ldots, u_m)[x_1, \ldots, x_n]$.*
3. *$\{1\}$ is the reduced Gröbner basis of the ideal*
$$\langle h_1, \ldots, h_n, 1 - yg \rangle \subseteq \mathbb{R}(u_1, \ldots, u_m)[x_1, \ldots, x_n, y].$$

Notice that (3) provides us with a straightforward computational way to determine whether g follows from the hypothesis set. We simply compute the reduced Gröbner basis of the ideal $\langle h_1, \ldots, h_n, 1-yg \rangle \subseteq \mathbb{R}(u_1, \ldots, u_m)[x_1, \ldots, x_n, y]$ and check whether it is the set $\{1\}$. If it is, then (1) is equivalently true, so the proposition above holds and g follows generically from the hypotheses.

This new method illuminates two problems. First, since we do not decompose the variety into irreducible components in this method, we learn very little about the variety itself. In particular, we don't have any information about the degenerate solutions that are present in the variety. The polynomial c in the proposition above contains information about the degenerate solutions. Intuitively, this makes sense, as g contains information related to the algebraically independant solutions. Since (3) of the corollary makes no mention of c, though, we need to do additional computations to determine c if we want information about the degenerate cases.

Secondly, it uncovers a major limitation of the Gröbner basis method of theorem proving. By extending the Gröbner basis method to the complex plane, we can show that g follows generically from a set of hypotheses in $\mathbb{R}[u_1, \ldots, u_m, x_1, \ldots, x_n]$ if and only if it also follows generically from the same hypotheses in $\mathbb{C}[u_1, \ldots, u_m, x_1, \ldots, x_n]$. Hence, the Gröbner basis method proves theorems that are true in the complex plane. As such, this method is unable to prove theorems that are true in $\mathbb{R}$ but false in $\mathbb{C}$.

This issue arises because $\mathbb{R}$ is not algebraically closed. In Chapter 4, the concept of a radical ideal was often strongly intertwined with the concept of an algebraically closed field. Thus, it is not surprising that the method proposed by (3) of the corollary is restricted to algebraically closed fields, as (1) and (2) of the corollary both make use of radical ideals.

2 Automatic Geometric Theorem Proving with Wu's Method

Wu's method is a method of automatic geometric theorem proving that predates the Gröbner basis method discussed above. It uses a method similar to row reduction in matricies to produce a modified version of the hypothesis set that has useful properties that allow us to quickly test a conclusion. Much like row reduction, we have a rule for how we can combine two entries in our set, which we call pseudodivision.

2.1 Pseudodivision

Pseudodivision behaves similar to one variable polynomial division, even though it happens on multivariate rings. Let f and g be polynomials in the field $k[x_1, \ldots, x_n, y]$. First, we select a variable y to pseudodivide with respect to. Then, we rewrite f and g in the form

$$f = c_p y^p + \cdots + c_1 y + c_0$$
$$g = d_m y^m + \cdots + d_1 y + d_0$$

This allows us to isolate y in each monomial. It is important to note that in normal one variable polynomial division, the coefficients of y are in the field k. In polynomial pseudodivision, however, the coefficients of y are polynomials in $k[x_1, \ldots, x_n]$.

We also need to add an additional condition to single variable polynomial division for pseudodivision to be useful. To ensure that the pseudodivision works, we multiply f by a power of d_m. We do this to ensure that the leading term of $d_m f$, the polynomial that we are dividing, is divisible by the leading term of g, the polynomial that we are dividing by. In other words, we do not know if c_p is divisible by d_m, but we do know that there exists some $s \geq 0$ such that $d_m^s \cdot c_p$ is divisible by d_m. While $s = 1$ will easily satisfy this relationship for the first leading terms of f and g, we may need to pick a larger s for terms later in the polynomial.

We can also think about the multiple of d_m^s in another way. Instead of multiplying by a power of d_m before dividing, we use one variable polynomial division with respect ot y to divide f by g and we don't worry about whether c_p is divisible by d_m. This means that some of the coefficients of

our pseudoquotient and pseudoremainder may end up being rational polynomials. To rectify this, we multiply both sides by the smallest power of d_m that will cancel the denominators that still remain in the pseudoquotient and pseudoremainder.

Proposition 3 (Pseudodivision)**.** *Let $f, g \in k[x_1, \ldots, x_n, y]$, where $deg(g, y) \leq deg(f, y)$. Then, in order to isolate y, rewrite f and g as follows:*

$$f = c_p y^p + \cdots + c_1 y + c_0$$
$$g = d_m y^m + \cdots + d_1 y + d_0$$

Then, there exists an equation of the form

$$d_m^s f = qg + r,$$

where $q, r \in k[x_1, \ldots, x_n, y]$, $s \geq 0$, and either $r = 0$ or $deg(r, y) < m$.

We denote the pseudoremainder r of f pseudodivided by g with respect to y as $Rem(f, g, y)$. One nice property of the pseudoremainder is that, as stated in the propositoin above, $deg(r, y) < deg(f, y) = m$. Thus, if we replace f with r and repeatedly pseudodivide, we can eliminate the variable y from an equation without losing other important information about it. This operation is similar to mapping a number to its equivalence class modulo n. In the case of integer equivalence classes, we are only interested in the remainder, which functionally behaves the same as the original integer under the operations of the ring.

2.2 Triangularization

Let $f'_1, \ldots, f'_n \in k[u_1, \ldots, u_m, x_1, \ldots, x_n]$ be the hypotheses derived from our geometric system. By using pseudodivision, we can reduce these hypotheses down to triangular form. To do this, we take the f'_i that has the lowest nonzero degree in x_n and swap it with f'_n. Then, for each f'_j where $j < n$, if f'_j depends on x_n, we replace f'_j with $Rem(f'_j, f_n, x_n)$. We repeat this process until f'_n is the only hypothesis that depends on x_n. Then, we find the f'_i that has the lowest nonzero degree in x_{n-1} and swap it with f'_{n-1}. We repeat this

process until we have reached a set of hypotheses of the form

$$\begin{aligned} f_1' &= f_1'(u_1, \ldots, u_m, x_1) \\ f_2' &= f_2'(u_1, \ldots, u_m, x_2) \\ &\vdots \\ f_n' &= f_n'(u_n, \ldots, u_m, x_1, \ldots, x_n) \end{aligned}$$

2.3 Successive Pseudodivision

A quick note about notation: unfortunately, the book uses f and g to introduce pseudodivision. We divide f by g in the definition of pseudodivision. It also uses the set of $f_1, \ldots, f_n$ to represent the triangularized polynomials and uses g to represent the conclusion of our proof. This means that we end up dividing g by the set of $f_1, \ldots, f_n$ in this section, which is the opposite of what we had before. This tripped me up on my first reading, so I wanted to clarify that point here. Anyway, back to the math.

Much like in the triangularization procedure, we can use pseudodivision on any polynomial in order to eliminate the variables $x_1, \ldots, x_n$ from its remainder. We do this with the following process.

Let $R_n = g$. Then, compute $R_{i-1} = Rem(R_i, f_i, x_i)$. Each successive pseudodivision will eliminate each x_i. Additionally, we know that each x_i will not be reintroduced in a later pseudodivision because the system has been triangularized. Thus, when we reach R_0, all of the x_is will have been eliminated and we will have a polynomial that depends only on $u_1, \ldots, u_m$.

2.4 Wu's Method

Wu's method lets us use pseudodivision to determine whether a conclusion g follows from a triangularized set of hypotheses $f_1, \ldots, f_n$. Once again, however, Wu's method is not guaranteed to work on reducible varieties. As such, decomposing a variety into its irreducible components and determining V' may still be required.

Theorem 4 (Wu's Method)**.** *Let R_0 be the final remainder computed by the successive pseudodivision of g by the set of triangularized hypotheses $f_1, \ldots, f_n$ as discussed above. Let d_i be the leading coefficient of each f_i. Then, there are nonnegative integers $s_1, \ldots, s_n$ and polynomials $A_1, \ldots, A_n$ in the ring $\mathbb{R}[u_1, \ldots, u_m, x_1 \ldots, x_n]$ such that*

$$d_1^{s_1} \cdot d_2^{s_2} \cdots d_n^{s_n} g = A_1 f_1 + \cdots + A_n f_n + R_0.$$

Additionally, if R_0 is the zero polynomial, then g is zero at every point of $V' \setminus \boldsymbol{V}(d_1 d_2 \cdots d_n) \subseteq \mathbb{R}^{m+n}$.

The proof of the first part of this theorem is rather intuitive. By rearranging the definition of pseudodivision, we can see that after our first step of successive pseudodivision, we can see that

$$R_{n-1} = d_n^{s_n} g - q_n f_n.$$

When we pseudodivide again, we get

$$\begin{aligned} R_{n-2} &= d_{n-1}^{s_{n-1}} (d_n^{s_n} g - q_n f_n) - q_{n-1} f_{n-1} \\ &= d_{n-1}^{s_{n-1}} d_n^{s_n} g - q_{n-1} f_{n-1} - d_{n-1}^{s_{n-1}} q_n f_n \end{aligned}$$

From here, it is apparent that the $d_i^{s_i}$ terms will accumulate in front of g and each f_i will also be multiplied by some polynomial.

If R_0 is the zero polynomial, then

$$d_1^{s_1} \cdots d_n^{s_n} g = A_1 f_1 + \cdots + A_n f_n.$$

Therefore, for every point on the variety $\mathbf{V}(f_1, \ldots, f_n)$, each $f_i = 0$, so at least one polynomial in the product on the left side of the equation must equal zero.

We know that since V' is an irreducible component of $\mathbf{V}(f_1, \ldots, f_n)$, it must hold that $V' \subseteq \mathbf{V}(f_1, \ldots, f_n)$. Once again, for all points on V', each $f_i = 0$, so it also holds on the points in V' that at least one polynomial on the left side of the equation must equal zero. Thus, by removing the points where any $d_i = 0$ from our variety, we find that $g = 0$ on $V' \setminus \mathbf{V}(d_1 d_2 \cdots d_n)$.

I wrote an implementation of the pseudodivision algorithm in the file `6.5.6.m2` and used it to prove Example 1 from Section 6.4. I used this implementation to follow Wu's method for triangularizing the hypotheses and testing the conclusions of the parallelogram proof. I found that once I had

set up my pseudodivision algorithm, which was a more arduous process than I had anticipated, the process went very smoothly and was much less time consuming to complete than the original method presented by the book in 6.4. One fortunate occurance was that, even though the variety V was not reduced into its components, Wu's method still worked on it. Wu's method is proven to work on irreducible varieties on which the u variables are algebraically independant. As such, Wu's method may work on an unreduced variety, but it is not a necessary condition. Moreover, there are various adaptions and modifications of Wu's method that do not require V to be irreducible in order to prove theorems on V.

3 Which method is better?

Both Wu's method and the Gröbner basis method presented above have the restriction that they only work on irreducible varieties. However, both methods also have more complicated counterparts that resolve these issues. As such, Wu's method and its variants are generally recognized as the computationally preferable automatic geometric theorem prover and are used much more often than the Gröbner basis method. This is because the triangularization and pseudodivision algorithms are significantly less computationally demanding than determining the Gröbner basis of an ideal. The Gröbner basis of an ideal preserves many more nice properties of the ideal than the triangularization procedure. It is unsurprising, then, that computing a Gröbner basis is a more intricate and time-consuming process than triangularizing the same set of functions. Since the added benefits of using a Gröbner basis are not used in automatic geometric theorem proving, it is more computationally efficient to use a solver like Wu's method.